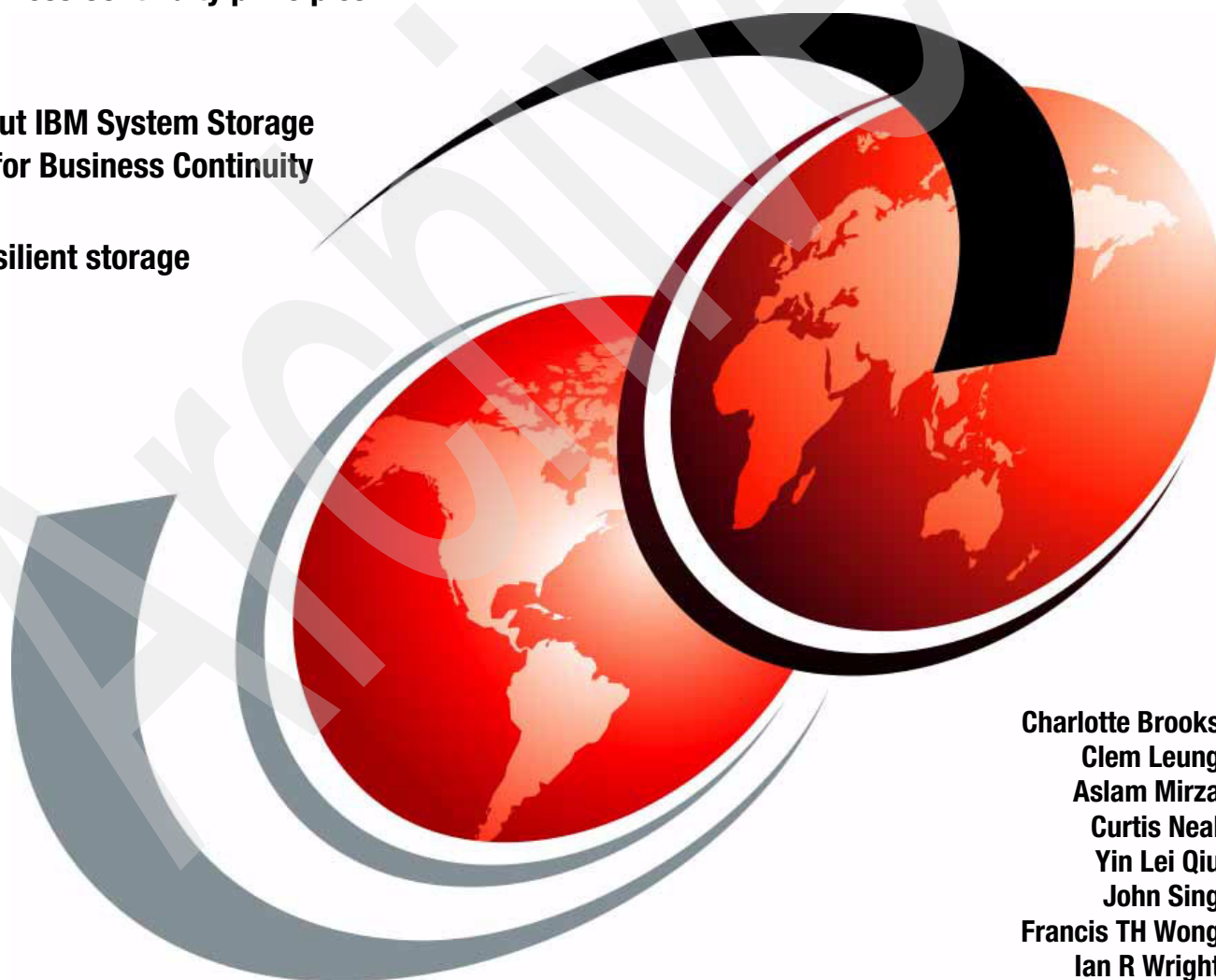


IBM System Storage Business Continuity: Part 2 Solutions Guide

Apply Business Continuity principles

Learn about IBM System Storage
products for Business Continuity

Design resilient storage
solutions



Charlotte Brooks
Clem Leung
Aslam Mirza
Curtis Neal
Yin Lei Qiu
John Sing
Francis TH Wong
Ian R Wright

Redbooks



International Technical Support Organization

**IBM System Storage Business Continuity:
Part 2 Solutions Guide**

February 2007

Archived

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

First Edition (February 2007)

This edition applies to IBM System Storage products current at the time of writing.

© Copyright International Business Machines Corporation 2007. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices	xi
Trademarks	xii
Preface	xiii
The team that wrote this redbook.	xiii
Become a published author	xvi
Comments welcome.	xvi
Part 1. Business Continuity solution offerings and solution segments	1
Chapter 1. Three Business Continuity solution segments defined	3
1.1 Tiers of Business Continuity and solution segments.	4
1.2 Blending tiers into an optimized solution.	4
1.2.1 Business continuity tiers mapped into solution segments.	5
1.2.2 Examples of Business Continuity solutions.	6
1.3 Summary.	7
Chapter 2. Continuous Availability	9
2.1 Geographically Dispersed Parallel Sysplex (GDPS).	10
2.1.1 System z Parallel Sysplex overview	10
2.1.2 Server Time Protocol (STP)	13
2.1.3 GDPS solution offerings: Introduction.	14
2.1.4 GDPS/PPRC overview	19
2.1.5 GDPS/PPRC HyperSwap Manager overview.	25
2.1.6 RCMF/PPRC overview	27
2.1.7 GDPS/XRC overview	28
2.1.8 RCMF/XRC overview	31
2.1.9 GDPS/Global Mirror (GDPS/GM) overview.	31
2.1.10 GDPS three site support.	34
2.1.11 TS7740 Grid Support (GDPS/PPRC and GDPS/XRC).	35
2.1.12 FlashCopy support (GDPS/PPRC, GDPS/XRC, GDPS/Global Mirror).	36
2.1.13 IBM Global Technology Services (GTS) offerings for GDPS overview.	36
2.1.14 Summary: Value of GDPS automation	37
2.2 GDPS components in more technical detail	38
2.2.1 GDPS/PPRC support of Consistency Group FREEZE.	38
2.2.2 HyperSwap function	41
2.2.3 GDPS Open LUN Management	46
2.2.4 GDPS/PPRC Multi-Platform Resiliency for System z	47
2.2.5 GDPS single site and multi-site workload.	49
2.2.6 GDPS extended distance support between sites	54
2.2.7 GDPS/XRC implementation	56
2.2.8 GDPS/Global Mirror in more detail	58
2.2.9 GDPS automation of System z Capacity Backup (CBU)	61
2.2.10 GDPS and TS7700 Grid support (GDPS/PPRC and GDPS/XRC)	62
2.2.11 GDPS FlashCopy support.	64
2.2.12 GDPS prerequisites	65
2.2.13 GDPS summary	66
2.2.14 Additional GDPS information	67
2.3 Geographically Dispersed Open Clusters (GDOC).	67

2.3.1	GDOC overview	67
2.3.2	GDOC in greater detail	69
2.4	HACMP/XD	71
2.4.1	HACMP/XD overview	71
2.4.2	HACMP/XD in greater detail	72
2.4.3	High Availability Cluster Multi-Processing (HACMP) for AIX	75
2.4.4	HACMP/XD for HAGEO	81
2.4.5	HAGEO cluster components	81
2.4.6	IBM Geographic Remote Mirror for AIX (GeoRM)	86
2.4.7	Cross-site LVM mirroring	88
2.4.8	Global Logical Volume Mirroring (GLVM)	89
2.4.9	AIX micro-partitions and virtualization	91
2.5	Continuous Availability for MaxDB and SAP liveCache	91
2.5.1	MaxDB and SAP liveCache hot standby overview	91
2.5.2	MaxDB and SAP liveCache hot standby in greater detail	93
2.6	Copy Services for System i	98
2.6.1	Answering the question: Why use external disk?	98
2.6.2	Independent Auxiliary Storage Pools (IASPs)	100
2.6.3	Copy Services with IASPs	101
2.6.4	Copy Services for System i	102
2.7	Metro Cluster for N series	103
Chapter 3. Rapid Data Recovery		107
3.1	System Storage Rapid Data Recovery: System z and mixed z+Open platforms (GDPS/PPRC HyperSwap Manager)	108
3.1.1	Resiliency Portfolio Positioning	108
3.1.2	Description	109
3.1.3	Additional information	112
3.2	System Storage Rapid Data Recovery for UNIX and Windows	112
3.3	System Storage Rapid Data Recovery for System z and mixed z+Distributed platforms using TPC for Replication	116
3.3.1	System Storage Resiliency Portfolio positioning	117
3.3.2	Solution description	117
3.3.3	Functionality of TPC for Replication	119
3.3.4	Functionality of TPC for Replication Two Site BC	119
3.3.5	Environment and supported hardware	120
3.3.6	Terminology	120
3.3.7	TPC for Replication session types and commands	123
3.3.8	Additional information	125
3.3.9	TPC for Replication architecture	125
3.4	IBM System Storage SAN Volume Controller (SVC)	127
3.4.1	SAN Volume Controller: FlashCopy services	128
3.4.2	SVC remote mirroring	132
3.4.3	SVC Metro Mirror	133
3.4.4	Global Mirror	139
3.4.5	Summary	151
3.5	System i storage introduction	151
3.5.1	System i storage architecture	152
3.5.2	Independent auxiliary storage pool (IASP)	152
3.5.3	System i and FlashCopy	153
3.5.4	Metro Mirror and Global Mirror	155
3.5.5	Copy Services Toolkit	157
3.5.6	System i high availability concepts: FlashCopy	158

3.5.7	System i high availability concepts: Metro Mirror	162
3.5.8	Cross Site Mirroring solutions	165
3.5.9	Summary	166
3.5.10	Additional information	166
3.6	FlashCopy Manager and PPRC Migration Manager	166
3.6.1	FlashCopy Manager overview	167
3.6.2	Tier level and positioning within the System Storage Resiliency Portfolio	167
3.6.3	FlashCopy Manager solution description	167
3.6.4	FlashCopy Manager highlights	167
3.6.5	FlashCopy Manager components	168
3.6.6	PPRC Migration Manager overview	168
3.6.7	Tier level and positioning within the System Storage Resiliency Portfolio	169
3.6.8	PPRC Migration Manager description	169
3.6.9	Diagnostic tools	170
3.6.10	Support for FlashCopy	170
3.6.11	Support for Metro Mirror Consistency Group FREEZE	170
3.6.12	Modes of operation	170
3.6.13	Use in a disaster recovery environment	171
3.6.14	PPRC Migration Manager prerequisites	171
3.6.15	Positioning of PPRC Migration Manager and GDPS	172
3.6.16	Summary	173
Chapter 4.	Backup and restore	175
4.1	An overview of backup and restore, archive and retrieve	176
4.1.1	What is backup and restore?	176
4.1.2	What is archive and retrieve?	176
4.1.3	Tape backup methodologies	176
4.1.4	Tape backup and recovery software	179
4.2	IBM DB2 for z/OS backup and recovery options	180
4.2.1	SET LOG SUSPEND	180
4.2.2	FlashCopy Manager	181
4.2.3	Backing up and restoring data in DB2 for z/OS	181
4.3	IBM Tivoli Storage Manager overview	182
4.3.1	IBM Tivoli Storage Manager solutions overview	182
4.3.2	IBM Tivoli Storage Manager solutions in detail	184
4.3.3	Tivoli Storage Manager for Copy Services: Data Protection for Exchange	195
4.3.4	Tivoli Storage Manager for Advanced Copy Services	199
4.3.5	Further information	205
4.3.6	Summary	205
4.4	IBM Data Retention 550	206
4.4.1	Retention-managed data	206
4.4.2	Storage and data characteristics	207
4.4.3	IBM strategy and key products	207
4.4.4	IBM System Storage DR550	209
4.5	System z backup and restore software	210
4.5.1	DFSMSdss	210
4.5.2	DFSMSHsm	211
4.5.3	z/VM utilities	211
4.6	Solution examples of backup, restore, and archive	211
4.6.1	BC Tier 1 — Manual off-site vaulting	212
4.6.2	BC Tier 2: Solution example	212
4.6.3	BC Tier 3: Solution example	213
4.7	Summary	214

Part 2. Business Continuity component and product overview	215
Chapter 5. Overview of IBM System Storage Resiliency Portfolio	217
5.1 System Storage Resiliency Portfolio	218
5.1.1 Reliable hardware infrastructure layer	219
5.1.2 Core technologies layer	220
5.1.3 Platform-specific integration layer	221
5.1.4 Application-specific integration layer	222
5.1.5 Summary	224
Chapter 6. Storage networking for IT Business Continuity	225
6.1 Storage networking overview	226
6.1.1 Storage attachment concepts	226
6.2 Storage Area Network (SAN)	229
6.2.1 Overview	229
6.2.2 Types of SAN implementations	231
6.2.3 Product portfolio	234
6.2.4 Disaster recovery and backup considerations	240
6.2.5 Additional information	245
6.3 Network Attached Storage (NAS)	245
6.3.1 Overview	245
6.4 iSCSI	250
6.4.1 Business Continuity considerations	251
6.4.2 Additional information	251
6.5 Booting from the SAN	251
Chapter 7. IBM System Storage DS6000, DS8000, and ESS	253
7.1 IBM System Storage DS6000 and DS8000	254
7.2 The IBM System Storage DS6000 series	255
7.2.1 Positioning	255
7.2.2 DS6000 models	255
7.2.3 Hardware overview	256
7.2.4 Storage capacity	258
7.2.5 DS management console	259
7.2.6 Supported servers environment	259
7.3 The IBM System Storage DS8000 series	260
7.3.1 Positioning	260
7.3.2 DS8000 models	260
7.3.3 Hardware overview	261
7.3.4 Storage capacity	263
7.3.5 Storage system logical partitions (LPARs)	264
7.4 The IBM TotalStorage ESS 800	265
7.4.1 ESS800 models	265
7.4.2 Hardware overview	265
7.4.3 Supported servers environment	267
7.5 Advanced Copy Services for DS8000/DS6000 and ESS	267
7.6 Introduction to Copy Services	269
7.7 Copy Services functions	270
7.7.1 Point-In-Time Copy (FlashCopy)	270
7.7.2 FlashCopy options	272
7.7.3 Remote Mirror and Copy (Peer-to-Peer Remote Copy)	277
7.7.4 Comparison of the Remote Mirror and Copy functions	286
7.7.5 What is a Consistency Group?	287
7.8 Interfaces for Copy Services	291

7.8.1 Storage Hardware Management Console (S-HMC)	291
7.8.2 DS Storage Manager Web-based interface	292
7.8.3 DS Command-Line Interface (DS CLI)	292
7.8.4 DS Open application programming Interface (API).	293
7.9 Interoperability with ESS	293
Chapter 8. The IBM System Storage DS4000	295
8.1 DS4000 series.	296
8.1.1 Introducing the DS4000 series hardware overview.	296
8.1.2 DS4000 Storage Manager Software	298
8.2 DS4000 copy functions	298
8.3 Introduction to FlashCopy	299
8.4 Introduction to VolumeCopy	300
8.5 Introduction to Enhanced Remote Mirroring	302
8.5.1 Metro Mirroring (synchronous mirroring).	303
8.5.2 Global Copy (asynchronous mirroring without write consistency group)	304
8.5.3 Global Mirroring (asynchronous mirroring with write consistency group)	306
8.6 Mirror repository logical drives	307
8.7 Primary and secondary logical drives	308
8.7.1 Logical drive parameters, roles, and maximum number of mirrored pairs	309
8.7.2 Host Accessibility of secondary logical drive.	309
8.7.3 Mirrored logical drive controller ownership	310
8.7.4 Enhanced Remote Mirroring and FlashCopy Premium Feature	310
8.7.5 Enhanced Remote Mirroring and VolumeCopy Premium Feature	310
8.7.6 Volume role compatibility	310
8.8 Data resynchronization process	311
8.9 SAN fabric and TCP/IP connectivity	313
8.9.1 SAN fabric and SAN zoning configuration	313
8.10 ERM and disaster recovery.	316
8.10.1 Role reversal concept.	317
8.10.2 Re-establishing Remote Mirroring after failure recovery.	317
8.10.3 Link interruptions.	318
8.10.4 Secondary logical drive error	318
8.10.5 Primary controller failure	319
8.10.6 Primary controller reset.	319
8.10.7 Secondary controller failure	319
8.10.8 Write Consistency Group and Unsynchronized State.	319
8.11 Performance considerations	319
8.11.1 Synchronization priority.	320
8.11.2 Synchronization performance and logical drive settings.	320
8.11.3 Mirroring mode and performance	321
8.11.4 Mirroring connection distance and performance.	321
8.12 Long-distance ERM.	322
8.13 Summary.	324
Chapter 9. IBM System Storage N series	325
9.1 N series hardware overview	326
9.1.1 System Storage N3700.	327
9.1.2 System Storage N5000 series	329
9.1.3 System Storage N7000 series	331
9.2 N series expansion units.	334
9.2.1 EXN1000 expansion unit	334
9.2.2 EXN2000 expansion unit	334

9.3 N series software overview	334
9.3.1 The IBM N series standard software features.	334
9.3.2 N series optional software features.	335
9.3.3 Details of advanced copy service functions	337
9.4 More information	350
Chapter 10. DS300 and DS400	351
10.1 DS300 and DS400 overview	352
10.2 Introduction	352
10.3 IBM TotalStorage DS400 and DS300	352
10.4 Data protection	355
10.4.1 DS300, DS400 copy functions	355
10.5 Disaster recovery considerations	357
Chapter 11. Storage virtualization products	359
11.1 Storage virtualization overview	360
11.1.1 Levels of storage virtualization	360
11.1.2 SNIA shared storage model	362
11.1.3 The Storage Management Initiative (SMI)	363
11.1.4 Multiple level virtualization example	364
11.2 IBM System Storage SAN Volume Controller	365
11.2.1 Glossary of commonly used terms	365
11.2.2 Overview	366
11.2.3 SVC copy functions.	369
11.2.4 Business Continuity considerations with SAN Volume Controller.	376
11.3 IBM System Storage N series Gateway	377
11.4 Volume Managers	381
11.4.1 Overview	381
11.5 Enterprise Removable Media Manager.	382
11.5.1 Introduction to Enterprise Removable Media Manager.	382
11.5.2 eRMM central reporting and monitoring	384
11.5.3 eRMM logical components and functions	385
11.5.4 eRMM control flow	387
11.5.5 Supported features (eRMM 1.2.4).	389
11.5.6 Supported platforms (eRMM 1.2.4).	390
11.5.7 Strategic fit and positioning.	390
Chapter 12. Storage management software	391
12.1 Storage management strategy	392
12.2 Key storage management areas	392
12.3 Storage management software solutions	393
12.4 IBM TotalStorage Productivity Center.	394
12.4.1 IBM TotalStorage Productivity Center for Data.	394
12.4.2 IBM TotalStorage Productivity Center for Fabric.	397
12.4.3 IBM TotalStorage Productivity Center for Disk	398
12.4.4 IBM TotalStorage Productivity Center for Replication.	399
12.4.5 Summary.	400
12.5 IBM Tivoli Storage Manager	400
12.5.1 Backup methods	401
12.5.2 Disaster Recovery Manager	403
12.5.3 Disaster Recovery for the Tivoli Storage Manager Server	404
12.5.4 IBM Tivoli Storage Manager for Databases	407
12.5.5 IBM Tivoli Storage Manager for Mail.	411
12.5.6 IBM Tivoli Storage Manager for Application Servers	413

12.5.7 IBM Tivoli Storage Manager for Enterprise Resource Planning	414
12.5.8 Tivoli Storage Manager for Advanced Copy Services.	416
12.6 Tivoli Storage Manager for Copy Services	417
12.6.1 IBM Tivoli Storage Manager for Space Management.	417
12.6.2 Tivoli Storage Manager for HSM for Windows	421
12.6.3 Tivoli Continuous Data Protection for Files.	421
12.6.4 IBM System Storage Archive Manager.	424
12.6.5 LAN-free backups: Overview	426
12.7 Bare Machine Recovery	427
12.7.1 Cristie Bare Machine Recovery (CBMR).	427
12.7.2 Tivoli Storage Manager support for Automated System Recovery	429
12.7.3 Tivoli Storage Manager for System Backup and Recovery.	430
12.8 DFSMS family of products	431
12.8.1 DFSMSdftp (data facility product)	431
12.8.2 DFSMSdftp Copy Services	432
12.8.3 DFSMSHsm (Hierarchical Storage Manager)	432
12.8.4 DFSMSHsm Fast Replication	433
12.8.5 DFSMSHsm Disaster Recovery using ABARS	434
12.9 IBM System Storage and TotalStorage Management Tools.	436
12.9.1 DFSMSStvs Transactional VSAM Services	436
12.9.2 CICS/VSAM Recovery	437
12.9.3 IBM TotalStorage DFSMSHsm Monitor.	438
12.9.4 Tivoli Storage Optimizer for z/OS	438
12.9.5 Mainstar Mirroring Solutions/Volume Conflict Rename (MS/VCR)	439
12.9.6 Mainstar Catalog RecoveryPlus	440
12.9.7 Mainstar FastAudit/390	441
12.9.8 Mainstar disaster recovery utilities	441
12.10 IBM TotalStorage Enterprise Tape Library (ETL) Expert	443
12.11 IBM TotalStorage Specialists	444
12.11.1 IBM TotalStorage Tape Library Specialist.	444
12.11.2 IBM TotalStorage Peer-to-Peer VTS Specialist	445
Chapter 13. Tape and Business Continuity	447
13.1 Positioning of tape in Business Continuity	448
13.2 Why tape requires a longer term commitment than disk.	449
13.2.1 Typical tape media migration plans.	449
13.3 What are the key attributes to a tape strategy?	449
13.3.1 Proprietary or Open Standard.	451
13.3.2 Proven technology endorsement	451
13.3.3 Performance	451
13.3.4 Capacity	451
13.3.5 Reliability.	452
13.3.6 Scalability	452
13.3.7 Investment protection	452
13.3.8 Application support	452
13.3.9 Simple to manage.	453
13.3.10 Total cost of ownership.	453
13.4 Tape strategies	453
13.4.1 Why tape backups are necessary.	453
13.4.2 Tape applications	454
13.5 Available tape technologies	456
13.5.1 Tape automation.	456
13.5.2 Tape drives	457

13.5.3	Criteria for selecting a tape drive	457
13.5.4	Tape cartridge usage considerations	459
13.5.5	Tape drive technology	459
13.6	Recent tape technology advancements	461
13.6.1	Tape drive enhancements	461
13.6.2	Tape automation enhancement	462
13.7	TS7740 Virtualization Engine overview	462
13.7.1	TS7740 Virtualization Engine and business continuity	463
13.7.2	TS7740 operation overview	466
13.7.3	TS7740 Grid overview	468
13.7.4	TS7740 and GDPS (Tier 7) Implementation	469
13.8	IBM Virtualization Engine TS7510 overview	473
13.8.1	TS7510 and Business Continuity	473
13.8.2	TS7510 configuration overviews	473
13.8.3	IBM Virtualization Engine TS7510 software	476
13.9	IBM System Storage TS1120 tape drive and tape encryption	479
13.9.1	TS1120 tape encryption and business continuity	480
13.10	Disaster recovery considerations for tape applications	485
Appendix A. System Storage resources and support		489
	Where to start	490
	Advanced Technical Support: System Storage	490
	IBM System Storage Solution Centers	490
	IBM System Storage Proven Solutions	491
	IBM Global Services: Global Technology Services	491
	Solutions integration	491
	GDPS Technical Consulting Workshop	491
	Broad-ranged assessment	492
	Business Resiliency & Continuity Services (BRCS)	493
	ClusterProven program	494
	Benefits of ClusterProven validation	494
	ClusterProven solutions	494
	Support considerations	494
	Solution Assurance	496
Appendix B. IBM System Storage Copy Services function comparison		499
	IBM System Storage point-in-time copy comparison	500
	FlashCopy function definitions	500
	IBM System Storage disk mirroring comparison	501
	Disk mirroring function definitions	502
Related publications		505
	IBM Redbooks	505
	Other publications	506
	How to get IBM Redbooks	506
	Help from IBM	506
Index		507

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

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

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


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

COPYRIGHT LICENSE:

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

Trademarks

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

AIX®	eServer™	Redbooks™
AIX 5L™	FICON®	Redbooks (logo)  ™
AS/400®	FlashCopy®	S/390®
BladeCenter®	GDPS®	ServeRAID™
Calibrated Vectors Cooling™	Geographically Dispersed Parallel Sysplex™	Sysplex Timer®
CICS®	HACMP™	System i™
ClusterProven®	HyperSwap™	System i5™
Common User Access®	IBM®	System p™
Cross-Site®	IMS™	System p5™
CUA®	IMS/ESA®	System x™
DB2®	Informix®	System z™
DB2 Universal Database™	iSeries™	System Storage™
DFSMSdfp™	i5/OS®	System Storage Proven™
DFSMSdss™	Lotus®	SysBack™
DFSMSHsm™	Magstar®	Tivoli®
DFSMSrmm™	MVS™	Tivoli Enterprise™
Domino®	NetView®	Tivoli Enterprise Console®
DS4000™	OS/390®	TotalStorage®
DS6000™	OS/400®	Virtualization Engine™
DS8000™	Parallel Sysplex®	VSE/ESA™
ECKD™	PowerPC®	WebSphere®
Enterprise Storage Server®	POWER™	z/OS®
Enterprise Systems Connection Architecture®	POWER5™	z/VM®
ESCON®	RACF®	zSeries®

The following terms are trademarks of other companies:

Oracle, JD Edwards, PeopleSoft, and Siebel are registered trademarks of Oracle Corporation and/or its affiliates.

mySAP, SAP, and SAP logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries.

Snapshot, SecureAdmin, WAFL, SyncMirror, SnapVault, SnapValidator, SnapRestore, SnapMover, SnapMirror, SnapManager, SnapDrive, MultiStore, FilerView, DataFabric, Data ONTAP, and the Network Appliance logo are trademarks or registered trademarks of Network Appliance, Inc. in the U.S. and other countries.

Java, Nearline, Solaris, StorageTek, Sun, Sun Microsystems, and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Excel, Internet Explorer, Microsoft, Outlook, PowerPoint, Windows NT, Windows Server, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

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

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

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

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

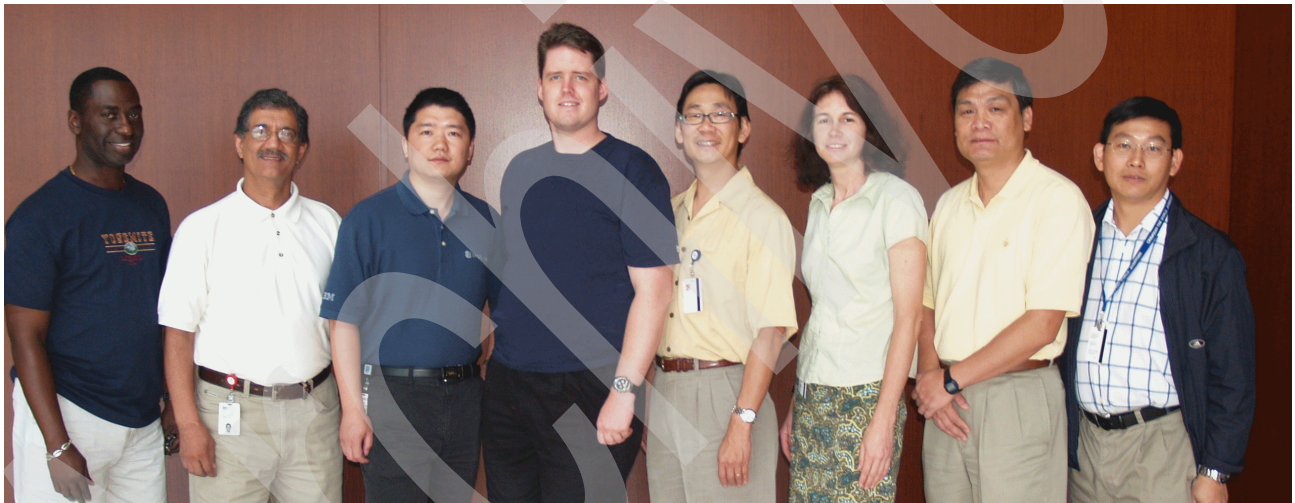
Preface

This IBM® Redbook is a companion to the *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547. IBM System Storage Business Continuity Guide: Part 1 Planning Guide We assume that the reader of this book has understood the concepts of Business Continuity planning described in that book.

In this book we explore IBM® System Storage™ solutions for Business Continuity, within the three segments of Continuous Availability, Rapid Recovery, and Backup and Restore. We position these solutions within the Business Continuity tiers. We describe, in general, the solutions available in each segment, then present some more detail on many of the products. In each case, we point the reader to sources of more information.

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.



The team: Curtis, Aslam, Yin Lei, Ian, John, Charlotte, Clem, and Francis

Charlotte Brooks is an IBM Certified IT Specialist and Project Leader for Storage Solutions at the International Technical Support Organization, San Jose Center. She has 15 years of experience with IBM in storage hardware and software support, deployment, and management. She has written many IBM Redbooks™, and has developed and taught IBM classes in all areas of storage and storage management. Before joining the ITSO in 2000, she was the Technical Support Manager for Tivoli® Storage Manager in the Asia Pacific Region.

Clem Leung is an Executive IT Architect with the IBM Global Small and Medium Business sector, supporting emerging and competitive clients. He specializes in IT infrastructure simplification and Business Continuity technologies and solutions. Previously, he was in worldwide technical sales support for IBM storage and storage networking solutions and products. Clem has worked for IBM for 25 years in various technical sales capacities, including networking, distributed computing, data center design, and more. He was a co-author of the IBM Redbooks, *IBM System Storage Business Continuity Solutions Overview*, SG24-6684 and *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.

Aslam Mirza is a Certified Senior Consulting Storage Specialist in New York, working as a pre-sales advisor for enterprise storage topics. He has more than 30 years of experience with IBM large systems, storage systems, tape systems and system storage resiliency portfolio. His area of expertise is strategy and design of storage solutions. He was a co-author of the IBM Redbooks, *IBM System Storage Business Continuity Solutions Overview*, SG24-6684 and *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.

Curtis Neal is a Senior IT Specialist working for the System Storage Group in San Jose, California. He has over 25 years of experience in various technical capacities including mainframe and open system test, design and implementation. For the past 6 years, he has led the Open Storage Competency Center, which helps clients and Business Partners with the planning, demonstration, and integration of IBM System Storage Solutions. He was a co-author of the IBM Redbooks, *IBM System Storage Business Continuity Solutions Overview*, SG24-6684 and *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.

Yin Lei Qiu is a senior IT specialist working for the Storage Systems Group in Shanghai, China. He is the leader of the storage technical team in East China and a pre-sales advisor, and provides technical support storage solutions to IBM professionals, Business Partners, and Clients. He has more than six years of solution design experience with IBM Enterprise Disk Storage Systems, Midrange Disk Storage Systems, NAS Storage Systems, Tape Storage Systems, Storage Virtualization Systems, and the System Storage Resiliency Portfolio. He was a co-author of the IBM Redbooks, *IBM System Storage Business Continuity Solutions Overview*, SG24-6684 and *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.

John Sing is a Senior Consultant with IBM Systems and Technology Group, Business Continuity Strategy and Planning. He helps with planning and integrating IBM System Storage products into the overall IBM Business Continuity strategy and product portfolio. He started in the Business Continuity arena in 1994 while on assignment to IBM Hong Kong, and IBM China. In 1998, John joined the IBM ESS planning team for PPRC, XRC, and FlashCopy®, and then in 2000, became the Marketing Manager for the ESS Copy Services. In 2002, he joined the Systems Group. John has been with IBM for 23 years. He was a co-author of the IBM Redbooks, *IBM System Storage Business Continuity Solutions Overview*, SG24-6684 and *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.

Francis TH Wong is a storage solution architect for Asia Pacific, where he provides training and technical support to the regional storage team, as well as designing client storage solutions. He has 20 years IT experience in various positions with IBM in both Australia and Hong Kong, including data center operations and S/390® storage support, as well as client sales, technical support, and services. His areas of expertise include Business Continuity solutions for mainframe and open systems, disk, tape, and virtualization. He was a co-author of the IBM Redbooks, *IBM System Storage Business Continuity Solutions Overview*, SG24-6684 and *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.

Ian R Wright is a Senior IT Specialist with Advanced Technical Support, in Gaithersburg, and is part of the Business Continuity Center of Competence. He holds a Bachelor of Science in Business Administration degree from Shippensburg University of Pennsylvania. He has 7 years of IT experience, encompassing Advanced Business Continuity Solutions, network connectivity, and GDPS® for the S/390 division. He has written educational material on Business Continuity and taught at the Business Continuity Top Gun. He was a co-author of the IBM Redbooks, *IBM System Storage Business Continuity Solutions Overview*, SG24-6684 and *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.

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

Gustavo Castets, Bertrand Dufrasne, Babette Haeusser, Emma Jacobs, Mary Lovelace,
Alex Osuna, Jon Tate, Yvonne Lyon
International Technical Support Organization, San Jose Center

Michael Stanek
IBM Atlanta

Steven Cook, Douglas Hilken, Bob Kern
IBM Beaverton

Tony Abete, David Sacks
IBM Chicago

Shawn Bodily, Dan Braden, Mike Herrera, Eric Hess, Judy Ruby-Brown, Dan Sunday
IBM Dallas

Bill Wiegand
IBM Fort Wayne

Craig Gordon, Rosemary McCutchen, David Petersen,
IBM Gaithersburg

Thomas Luther
IBM Germany

Manny Cabezas
IBM Miami

Nick Clayton
IBM Portsmouth

Noshir Dhondy, Scott Epter, David Raften
IBM Poughkeepsie

John Foley, Harold Pike
IBM Raleigh

Selwyn Dickey
IBM Rochester

Jeff Barckley, Charlie Burger, Don Chesarek, Pete Danforth, Scott Drummond, John Hulse,
Tricia Jiang, Sathees Kodi, Vic Peltz, John Power, Peter Thurston
IBM San Jose

Greg Gendron
IBM San Ramon

Chooi Ling Lee
IBM Singapore

Mark S. Putnam
IBM St Louis

Thomas Maher
IBM Southfield

Matthias Werner
IBM Switzerland

Bob Bartfai, Ken Boyd, James Bridges, Ken Day, Brad Johns, Carl Jones, Greg McBride, JD Metzger, Jon Peake, Tony Pearson, Gail Spear, Paul Suddath, Steve West
IBM Tucson

Patrick Keyes
IBM UK

Cristian Svensson
Cristie

Tom and Jenny Chang and their staff
Garden Inn, Los Gatos

Become a published author

Join us for a two- to six-week residency program! Help write an IBM Redbook dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You will have the opportunity to team with IBM technical professionals, Business Partners, and Clients.

Your efforts will help increase product acceptance and client satisfaction. As a bonus, you will develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:
ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

ibm.com/redbooks

- Send your comments in an email to:

redbooks@us.ibm.com

- Mail your comments to:

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



Part 1

Business Continuity solution offerings and solution segments

In this part of the book, we describe the three segments of Business Continuity and the solutions contained in that segment.

We cover the following topics:

- ▶ The three Business Continuity Solution segments, defined:
 - Backup/Restore
 - Rapid Data Recovery
 - Continuous Availability
- ▶ Business Continuity Solution offerings for each segment

Archived



Three Business Continuity solution segments defined

The concept of having tiers of Business Continuity can help us to organize the various technologies into useful subsets that are much easier to evaluate and manage. In this chapter we define the three solution segments for Business Continuity.

1.1 Tiers of Business Continuity and solution segments

As we have seen, the tiers of Business Continuity concept, shown in Figure 1-1, provides a generalized view of the continuum of today's business continuity technologies. The tiers concept assists us in organizing the various technologies into useful subsets that are much easier to evaluate and manage.

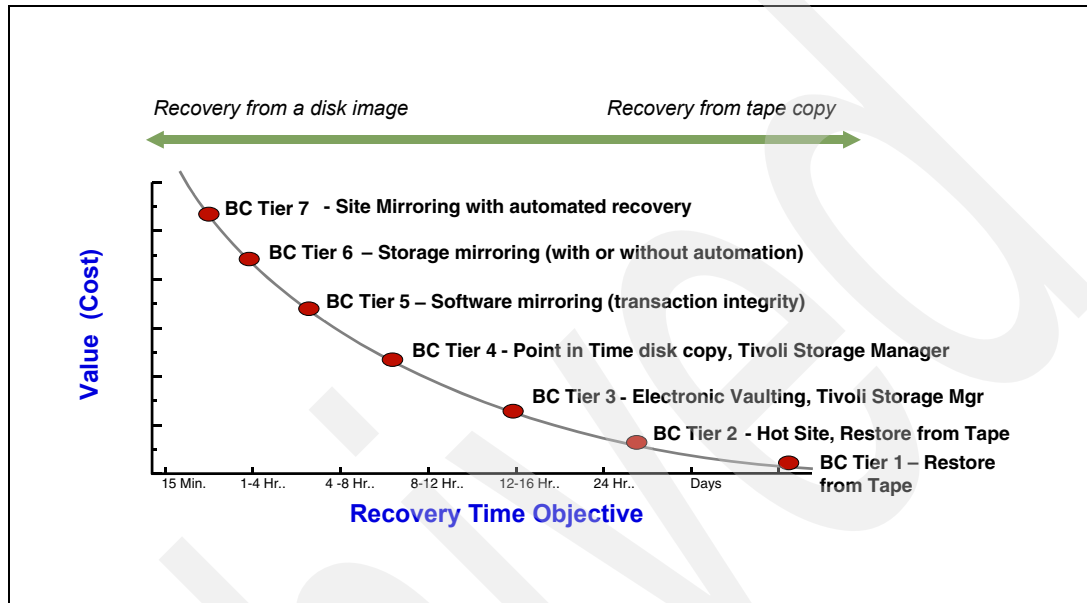


Figure 1-1 Tiers chart with representative technologies

Most organizations, of course, do not use all seven tiers. It is reasonable to assume that a selected blend of the Backup/Restore, Rapid Data Recovery, and Continuous Availability solutions, tiered and blended, is an appropriate strategy to achieve a cost-optimized strategic solution, across a set of applications that have been segmented into their appropriate tier of Business Continuity recovery.

These sets of solutions can fit the specific tier requirement of any given client; and can be used in a building block fashion for future improvements in Business Continuity.

1.2 Blending tiers into an optimized solution

To use the tiers to derive a blended, optimized enterprise business continuity architecture, we suggest the following steps:

1. Categorize the business' entire set of business processes and associated applications into three bands:
 - Low Tolerance to Outage (Continuous Availability)
 - Somewhat Tolerant to Outage (Rapid Data Recovery)
 - Very Tolerant to Outage (Backup/Restore)

Some applications, which are not in and of themselves critical, feed the critical applications. Therefore, those applications would have to be included in the higher tier.

2. Once we have segmented the business processes and applications (as best we can) into the three bands, we usually select one best strategic business continuity methodology for

that band. The contents of the tiers are the *candidate technologies* from which the strategic methodology is chosen.

The most common result, from an enterprise standpoint, is a strategic architecture of three solution segments in a blended Business Continuity solution. Three solution segments generally appear as an optimum number. At the enterprise level, two tiers generally are insufficiently optimized (in other words, overkill at some point and underkill at others), and four tiers are more complex but generally do not provide enough additional strategic benefit.

1.2.1 Business continuity tiers mapped into solution segments

To match this best practice of separating applications by tier, the IBM Business Continuity solutions are mapped into this three-band segmentation, as shown in Figure 1-2.

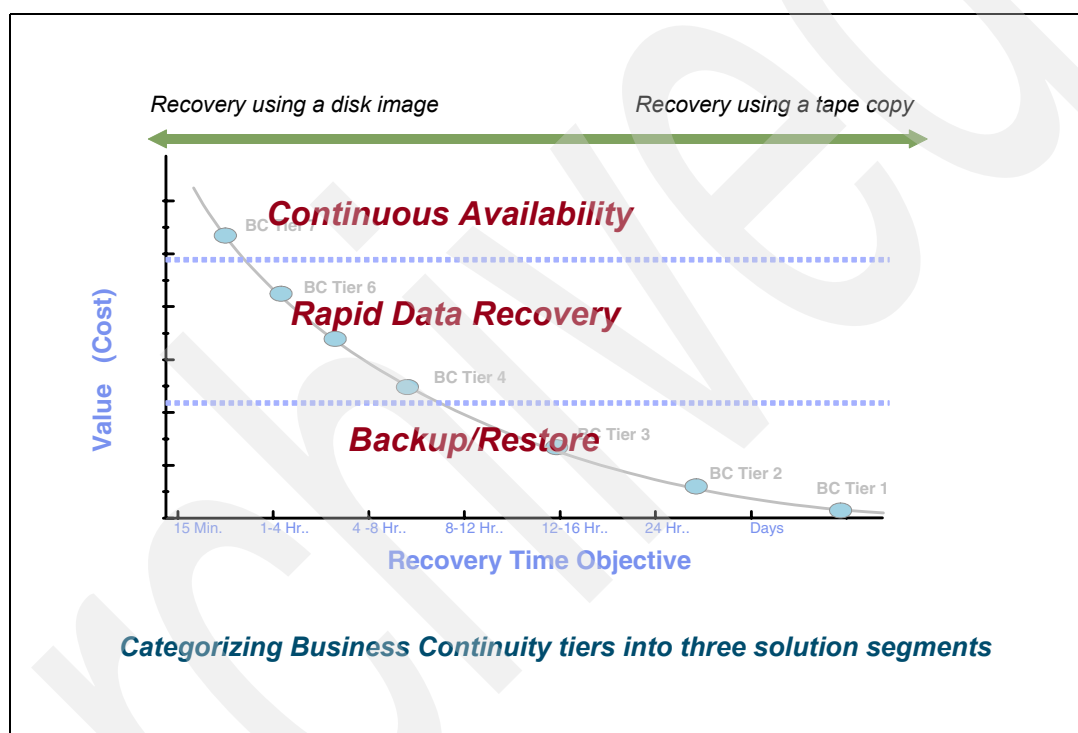


Figure 1-2 Categorize the tiers into Business Continuity Solution Segments

The three solution segments are defined in the following paragraphs.¹

Continuous Availability

This segment includes the following characteristics:

- ▶ 24x7 application and data availability (server, storage, network availability)
- ▶ Automated failover of total systems / site failover
- ▶ Very fast and transparent recovery of servers, storage, network
- ▶ Ultimate Disaster Recovery - protection against site disasters, system failures
- ▶ General RTO guideline: minutes to < 2 hours

¹ Note that the stated Recovery Time Objectives in the chart's Y-axis and given in the definitions are **guidelines** for comparison only. RTO can and do vary depending on the size and scope of the solution.

Rapid Data Recovery

This segment includes the following characteristics:

- ▶ High availability of data and storage systems (storage resiliency)
- ▶ Automated or manual failover of storage systems
- ▶ Fast recovery of data/storage from disasters or storage system failures
- ▶ Disaster Recovery from replicated disk storage systems
- ▶ General RTO guideline: 2 to 8 hours

Backup/Restore

This segment includes the following characteristics:

- ▶ Backup and restore from tape or disk
- ▶ Disaster Recovery from tape
- ▶ RTO= 8 hours to days

1.2.2 Examples of Business Continuity solutions

Each of these solution segments has a series of packaged, integrated IBM System Storage Business Continuity solutions.

An IBM System Storage Business Continuity solution is made up of the components listed in Example 1-1.

Example 1-1 Definition of an IBM System Storage Business Continuity solution

1. A solution name
 2. A tier level
 3. A specifically selected set of solution components
 4. Code that integrates and automates the solution components
 5. Services and skills to implement and tailor this named solution
-

As an overview, IBM System Storage Business Continuity solutions in the **Continuous Availability** solution segment are Tier 7 solutions. They include (but are not limited to):

- ▶ **System p™**: AIX HACMP/XD
- ▶ **System z™**: GDPS

IBM System Storage Business Continuity solutions in the **Rapid Data Recovery** solution segment are Tier 4 to Tier 6 solutions. They include (but are not limited to):

- ▶ **Heterogeneous high end disk replication**: TotalStorage® Productivity Center for Replication
- ▶ **Heterogeneous open system disk vendor mirroring**: IBM SAN Volume Controller Metro Mirror
- ▶ **System z**: GDPS HyperSwap™ Manager

IBM System Storage Business Continuity solutions in the **Backup/Restore** solution segment are Tier 1 to Tier 4. They include (but are not limited to):

- ▶ Tivoli Storage Manager for Copy Services
- ▶ Tivoli Storage Manager for Advanced Copy Services
- ▶ Tivoli Storage Manager for Databases
- ▶ Tivoli Storage Manager for Mail
- ▶ Tivoli Storage Manager for ERP
- ▶ N series Snapshot™ software integration products

The foregoing solutions are just some examples of the range of solutions available. In the rest of this book, we provide details of most (if not all) of the solutions in each segment.

1.3 Summary

The IBM System Storage Resiliency Portfolio is made up of products from various product brand families and solutions that directly map to the roadmap to IT Business Continuity.

Through provision of an integrated set of business continuity solutions, the System Storage Resiliency Portfolio is designed to provide a spectrum of integrated solutions that can provide business continuity capability for any level of desired recovery and budget.

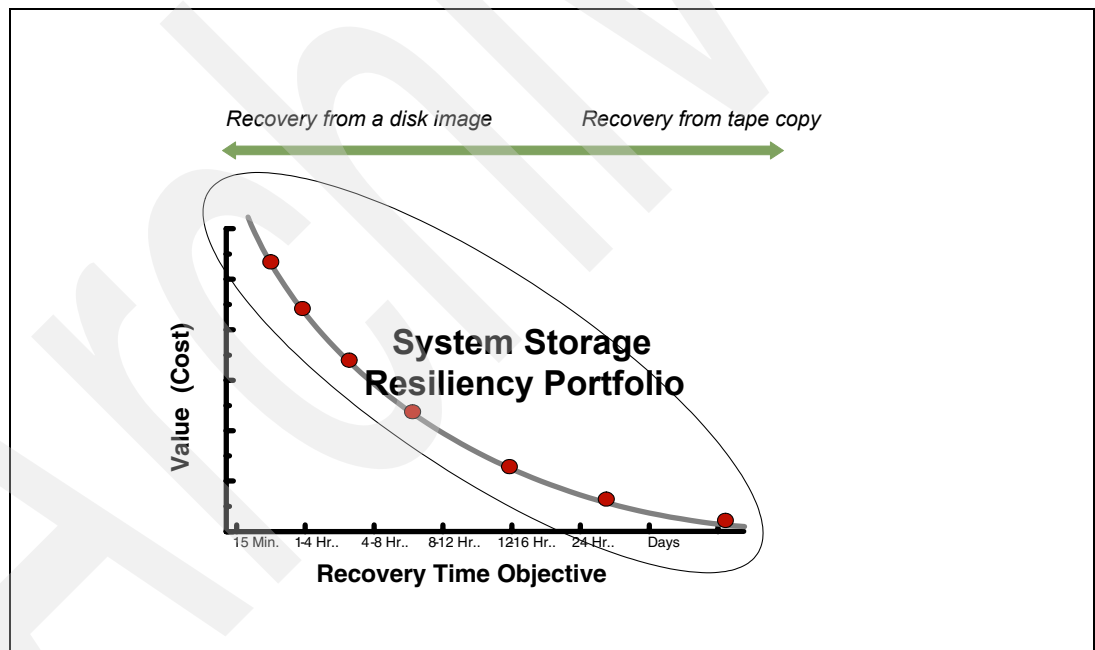


Figure 1-3 System Storage Resiliency Portfolio summary

In the following chapters, we go into more detail on key solutions that are contained with the solution segments of the System Storage Resiliency Portfolio.

Archived

Continuous Availability

At a minimum, personnel in all environments who are interested in maintaining application availability require procedures on how to do so. They have to understand how, and in what order, the applications, server, and storage are to be brought down in an orderly way. They have to understand how to recover the disk and how to bring up the applications. In disasters, they have to be prepared to recover any data that might have been rendered inconsistent in the recovery location due to a rolling disaster. Then, when it is no longer necessary to be in the alternate, they have to be able to roll the entire operation back.

If this sounds complex, it really is. The risk increases even more because of natural confusion in a disaster situation, particularly if it has directly affected some key personnel. If people are focused on other priorities during a disaster, such as their personal shelter or safety, they are distracted from working on the recovery process. Also, some employees who would have otherwise worked to aid in the recovery might be rendered unavailable one way or another.

Based on all of these factors, the human element represents risk. While they do not replace the necessity for good planning and policies, Continuous Availability technologies help mitigate that risk in both planned and unplanned outages by reducing the amount of human interaction required, by using well planned and supported control software. This enables servers and applications to be kept available more reliably and faster than they would be through a manual process.

In this chapter we describe the following Continuous Availability Technologies:

- ▶ Geographically Dispersed Parallel Sysplex™ (GDPS)
- ▶ Geographically Dispersed Open Clusters (GDOC)
- ▶ HACMP/XD
- ▶ Continuous Availability for MaxDB and SAP® liveCache
- ▶ Copy Services for System i™
- ▶ Metro Cluster for N series

For general information about Continuous Availability solutions, see the Web site:

- ▶ http://www-03.ibm.com/servers/storage/solutions/business_continuity/continuous_availability/technical_details.html.

2.1 Geographically Dispersed Parallel Sysplex (GDPS)

Geographically Dispersed Parallel Sysplex (GDPS) offerings are a family of System z-based Business Continuity solutions that include the ability to do Continuous Availability or Rapid Data Recovery, depending on the nature of the outage and the type of GDPS in use.

Our discussion of GDPS is organized in the following way:

- ▶ A review of the base System z Parallel Sysplex technology, which provides a key Business Continuity technology in its own right for System z
- ▶ A GDPS overview and introduction, in “GDPS solution offerings” on page 15
- ▶ A discussion of selected GDPS technology components in more detail, in 2.2, “GDPS components in more technical detail” on page 38

2.1.1 System z Parallel Sysplex overview

IBM clients of every size, and across every industry, are looking for ways to make their businesses more productive and more resilient in the face of change and uncertainty. They require the ability to react to rapidly changing market conditions, manage risk, outpace their competitors with new capabilities, and deliver clear returns on investment. The System z server family delivers the mainframe expectations in terms of security, reliability, availability, manageability and recoverability. In a System z Central Electronic Complex (CEC), you can run a single z/OS® image or multiple z/OS images. With IBM Parallel Sysplex technology, you can harness the power of up to 32 z/OS systems, making them behave like a single system. You can have a Parallel Sysplex either in a multi-image single CEC or in a multi-image multi-CEC environment.

The IBM Parallel Sysplex technology is a highly advanced z/OS clustered system that aggregates the capacity of multiple z/OS systems to execute common workloads.

Parallel Sysplex is not required for the GDPS/XRC or GDPS/Global Mirror environments, but is supported in their installations.

Parallel Sysplex is required for full GDPS/PPRC, and for GDPS/PPRC HyperSwap Manager.

System z Parallel Sysplex

There are three main types of clustering environments.

The first type is a high-availability or *failover* type, where in case of a failure of the primary server, the processing is transferred to the backup server (see Figure 2-1). For the maximum level of readiness, the backup server should be kept idle, which raises the cost.

The second type uses *parallel clusters*, where clusters share the workload and provide availability and scalability. The major drawback of this clustering technique is that the workload should be carefully distributed between clusters or you could incur under utilization or over utilization. Other drawbacks include database re-partitioning and distributing paths to data that could require a lot of work during the initial configuration and afterwards, when your workload grows or your business requirements change.

The most efficient form of clustering is the third type, where the clustered servers appear as a *single system* to the applications (see Figure 2-2). This is a *shared data, shared access* approach where every cluster node has access to all of the data and could run any part of your business transactions or batch jobs in parallel on the available processor capacity. Other benefits include automatic workload balancing for continuous availability, continuous operation and scalability, and a single point of control.

The IBM Parallel Sysplex technology is the System z implementation of the single system image type of clustering.

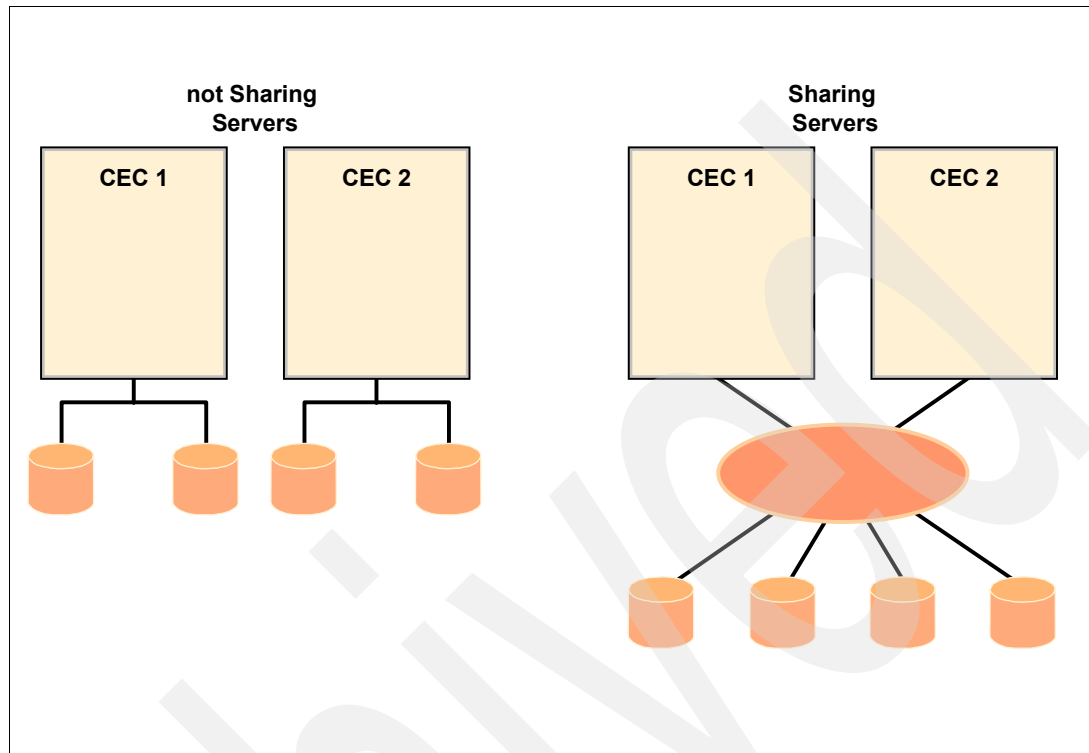


Figure 2-1 Independent servers and shared data, shared access servers

Components of a System z Parallel Sysplex

In this section, we briefly describe the components of a Parallel Sysplex:

- ▶ System z server or CEC (Central Electronic Complex)
- ▶ Single/multiple z/OS images
- ▶ ICF/CF (Internal Coupling Facility/Coupling Facility)
- ▶ Sysplex timer or Server Timer Protocol

The coupling facility is the repository of all the data structures used to allow the z/OS images to share resources and data. It can be redundant and duplexed for maximum availability. It consists of special microcode running on a general purpose or dedicated processor. It can reside on standalone CF hardware, a CF LPAR on a CEC, or be defined as an internal CF using dedicated processors.

The Sysplex timer is a dedicated hardware component used to give time consistency to the entire Parallel Sysplex complex.

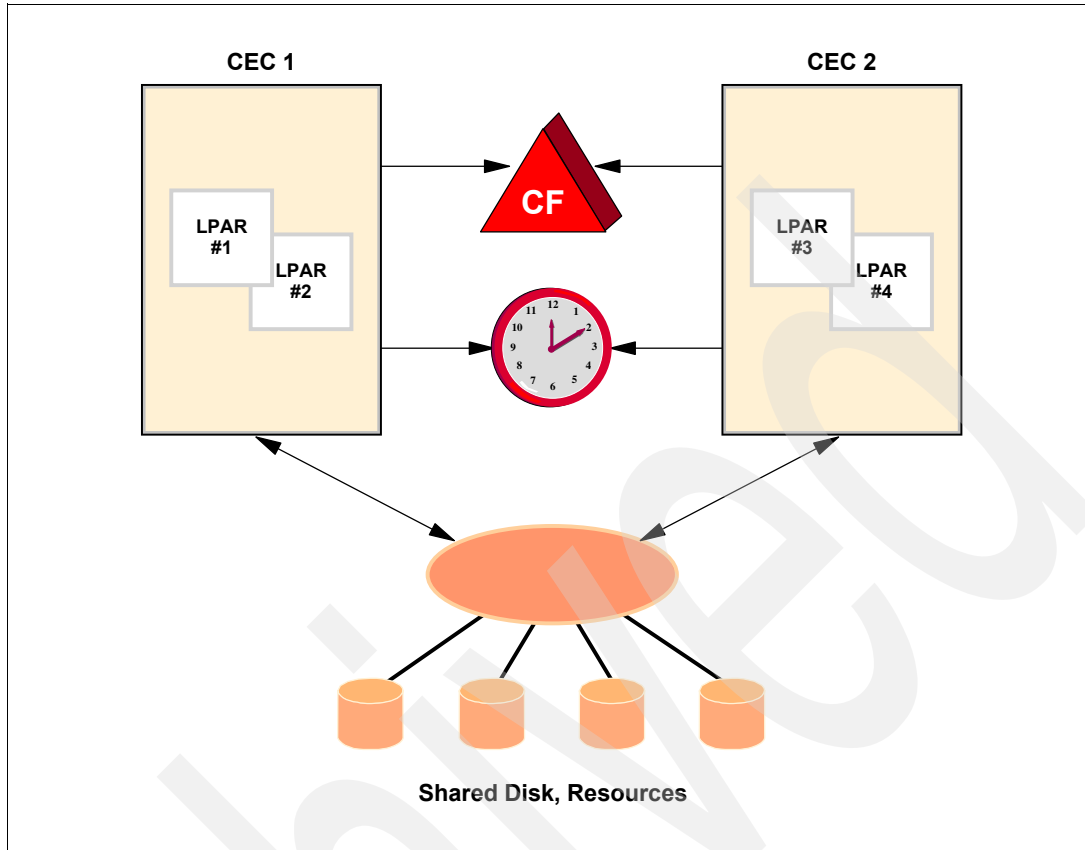


Figure 2-2 Shared data, shared access and shared resources

Benefits of a System z Parallel Sysplex

The Parallel Sysplex is a way of managing this multisystem environment, providing benefits that include:

- ▶ Continuous availability
- ▶ Capacity
- ▶ Dynamic workload balancing
- ▶ Ease of use
- ▶ Single system image
- ▶ Nondisruptive growth
- ▶ Application compatibility

Within a Parallel Sysplex cluster it is possible to construct a parallel processing environment, and by providing redundancy in the hardware configuration, a significant reduction in the total single points of failure is possible (see Figure 2-3). Even though they work together and present a single image, the nodes in a Parallel Sysplex cluster remain individual systems, making installation, operation, and maintenance nondisruptive. The Parallel Sysplex environment can scale nearly linearly from two to 32 systems. This can be a mix of any servers that support the Parallel Sysplex environment. Just as work can be dynamically distributed across the individual processors within a single SMP server, so, too, can work be directed to any node in a Parallel Sysplex cluster having the available capacity.

The Parallel Sysplex solution satisfies a major client requirement for continuous 24-hours-a-day, 7-days-a-week, 365-days-a-year (24x7x365) availability.

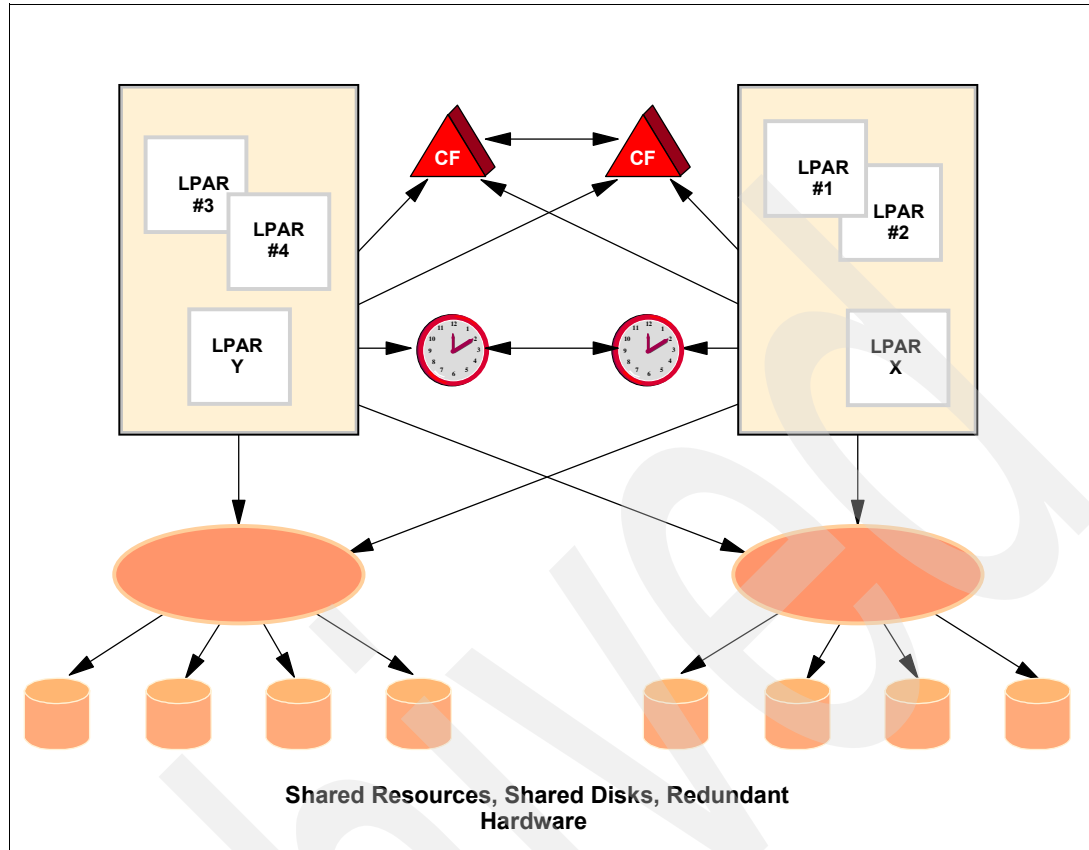


Figure 2-3 Shared data, shared access, shared resources, and redundant hardware

In summary, the System z Parallel Sysplex is a key Business Continuity technology in its own right. Parallel Sysplex is also a key technology for support of various GDPS family solution offerings

2.1.2 Server Time Protocol (STP)

The 9037 Sysplex Timers have been a functional and reliable method of coordinating time between CECs in a Sysplex. However, there were certain limitations, particularly in multisite Parallel Sysplex implementations. They require a separate footprint, separate links, and are limited in distance.

The Server Time Protocol can reduce or eliminate these limitations by serving as a replacement for the Sysplex Timer®.

Details

The Server Time Protocol runs from a function in the Licensed Internal Code (LIC) of the server in order to present a single view of time to all connected processor resources and the system manager. All messages from the server time protocol are passed across the ISC-3, ICB-3, or ICB-4 links that are already used for Coupling Facility Links.

The Server Time Protocol supports a *Coordinated Timing Network (CTN)*. The CTN is a collection of servers and coupling facilities that are synchronized to a time value called *Coordinated Server Time*.

The simplified structure is shown in Figure 2-4.

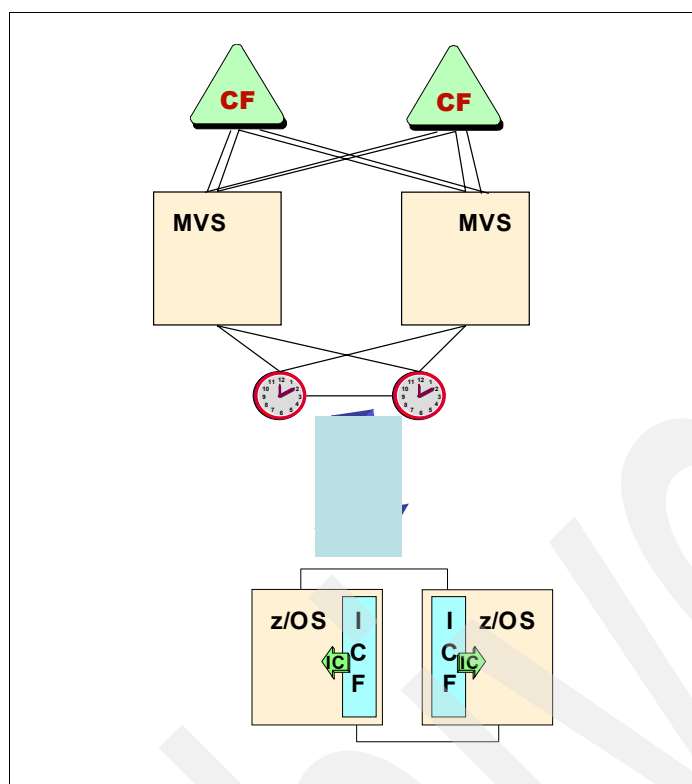


Figure 2-4 Greatly simplified infrastructure of a STP environment

STP and CTN support the following functions:

- ▶ Initialize the time manually, including Time Zone offset, Daylight Saving Time (DST) offset, Leap Seconds offset.
- ▶ Initialize the time by dialing out to a time service so that coordinated server time can be set within 100ms of an international time standard such as Coordinated Universal Time (UTC).
- ▶ Schedule periodic dial-outs so that coordinated server time can be gradually steered to an international tie standard.
- ▶ Adjust coordinated server time by up to +/- 60 seconds. This improves upon the 9037 Sysplex Timer's capability of adjusting by up to +/- 4.999 seconds
- ▶ Schedule changes to the offsets, such as DST. This was not possible with the Sysplex Timer.

2.1.3 GDPS solution offerings: Introduction

GDPS is a family of offerings, for single site or multi-site application availability solutions, which can manage the remote copy configuration and storage systems, automate z/OS operational tasks, manage and automate planned reconfigurations, and do failure recovery from a single point of control.

GDPS is an integrated end-to-end solution composed of software automation, software, servers and storage, networking, and IBM Global Technology Services to configure and deploy the solution, as shown in Figure 2-5 (the GDPS solution has components in the areas denoted by dark shading).

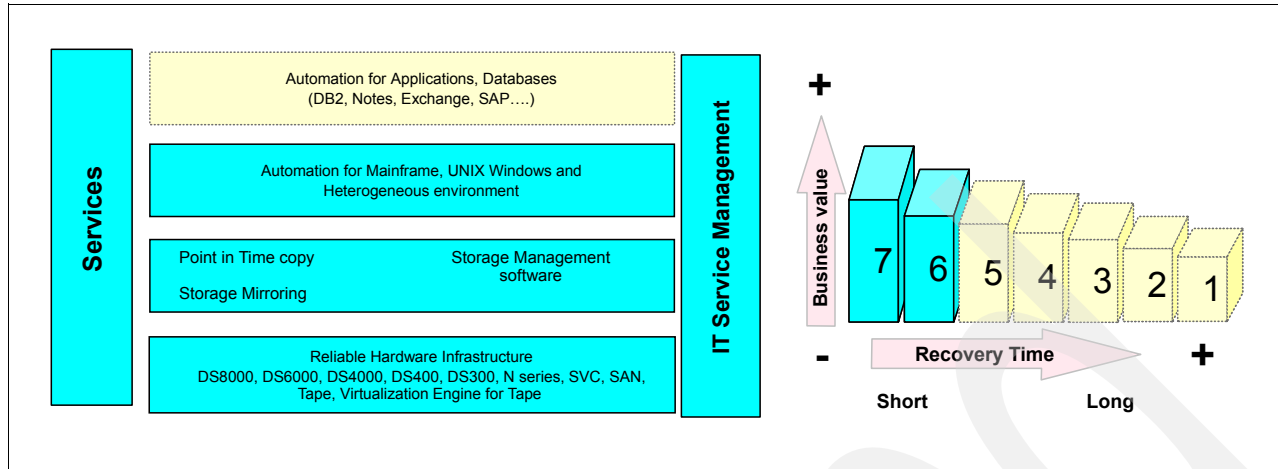


Figure 2-5 The positioning of the GDPS family of solutions

GDPS is a highly reliable, highly automated IT infrastructure solution with a System z-based control point, which can provide very robust application availability. GDPS is independent of the transaction managers (such as CICS® TS, IMS™, and WebSphere®) or database managers (such as DB2®, IMS, and VSAM) being used, and is enabled by means of key IBM technologies and architectures, including:

- ▶ System z processor technology and System z Parallel Sysplex
- ▶ Tivoli NetView® for z/OS
- ▶ Tivoli System Automation
- ▶ IBM System Storage DS6000™, DS8000™, and ESS
- ▶ TS7740 Grid
- ▶ Optical Dense or Coarse Wavelength Division Multiplexer (DWDM or CWDM)
- ▶ Metro Mirror architecture for GDPS/PPRC and GDPS/PPRC HyperSwap Manager
- ▶ z/OS Global Mirror architecture for GDPS/XRC
- ▶ Global Mirror for GDPS/Global Mirror
- ▶ TS7700 Grid architecture

Depending on the client requirements, only the required components from the foregoing list are selected.

GDPS supports both the synchronous (Metro Mirror) as well as the asynchronous (z/OS Global Mirror and Global Mirror) forms of remote copy. GDPS also supports TS7700 Grid for remote copying tape data in GDPS/PPRC and GDPS/XRC environments. The GDPS solution is a non-proprietary solution, working with IBM as well as other disk vendors, as long as the vendor meets the specific functions of the Metro Mirror, Global Mirror, and z/OS Global Mirror architectures required to support GDPS functions.

GDPS automation manages and protects IT services by handling planned and unplanned exception conditions. Depending on the GDPS configuration selected, availability can be storage resiliency only, or can provide near-continuous application and data availability. Regardless of the specific configuration variation, GDPS provides a System z-based Business Continuity automation and integration solution to manage both planned and unplanned exception conditions.

GDPS solution offerings

Figure 2-6 shows some of the many possible components supported in a GDPS configuration. GDPS is configured to meet the specific client requirements; not all components shown are necessary in all GDPS offerings or GDPS solutions.

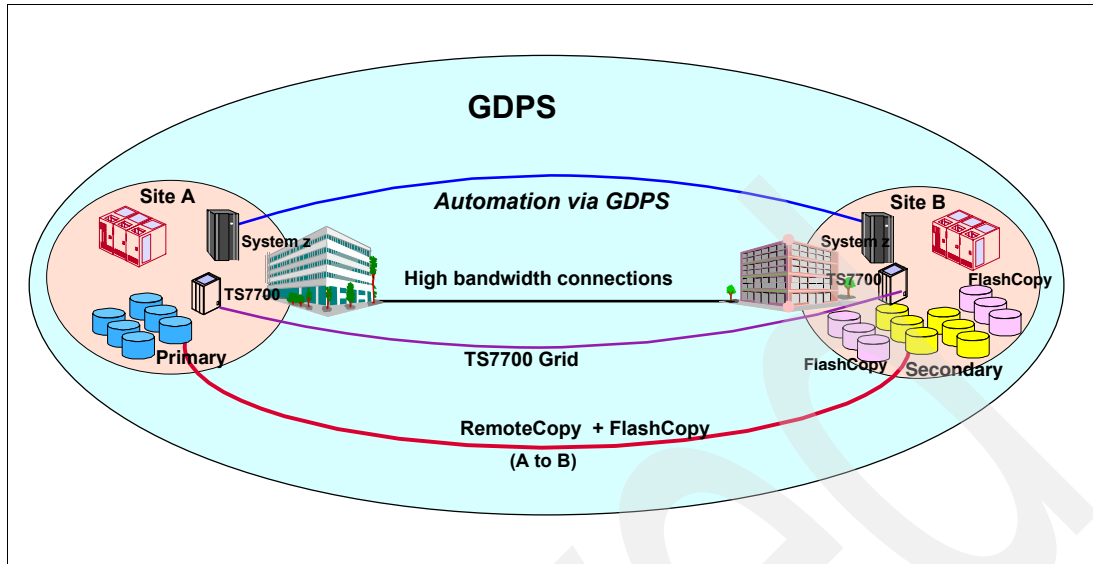


Figure 2-6 Many components are integrated within a GDPS solution

The GDPS family of System z Business Continuity solutions consists of two major offering categories, and each category has several sub-offerings. Each GDPS solution is delivered through IBM Global Technology Services, specifically tailored to fit a specific set of client requirements in the areas of recovery, budget, physical distance, infrastructure, and other factors. All forms of GDPS are application independent and, as such, can be used with any of the applications within their particular environments.

The two major categories of GDPS technologies are:

- ▶ Synchronous GDPS technologies, based on IBM Metro Mirror (formerly known as PPRC)
- ▶ Asynchronous GDPS technologies, based on IBM z/OS Global Mirror (formerly known as Extended Remote Copy (XRC), or IBM Global Mirror

The following sub-offerings are included within these two major GDPS categories.

Synchronous GDPS:

Synchronous GDPS includes:

- ▶ GDPS/PPRC, a fully automated and integrated technology for near continuous availability and disaster recovery
- ▶ GDPS/PPRC HyperSwap Manager, a subset of GDPS/PPRC which is used for near continuous availability and disaster recovery, but with more limitations than GDPS/PPRC
- ▶ RCMF/PPRC, a remote copy management technology for Metro Mirror

Asynchronous GDPS:

Asynchronous GDPS includes:

- ▶ GDPS/XRC, an unlimited distance, System z only technology for fully automated and integrated disaster recovery using z/OS Global Mirror (zGM) to send the data to the remote location.
- ▶ GDPS/Global Mirror, an unlimited distance technology which automates the mirroring environment for System z and open data while providing a fully automated and integrated disaster recovery.
- ▶ RCMF/XRC, a remote copy management solution for zGM

Some of these GDPS offerings can also be combined into three site GDPS solutions, providing near continuous availability within synchronous distances, as well as disaster recovery at unlimited distances.

GDPS/PPRC introduction

GDPS/PPRC is designed with the attributes of a *continuous availability and disaster recovery* technology and is based on the Metro Mirror disk replication technology.

GDPS/PPRC complements a multisite Parallel Sysplex implementation by providing a single, automated solution to dynamically manage storage system mirroring (disk and tape), processors, and network resources designed to help a business to attain continuous availability with no or minimal data loss in both planned and unplanned outages. It is designed to minimize and potentially eliminate the impact of any failure including disasters, or a planned outage.

GDPS/PPRC has a number of tools at its disposal in order to guarantee consistency of data and enable recovery at its alternate location including the FREEZE function and HyperSwap.

GDPS/XRC introduction

GDPS/XRC has the attributes of a *disaster recovery* technology and is based on. z/OS Global Mirror (zGM).

In GDPS/XRC, the production systems located in site 1 can be a single system, multiple systems sharing disk, or a base or Parallel Sysplex cluster. GDPS/XRC provides a single, automated interface, designed to dynamically manage storage system mirroring (disk and tape) to allow a business to attain a *consistent* disaster recovery through a single interface and with minimal human interaction.

GDPS/Global Mirror introduction

GDPS/Global Mirror is a disaster recovery technology and is based on the Global Mirror remote mirror and copy technology.

With GDPS/Global Mirror, the production systems located in site 1 can be a single system, multiple systems sharing disk, or a base or parallel Sysplex cluster. GDPS/GM provides a single automated interface with the Disk Systems and recovery processors in order to provide support for mirroring both System z and other open systems data through the Global Mirror technology. Additionally, GDPS/Global Mirror provides a single interface for performing recovery tasks in the System z environment including automation of Capacity Backup and System z server reconfiguration.

GDPS is delivered through Global Technology Services

The various GDPS offerings are not products; they are delivered as Global Technology Services offerings. GDPS is an end-to-end solution in which Global Technology Services tailors and installs the specific combination of components, integrated within the client's environment. GDPS integration work, including education, is done by Global Technology Services (GTS), as well as the planning and installation of the code. This assures that the GDPS solution provides the intended value to all parts of the business and IT processes.

Introduction summary

The differences of the various GDPS implementations implementation should be carefully reviewed before deciding which implementation best fits your business objectives.

Note: The original GDPS technologies still use the original IBM Advanced Copy Services names for disk mirroring:

- ▶ GDPS/PPRC refers to the use of IBM Metro Mirror remote mirror and copy, which was previously known as PPRC.
- ▶ GDPS/XRC refers to the use of z/OS Global Mirror remote mirror and copy, which was previously known as XRC.

In the following sections, we first describe some of the end-to-end server, storage, software, and automation integration requirements that are solved by the GDPS technologies.

We then provide more information about the various GDPS offerings.

Finally, we discuss in detail some of the major technology components of various GDPS configurations.

Requirement for data consistency

As noted in Chapter 6 “Planning for Business Continuity in a heterogeneous IT environment” in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547 data consistency across all secondary volumes, spread across any number of storage systems, is essential in providing data integrity and the ability to do a normal database restart in the event of a disaster.

The mechanism for data consistency in GDPS/XRC and GDPS/GM lies within their respective remote copy technologies, which are described in Chapter 6 “Planning for Business Continuity in a heterogeneous IT environment” in the *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547. While the requirement for automation in Metro Mirror environments is also discussed in that chapter, it is important to understand how the GDPS/PPRC *FREEZE function* behaves.

GDPS/PPRC uses a combination of storage system and Parallel Sysplex technology triggers to capture, at the first indication of a potential disaster, a data consistent secondary site (site 2) copy of the data, using the IBM Metro Mirror FREEZE function. The fact that this function is a combination of Parallel Sysplex technology and Disk System technology is a key piece of the GDPS/PPRC value.

Most similar functions look for an SNMP message from the Disk System indicating that an error has prevented a volume from mirroring and react to that message. In GDPS/PPRC, the dedicated K-System is able to see a variety of potential errors from the FICON® channel perspective and can trigger a FREEZE based on any of them.

Note: A K-System is the LPAR which runs the code for GDPS/PPRC and GDPS/XRC (and their derivatives). K is the first letter in the Swedish word for Control and is, thus, the Controlling System.

GDPS/Global Mirror makes use of a K-System but also makes use of an R (for Recovery) System in the Recovery site.

The FREEZE function, initiated by automated procedures, is designed to freeze the image of the secondary data at the very first sign of a disaster, even before the DBMS is made aware of I/O errors. This can prevent the logical contamination of the secondary copy of data that would occur if any storage system mirroring were to continue after a failure that prevents some, but not all secondary volumes from being updated.

Providing data consistency enables the secondary copy of data to perform normal restarts (instead of performing database manager recovery actions). This is the essential design element of GDPS in helping to minimize the time to recover the critical workload, in the event of a disaster in site 1.

GDPS systems

GDPS consists of production systems and controlling systems. The production systems execute the mission critical workload. There must be sufficient processing resource capacity (typically in site 2), such as processor capacity, main storage, and channel paths available that can quickly be brought online to restart a system's or site's critical workload (typically by terminating one or more systems executing expendable [non-critical] work and acquiring its processing resource).

The Capacity backup (CBU) feature, available on the System z server could provide additional processing power, which can help achieve cost savings. The CBU feature can increase capacity temporarily, when capacity is lost elsewhere in the enterprise. CBU adds Central Processors (CPs) to the available pool of processors and is activated only in an emergency. GDPS-CBU management automates the process of dynamically adding reserved Central Processors (CPs), thereby helping to minimize manual intervention and the potential for errors. Similarly, GDPS-CBU management can also automate the process of dynamically returning the reserved CPs when the temporary period has expired.

The controlling system coordinates GDPS processing. By convention, all GDPS functions are initiated and coordinated by the controlling system.

All GDPS systems run GDPS automation based upon Tivoli NetView for z/OS and Tivoli System Automation for z/OS. Each system can monitor the Sysplex cluster, Coupling Facilities, and storage systems and maintain GDPS status. GDPS automation can coexist with an enterprise's existing automation product.

2.1.4 GDPS/PPRC overview

GDPS/PPRC is designed to manage and protect IT services by handling planned and unplanned exception conditions, and maintaining data integrity across multiple volumes and storage systems. By managing both planned and unplanned exception conditions, GDPS/PPRC can help to maximize application availability and provide business continuity.

GDPS/PPRC can provide:

- ▶ Near Continuous Availability
- ▶ Near transparent D/R
- ▶ Recovery Time Objective (RTO) less than an hour
- ▶ Recovery Point Objective (RPO) of zero (optional)
- ▶ Protection against localized area disasters (distance between sites limited to 100 km fiber)

The GDPS/PPRC solution offering combines System z Parallel Sysplex capability and DS6000 or DS8000 Metro Mirror disk mirroring technology to provide a Business Continuity solution for IT infrastructures with System z at the core. GDPS/PPRC offers efficient workload management, system resource management, Business Continuity or Disaster Recovery for z/OS servers and open system data, and providing data consistency across all platforms using the Metro Mirror Consistency Group function.

The GDPS solution uses automation technology to provide end-to-end management of System z servers, disk mirroring, tape mirroring, and workload shutdown and startup. GDPS manages the infrastructure to minimize or eliminate the outage during a planned or unplanned site failure. Critical data is disk mirrored, and processing is automatically restarted at an alternate site in the event of a primary planned site shutdown or site failure.

GDPS/PPRC automation provides scalability to insure data integrity at a very high large number of volumes, across hundreds or thousands of Metro Mirror pairs.

Zero data loss

The GDPS/PPRC FREEZE function grants zero data loss at the hardware block level if combined with a “FREEZE and stop” or a HyperSwap (discussed in 2.1.5, “GDPS/PPRC HyperSwap Manager overview” on page 25). It is an easily misunderstood concept, however, and as such deserves a further review.

The GDPS/PPRC FREEZE arranges for zero data loss, again, at the hardware block level. What this means is that up to the point where an event was detected as a possible disaster, all data was mirrored. However, this FREEZE does not necessarily cause updates to halt at a “transaction boundary”. As a result, some transactions are probably incomplete at this stage. This is shown in Figure 2-7.

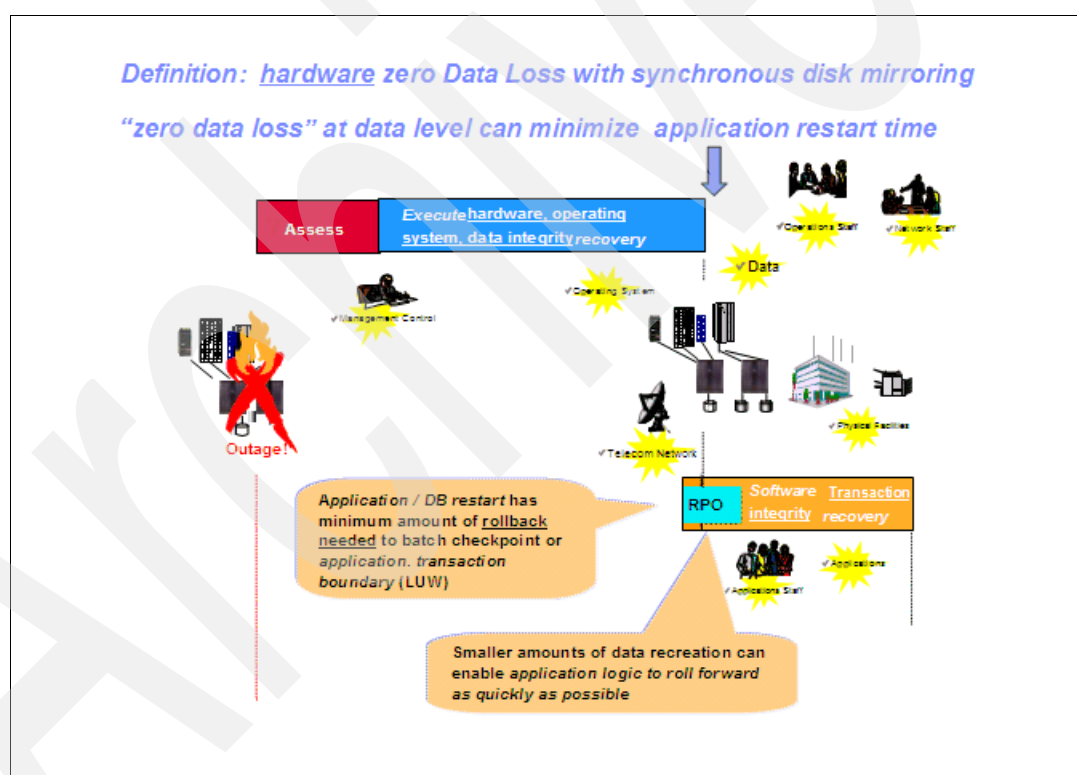


Figure 2-7 How hardware block level zero data loss works alongside DBMS transaction integrity

When the database restart occurs, the first thing that occurs in the DBSMS is that these incomplete transactions are backed out. As a result, all data is realigned on a transaction boundary. The next step might be a roll forward (as possible or as permitted by the DBMS) based on logger data, which might be more up to date than the data itself.

As we can see, GDPS/PPRC can maintain zero data loss at a hardware block level, however, this is not the same as a transaction level. Transaction level zero data loss is still possible in these environments, but ultimately, that level of recovery is up to the DBMS itself.

GDPS/PPRC topology

The physical topology of a GDPS/PPRC, shown in Figure 2-8, consists of a base or Parallel Sysplex cluster spread across two sites (known as site 1 and site 2) separated by up to 100km of fiber – approximately 62 miles – with one or more z/OS systems at each site.

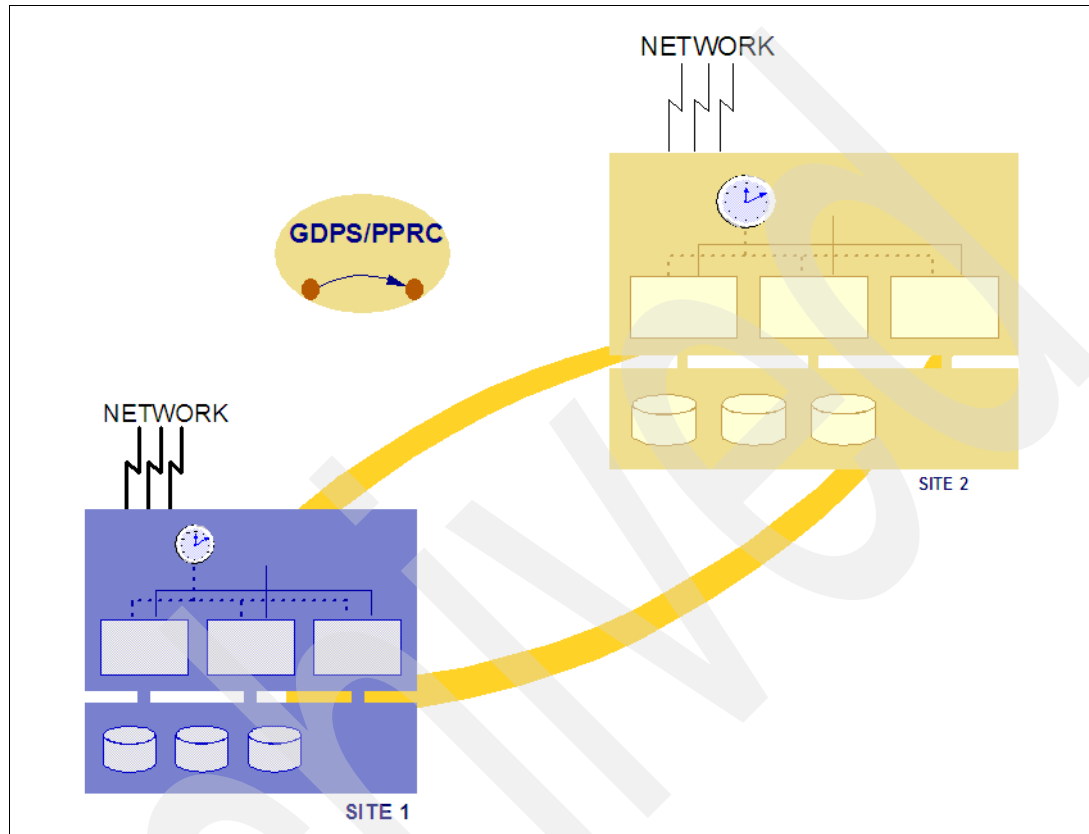


Figure 2-8 GDPS/PPRC Topology

The multisite Sysplex cluster must be configured with redundant hardware (for example, a Coupling Facility and a Sysplex Timer in each site) and the cross-site connections must be redundant. All critical data resides on storage systems in site 1 (the primary copy of data) and is mirrored to the storage systems in site 2 (the secondary copy of data) via Metro Mirror.

In order to maintain the 100km maximum distance for GDPS/PPRC, it is a requirement to maintain a distance of no more than 40 km between the Sysplex Timer CLO links, by means of using an alternate site between data centers or potentially placing both Sysplex Timers within the same data center's campus. The other alternative is the new Server Timer Protocol which puts the Sysplex Timer functionality within the System z server, and allows for a maximum native distance of 100km.

An immediate advantage of this extended distance is to potentially decrease the risk that the same disaster might affect both sites, thus permitting enterprises to recover production applications at another site. This should be balanced against the potential performance impacts that would be incurred if the intent is to share workload between the two sites, because longer distances are typically found in single site workload environments.

Near Continuous Availability of data with HyperSwap

Exclusive to GDPS in the Metro Mirror environment is HyperSwap. This function is designed to broaden the near continuous availability attributes of GDPS/PPRC by extending the

Parallel Sysplex redundancy to disk systems. The HyperSwap function can help significantly reduce the time required to switch to the secondary set of disks while keeping the z/OS systems active, together with their applications.

All current versions of the function exploit the Metro Mirror Failover/Failback (FO/FB) function. For planned reconfigurations, FO/FB can reduce the overall elapsed time to switch the disk systems, thereby reducing the time that applications might be unavailable to users.

This is demonstrated by the benchmark measurements discussed below in Table 2-1 on page 23.

For unplanned reconfigurations, Failover/Failback allows the secondary disks to be configured in the suspended state after the switch and record any updates made to the data. When the failure condition has been repaired, resynchronizing back to the original primary disks requires only the changed data to be copied, thus eliminating the requirement to perform a full copy of the data. The window during which critical data is left without Metro Mirror protection following an unplanned reconfiguration is thereby minimized.

As of GDPS 3.3, it is possible to add a user defined HyperSwap trigger based on IOS. This means that if I/O response times exceed a user defined threshold, a HyperSwap occurs.

GDPS/PPRC planned reconfiguration support

GDPS/PPRC planned reconfiguration support automates procedures performed by an operations center. These include standard actions to:

- ▶ Quiesce a system's workload and remove the system from the Parallel Sysplex cluster (for example, stop the system prior to a hardware change window).
- ▶ IPL a system (for example, start the system after a hardware change window).
- ▶ Quiesce a system's workload, remove the system from the Parallel Sysplex cluster, and re-IPL the system (for example, recycle a system to pick up software maintenance).

Standard actions can be initiated against a single system or a group of systems. With the introduction of HyperSwap, you now can perform disk maintenance and planned site maintenance without requiring applications to be quiesced. Additionally, GDPS/PPRC provides customizable scripting capability for user defined actions (for example, planned disk maintenance or planned site switch in which the workload is switched from processors in site 1 to processors in site 2).

All GDPS functions can be performed from a single point of control, which can help simplify system resource management. Panels are used to manage the entire remote copy configuration, rather than individual remote copy pairs. This includes the initialization and monitoring of the remote copy volume pairs based upon policy and performing routine operations on installed storage systems – disk and tape. GDPS can also perform standard operational tasks, and monitor systems in the event of unplanned outages.

The Planned HyperSwap function provides the ability to transparently switch all primary disk systems with the secondary disk systems for planned reconfigurations. During a planned reconfiguration, HyperSwap can perform disk configuration maintenance and planned site maintenance without requiring any applications to be quiesced. Large configurations can be supported, as HyperSwap provides the capacity and capability to swap large number of disk devices very quickly. The important ability to re-synchronize incremental disk data changes, in both directions, between primary and secondary disks is provided as part of this function.

Benchmark figures: HyperSwap for planned reconfiguration

Table 2-1 gives some IBM benchmark results from an IBM test facility for GDPS/PPRC with HyperSwap for planned and unplanned reconfigurations.

Table 2-1 Benchmark measurements using HyperSwap for reconfiguration

Configuration	Planned reconfiguration switch time with failover/failback	Unplanned reconfiguration switch time with failover/failback
6545 volume pairs 19.6 TB, 46 LSSs	15 seconds	13 seconds

** 3390-9 device type volumes

Note 1: HyperSwap prerequisites are described in 2.2.12, "GDPS prerequisites" on page 65

For further information about these performance benchmarks, you can contact your IBM GDPS representative.

GDPS/PPRC unplanned reconfiguration support

GDPS/PPRC unplanned reconfiguration support not only can automate procedures to handle site failures, but can also help minimize the impact and potentially mask a z/OS system, processor, Coupling Facility, disk or tape failure, based upon GDPS/PPRC policy. If a z/OS system fails, the failed system and workload can be automatically restarted. If a processor fails, the failed systems and their workloads can be restarted on other processors.

The Unplanned HyperSwap function can transparently switch to use secondary disk systems that contain mirrored data consistent with the primary data, in the event of unplanned outages of the primary disk systems or a failure of the site containing the primary disk systems (site 1).

With Unplanned HyperSwap support:

- ▶ Production systems can remain active during a disk system failure. Disk system failures no longer constitute a single point of failure for an entire Sysplex.
- ▶ Production systems can remain active during a failure of the site containing the primary disk systems (site 1), if applications are cloned and exploiting data sharing across the 2 sites. Even though the workload in site 2 must be restarted, an improvement in the Recovery Time Objective (RTO) is accomplished.

Benchmark figures: HyperSwap for Unplanned Reconfiguration

An unplanned disk reconfiguration test using HyperSwap with IBM Metro Mirror Failover/Failback, conducted at the GDPS solution center, demonstrated that the user impact time was only 15 seconds to swap a configuration of 2,900 volumes of IBM Enterprise Storage Server® disks while keeping the applications available, compared to typical results of 30-60 minutes without HyperSwap.

In addition to this timing improvement, the Failover/Failback capability is also designed to significantly reduce the elapsed time to resynchronize and switch back, because only the incremental changed data (instead of the entire disk) has to be copied back during the resynchronization process. This can save significant time and network resources.

GDPS/PPRC support for heterogeneous environments

GDPS/PPRC has been extended to support heterogeneous operating systems platforms, including other System z operating systems such as zLinux, and Open Systems operating systems.

Management of System z operating systems

In addition to managing images within the base or Parallel Sysplex cluster, GDPS can now also manage a client's other System z production operating systems – these include z/OS, Linux® for System z, z/VM®, and VSE/ESA™. The operating systems have to run on servers

that are connected to the same Hardware Management Console (HMC) LAN as the Parallel Sysplex cluster images. For example, if the volumes associated with the Linux images are mirrored using Metro Mirror, GDPS can restart these images as part of a planned or unplanned site reconfiguration. The Linux for System z images can either run as a logical partition (LPAR) or as a guest under z/VM.

GDPS/PPRC management for open systems LUNs

GDPS/PPRC technology has been extended to manage a heterogeneous environment of z/OS and Open Systems data. If installations share their disk systems between the z/OS and Open Systems platforms, GDPS/PPRC can manage the Metro Mirror and Metro Mirror FREEZE for open systems storage, thus providing a common Consistency Group and data consistency across both z/OS and open systems data. This allows GDPS to be a single point of control to manage business resiliency across multiple tiers in the infrastructure, improving cross-platform system management and business processes.

GDPS/PPRC multi-platform resiliency for System z

GDPS/PPRC *multi-platform resiliency for System z* is especially valuable if data and storage systems are shared between z/OS and z/VM Linux guests on System z – for example, an application server running on Linux on System z and a database server running on z/OS.

z/VM 5.1 and above supports the HyperSwap function, so that the virtual device (for example, associated with System z Linux guests running under z/VM) associated with one real disk, can be swapped transparently to another disk. HyperSwap can be used to switch to secondary disk systems mirrored by Metro Mirror. If there is a hard failure of a storage device, GDPS coordinates the HyperSwap with z/OS for continuous availability spanning the multi-tiered application. For site failures, GDPS invokes the FREEZE function for data consistency and rapid application restart, without the necessity of data recovery. HyperSwap can also be helpful in data migration scenarios to allow applications to migrate to new disk volumes without requiring them to be quiesced.

GDPS/PPRC provides the reconfiguration capabilities for the Linux on System z servers and data in the same manner as for z/OS systems and data. To support planned and unplanned outages, GDPS provides the recovery actions such as the following examples:

- ▶ Re-IPL in place of failing operating system images
- ▶ Site takeover/failover of a complete production site
- ▶ Coordinated planned and unplanned HyperSwap of disk systems, transparent to the operating system images and applications using the disks.
- ▶ Linux node or cluster failures
- ▶ Transparent disk maintenance or failure recovery with HyperSwap across z/OS and Linux applications
- ▶ Data consistency with FREEZE functions across z/OS and Linux

GDPS/PPRC summary

GDPS/PPRC can provide a solution for many categories of System z clients, including (but not limited to):

- ▶ Clients who can tolerate an acceptable level of synchronous disk mirroring performance impact (typically an alternate site at metropolitan distance, can be up to 100 KM)
- ▶ Clients that require as close to near zero data loss as possible
- ▶ Clients that desire a fully automated solution that covers the System z servers, System z Coupling Facilities, System z reconfiguration, and so on, in addition to the disk recovery

A given GDPS/PPRC implementation can feature:

- ▶ A highly automated, repeatable site takeover managed by GDPS/PPRC automation
- ▶ High performance synchronous remote copy
- ▶ Hardware data loss zero or near zero
- ▶ Data consistency and data integrity assured to insure fast, repeatable database restart
- ▶ Support for metropolitan distances
- ▶ Automation of System z Capacity Back Up (CUB)
- ▶ Single point of control for disk mirroring and recovery
- ▶ Support for consistency across both open systems and system z data
- ▶ Support of the GDPS HyperSwap functionality

2.1.5 GDPS/PPRC HyperSwap Manager overview

The GDPS/PPRC HyperSwap Manager solution is a subset of the full GDPS/PPRC solution, designed to provide an affordable entry point to the full family of GDPS/PPRC offerings.

GDPS/PPRC HyperSwap Manager (GDPS/PPRC HM) expands System z Business Resiliency by providing a single-site near continuous availability solution as well as a multi-site entry-level disaster recovery solution.

Within a single site, GDPS/PPRC HyperSwap Manager extends Parallel Sysplex availability to disk systems by masking planned and unplanned disk outages caused by disk maintenance and disk failures. It also provides management of the data replication environment and automates switching between the two copies of the data without requiring an IPL of the System z server, thereby providing near-continuous access to data. Figure 2-9 shows an example of a GDPS/PPRC HM configuration. As you can see, the Unit Control Block switches from pointing to one disk system to the other when the swap occurs.

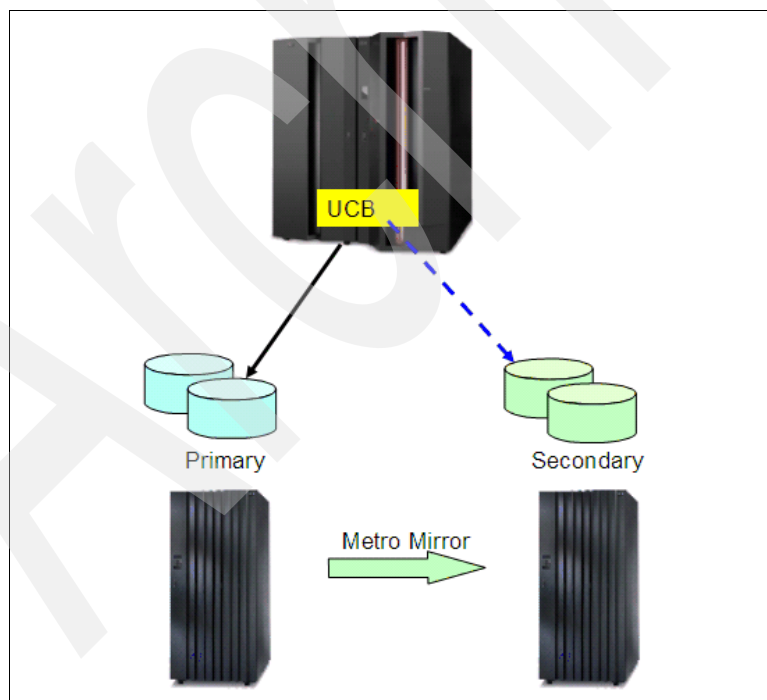


Figure 2-9 GDPS/PPRC HyperSwap Manager

In the multisite environment, GDPS/PPRC HyperSwap Manager provides an effective entry-level disaster recovery offering for System z clients who have to guarantee data consistency but are willing to manually restore operations in the case of a site failure. In

multi-site environments where the secondary disk volumes are within acceptable distances (for application performance purposes), the HyperSwap occurs as normal in planned or unplanned disk outages.

Note: The GDPS HyperSwap Manager offering features specially priced, limited function Tivoli System Automation and NetView software products, thus satisfying the GDPS software automation prerequisites with a lower initial price and a cost-effective initial entry point to the GDPS family of GDPS offerings.

Users who already have the full function Tivoli System Automation and NetView software products, can continue to use them as the prerequisites for GDPS/PPRC HyperSwap Manager.

For more information, see 2.2.12, “GDPS prerequisites” on page 65.

In addition, a client can migrate to the full function GDPS/PPRC capability across multiple sites as business requirements demand shorter Recovery Time Objectives provided by a second site. The initial investment in GDPS/PPRC HM is protected when moving to full-function GDPS/PPRC by leveraging existing GDPS/PPRC HM implementation and skills.

Note: The limited function Tivoli System Automation and NetView software must be upgraded to the full function versions when migrating from an entry-level GDPS HyperSwap Manager to full GDPS/PPRC.

GDPS/PPRC HM simplifies the control and management of the Metro Mirror environment for both z/OS and Open Systems data. This reduces storage management costs while reducing the time required for remote copy implementation.

GDPS/PPRC HM supports FlashCopy in the NOCOPY mode only. GDPS/PPRC HM can be set up to automatically take a FlashCopy of the secondary disks before resynchronizing the primary and secondary disks following a Metro Mirror suspension event, ensuring a consistent set of disks are preserved should there be a disaster during the re-synch operation.

Near Continuous Availability within a single site

A Parallel Sysplex environment has been designed to reduce outages by replicating hardware, operating systems and application components. In spite of this redundancy having only one copy on the data is an exposure. GDPS/PPRC HM provides Continuous Availability of data by masking disk outages caused by disk maintenance and failures. For example, if normal processing is suddenly interrupted when one of the disk systems experiences a hard failure, thanks to GDPS the applications are masked from this error because GDPS detects the failure and autonomically invokes HyperSwap. The production systems continue using data from the mirrored secondary volumes. Disk maintenance can also be similarly performed without application impact by executing HyperSwap command.

Near Continuous Availability and D/R solution at metro distances

In addition to the single site capabilities, in a two site configuration GDPS/PPRC HyperSwap Manager provides an entry-level disaster recovery capability at the recovery site. GDPS/PPRC HM uses the FREEZE function described in “Requirement for data consistency” on page 18. The FREEZE function is designed to provide a consistent copy of data at the recovery site from which production applications can be restarted. The ability to simply restart applications helps eliminate the requirement for lengthy database recovery actions.

Automation to stop and restart the operating system images available with the full-function GDPS/PPRC is not included with GDPS/PPRC HM.

GDPS/PPRC HyperSwap Manager summary

GDPS/PPRC HM manages the disk mirroring and transparent switching to the secondary disk. When a HyperSwap trigger is detected by the HyperSwap code, it stops the I/O and switches to the secondary devices, and resumes the I/O before the application times out. The intended result is to mask the applications from an outage.

GDPS/PPRC HM manages the disk data for fast recovery, insuring a data consistent copy of the data at the remote site, suitable for database restart.

The GDPS/PPRC HyperSwap function can help significantly reduce the time required to transparently switch disks between primary and secondary disk systems. GDPS/PPRC HM function can be controlled by automation, allowing all aspects of the disk system switch to be controlled via GDPS.

The GDPS Open LUN Management capability is available with GDPS/PPRC HM, supporting open systems servers and their data, to be mirrored with data consistency maintained across platforms.

Note: The HyperSwap function is only available on the System z platform. The GDPS/PPRC HM provides data consistency on all other platforms, however, restart of open systems platform applications is required.

GDPS/PPRC HM (see Figure 2-9 on page 25) is intended for:

- ▶ Clients with a Parallel Sysplex in one site, and wanting to extend Parallel Sysplex availability to the disk systems
- ▶ Clients that require an entry level support for disaster recovery capabilities and Continuous Availability, and might desire the future ability to upgrade to the full GDPS/PPRC
- ▶ Clients that can tolerate an acceptable amount of performance impact due to the synchronous type of mirroring

The main highlights of the GDPS/PPRC HM solution are that it:

- ▶ Supports HyperSwap
- ▶ Supports high performance synchronous remote copy
- ▶ Supports zero or near zero data loss
- ▶ Provides time consistent data at the recovery site
- ▶ Provides a single point of control for disk mirroring functions
- ▶ Offers entry point pricing
- ▶ Is positioned to upgrade to full GDPS/PPRC at a later date

2.1.6 RCMF/PPRC overview

Remote Copy Management Facility / PPRC (RCMF/PPRC) is the name given to a subset of the GDPS/PPRC offerings. RCMF/PPRC includes the storage interface management functions only.

RCMF/PPRC provides panels and code that execute under NetView and provides an operator interface for easier management of a remote copy configuration, in setup, initialization, and any planned outage operational mode. This provides benefits for businesses looking to improve their management of Metro Mirror for normal running circumstances.

Note that RCMF/PPRC does not provide a monitoring capability, and is not designed to notify the operator of an error in the remote copy configuration. Thus, RCMF does not support the FREEZE function for GDPS/PPRC, FlashCopy automation, or the unplanned outage functions available through the other versions of GDPS.

RCMF/PPRC is positioned as a remote copy management control tool, designed to make much easier the task for operators to stop and start remote copy sessions.

RCMF/PPRC highlights

The main highlights of RCMF/PPRC are that it:

- ▶ Offers a central point of control - full screen
- ▶ Has Metro Mirror configuration awareness
- ▶ Provides a functional, tree-structured interface
- ▶ Does not require TSO commands
- ▶ Can initialize and maintain a Remote Copy configuration
- ▶ Provides single key stroke function invocation
- ▶ Can initiate functions per pair, disk system, or all
- ▶ Automatically establishes target configuration at system startup
- ▶ Supports adding, moving, removing pairs, systems and links
- ▶ Does not manage unplanned outage secondary data consistency
- ▶ Drives Peer-to-Peer/Dynamic Address Switching (P/DAS)
- ▶ Offers user-initiated status and exception reporting
- ▶ Can run as a NetView application - System Automation not required

2.1.7 GDPS/XRC overview

IBM z/OS Global Mirror (zGM) — formerly Extended Remote Copy (XRC) — is a combined hardware and z/OS software asynchronous remote copy solution for System z data. z/OS Global Mirror provides premium levels of scalability, reaching into the tens of thousands of z/OS volumes (including z Linux volumes). Consistency of the data is maintained via the Consistency Group function within the System Data Mover.

GDPS/XRC includes automation to manage z/OS Global Mirror remote copy pairs and automates the process of recovering the production environment with limited manual intervention, including invocation of CBU (Capacity Backup), thus providing significant value in reducing the duration of the recovery window and requiring less operator interaction.

GDPS/XRC has the following attributes:

- ▶ Disaster recovery solution
- ▶ RTO between an hour to two hours
- ▶ RPO of seconds, typically 3-5 seconds
- ▶ Unlimited distance between sites
- ▶ Minimal application performance impact

GDPS/XRC topology

The GDPS/XRC physical topology, shown in Figure 2-10, consists of production systems in site 1. The production systems could be a single system, multiple systems sharing disk, or a base or Parallel Sysplex cluster¹. Site 2 (the recovery site) can be located at a virtually unlimited distance from site 1 (the production site).

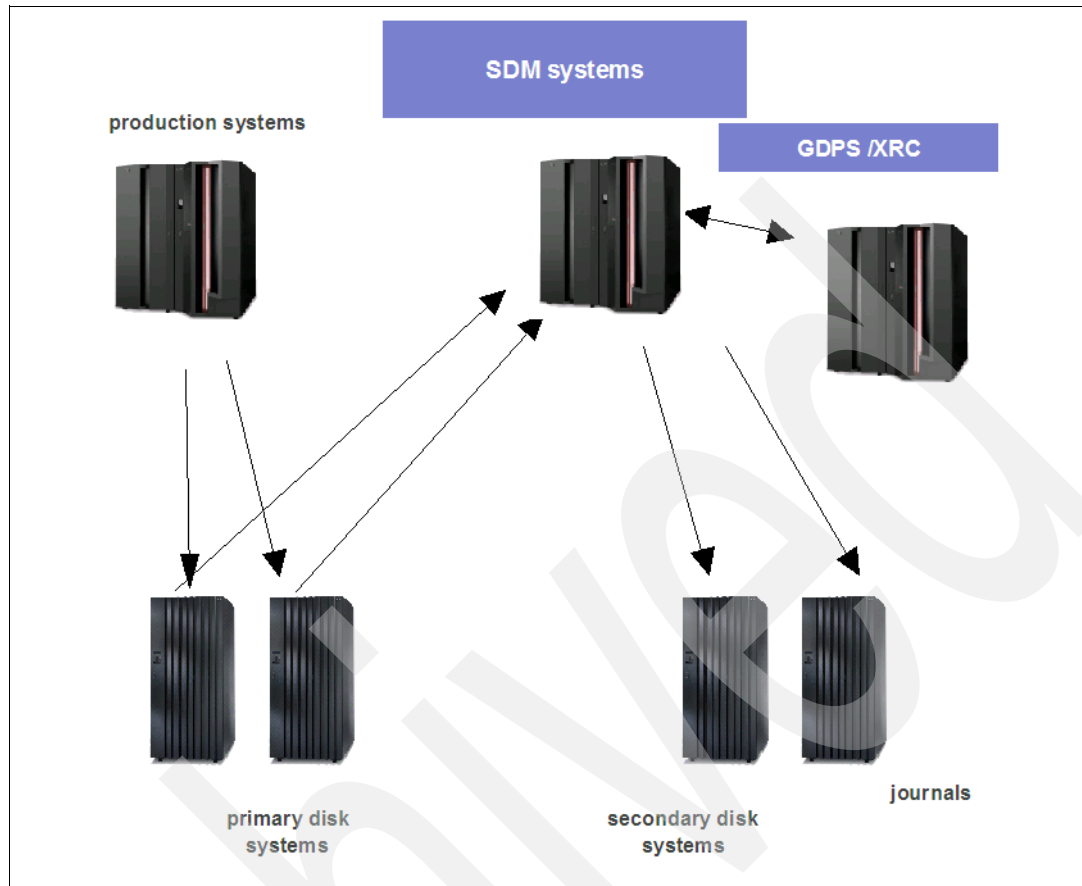


Figure 2-10 GDPS/XRC Topology

During normal operations, the System Data Movers (one or more) execute in site 2 and are in a Base Sysplex with the GDPS controlling system. All critical data resides on storage systems in site 1 (the primary copy of data) and is mirrored to the storage systems in site 2 (the secondary copy of data) via z/OS Global Mirror (zGM) asynchronous remote copy.

GDPS/XRC planned and unplanned reconfiguration support

Support for planned and unplanned reconfiguration actions includes the ability to:

- ▶ Quiesce a system's workload and remove the system from the Parallel Sysplex cluster (for example, stop the system prior to a site switch)
- ▶ Reconfigure and IPL systems
- ▶ Quiesce a system's workload, remove the system from the Parallel Sysplex cluster, and re-IPL the systems

GDPS/XRC provides a highly automated solution, to minimize the dependency on key human skills being available at the remote site, to recover from a disaster. GDPS/XRC automates the recovery process of the production environment with minimal manual intervention, which can provide significant value in minimizing the duration of the recovery window.

Coupled System Data Mover support

GDPS/XRC supports the z/OS Global Mirror support for Coupled System Data Movers. This expands the capability of z/OS Global Mirror so that configurations with thousands of primary volumes can recover all their volumes to a consistent point in time. A single SDM can typically manage approximately 1000 to 2000 volume pairs (based on the write I/O rate); Coupled System Data Movers can provide the scalability required to support larger z/OS Global Mirror (zGM) configurations.

The first level of coupling occurs within an LPAR. Up to 13 SDMs can be clustered together within a single LPAR. On top of that, up to 14 of these LPARs can be Coupled. This gives support for even the largest z/OS mirroring environments.

Commands can be executed in parallel across multiple SDMs in a GDPS/XRC configuration, providing improved scalability. The parallelism is across multiple SDMs, provided there is only one SDM per z/OS image. If there are multiple SDMs per z/OS image, processing is sequential by SDM within the z/OS image.

GDPS/XRC summary

GDPS automation technology provides automation to manage the mirroring of a z/OS Global Mirror configuration, z/OS Global Mirror SDM-Sysplex, and the bring up of the application systems at the remote site. Critical data is mirrored using z/OS Global Mirror, and processing automatically restarted at an alternate site in the event of a primary planned site shutdown or site failure. With the GDPS/XRC solution, there is virtually no limit to the distance between the primary and alternate sites.

GDPS/XRC typically accommodates:

- ▶ Clients that cannot tolerate the performance impact of a synchronous remote copy technology, such as Metro Mirror.
- ▶ Clients that require more distance between their production and failover site than can be accommodated by GDPS/PPRC (which depends upon the underlying hardware).

These are the main highlights of the GDPS/XRC solution:

- ▶ Site failover managed by GDPS/XRC
- ▶ High performance asynchronous remote copy
- ▶ Data loss measured in seconds
- ▶ Highly scalable
- ▶ Designed for unlimited distances
- ▶ Time consistent data at the recovery site
- ▶ Automatic Capacity Backup (CBU) System z CPU activation
- ▶ Single point of control
- ▶ Supports System z data
- ▶ Supports selected Linux z/VM environments¹.

¹ Check with your distribution of Linux on System z to determine if it can be supported by GDPS/XRC. For more information, contact your IBM representative.

2.1.8 RCMF/XRC overview

Remote Copy Management Facility / XRC (RCMF/XRC) is a subset of GDPS/XRC that can manage the z/OS Global Mirror (zGM) remote copy configuration for planned and unplanned outages. RCMF/XRC is not intended to manage the system and workload environment. RCMF/XRC provides a central point of control of z/OS Global Mirror disk mirroring configurations where the environment does not require more than one SDM. Those environments that require multiple SDMs and coupled SDMs require the full GDPS/XRC.

RCMF/XRC solution highlights include:

- ▶ Functional full screen interface
- ▶ Capability to initialize and maintain z/OS Global Mirror Remote Copy configuration
- ▶ User initiated status collection and exception monitoring
- ▶ As a NetView application, does not require System Automation for z/OS

RCMF/XRC provides panels and code that execute under NetView, and provides an operator interface for easier management of a remote copy z/OS Global Mirror configuration, in setup, initialization, and any planned outage operational mode. This provides benefits for businesses looking to improve their management of XRC for normal running circumstances.

RCMF/XRC is positioned as a remote copy management control tool, designed to make much easier the task for operators to stop and start a single SDM z/OS Global Mirror remote copy session.

2.1.9 GDPS/Global Mirror (GDPS/GM) overview

IBM System Storage Global Mirror is an asynchronous mirroring solution that can replicate both System z and open systems data.

GDPS/GM provides a link into the System z environment in order to enhance the remote copy interface for more efficient use of mirroring with fewer opportunities for mistakes, and the automation and integration necessary to perform a complete disaster recovery with minimal human intervention.

GDPS/GM has the following attributes:

- ▶ Disaster recovery technology
- ▶ RTO between an hour to two hours
- ▶ RPO less than 60 seconds, typically 3-5 seconds
- ▶ Protection against localized or regional disasters (distance between sites is unlimited)
- ▶ Minimal remote copy performance impact
- ▶ Improved and supported interface for issuing remote copy commands
- ▶ Maintaining multiple Global Mirror sessions and multiple RPOs

GDPS/GM topology

The GDPS/GM physical topology, shown in Figure 2-10, consists of production systems in site 1. The production systems could be a single system, multiple systems sharing disk, or a base or Parallel Sysplex cluster. Site 2, (the recovery site) can be located at a virtually unlimited distance from site 1 (the production site) and, again, is not actually required to be a Parallel Sysplex cluster.

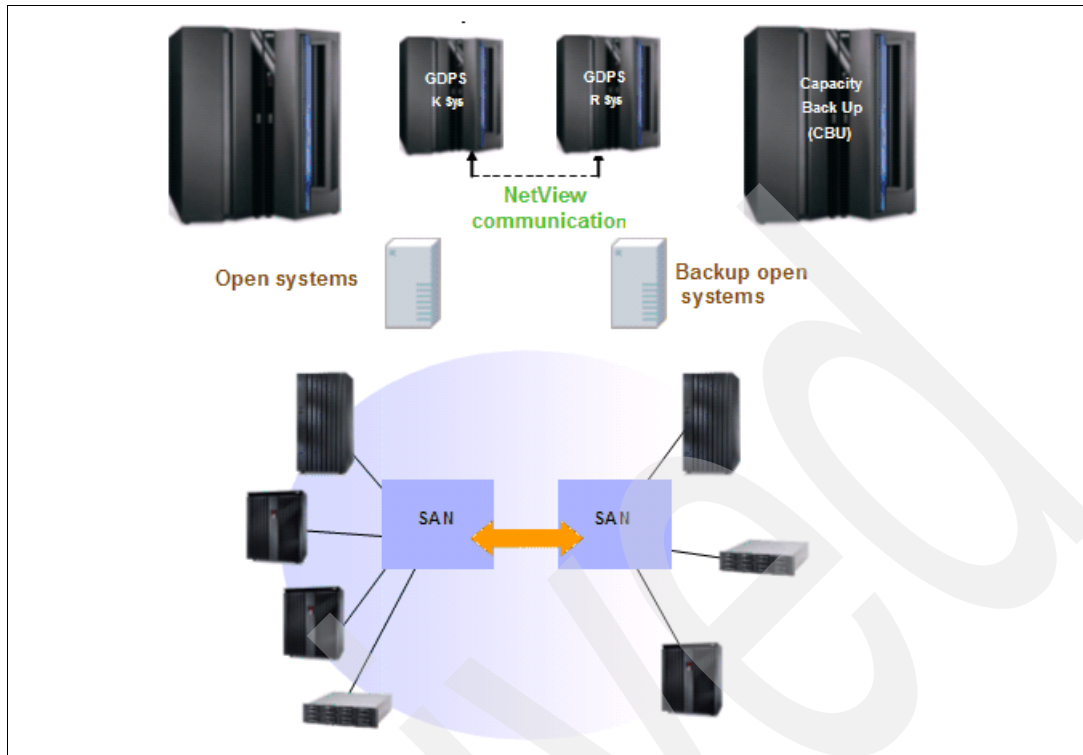


Figure 2-11 Topology of a GDPS/GM environment

As noted in Figure 2-11, like other GDPS environments, the GDPS/GM environment contains a K-System. The K-System, in this case, is contained within the production site and is the primary control point for interactions with the GDPS environment. Additionally, however, the GDPS/GM environment has a R-System.

The R-System communicates with the GDPS/GM K-System via NetView communications links. By doing this, GDPS/GM is able to verify the state of operations in each location and, if communications fail, GDPS/GM is able to notify the user of a potential problem.

Importantly, the R-System LPAR does not have to be a part of the recovery site Sysplex — it can act as stand-alone, but is still able to use its automation to activate additional engines via CBU and perform reconfiguration of LPARs during a failover.

As is the case with zGM, consistency of data is not controlled by GDPS. Instead, Global Mirror itself maintains its own data consistency.

Open LUN support

As is the case with GDPS/PPRC, GDPS/GM can maintain the mirroring environment for Open Systems data in addition to System z data. Data consistency is maintained by the Global Mirror replication technology, with GDPS serving as a front end to create an easy to use management environment and single point of control for all mirror related commands.

Also, as with GDPS/PPRC, GDPS/GM only maintains the mirroring environment for Open Systems. Recovery of the non-System z servers is left to the administrators and whatever procedures and scripts they have at their disposal.

GDPS/GM planned and unplanned reconfiguration support

Planned and unplanned reconfiguration actions include the capability to:

- ▶ Quiesce a system's workload and remove the system from the Parallel Sysplex cluster (for example, stop the system prior to a site switch)
- ▶ Reconfigure and IPL systems
- ▶ Quiesce a system's workload, remove the system from the Parallel Sysplex cluster, and re-IPL the systems

GDPS/GM provides a highly automated solution, to minimize the dependency on key human skills being available at the remote site, to recover from a disaster. GDPS/GM automates the process of recovering the production environment with minimal manual intervention, which can provide significant value in minimizing the duration of the recovery window.

GDPS/GM summary

GDPS/GM control software provides a single point of control for Global Mirror as well as restarting and reconfiguring the System z Application Systems at the Recovery site, while Global Mirror itself maintains the consistency of data. A K-System is used in the production site as the main control point while an R-System is maintained in the recovery site. The two systems maintain points of control as well as a heartbeat to notify the administrator that the two sites are healthy or if a possible event has occurred.

GDPS/GM typically accommodates these kinds of environments:

- ▶ Environments that cannot tolerate the performance impact of a synchronous remote copy technology, such as Metro Mirror.
- ▶ Environments that require more distance between their production and failover site than can be accommodated by GDPS/PPRC (which depends upon the underlying hardware).
- ▶ Environments that require a single point of control for mirroring both System z and open system data
- ▶ Environments that do not require TS7700 Grid control from the GDPS interface
- ▶ Environments where bandwidth might be limited

Here are some GDPS/GM solution highlights:

- ▶ Managed site failover
- ▶ High performance asynchronous remote copy
- ▶ Data loss measured in seconds
- ▶ Designed for unlimited distances
- ▶ Does not necessarily have to be sized for peak bandwidth
- ▶ Automatic Capacity Backup (CBU) System z CPU activation
- ▶ Single point of control for disk mirroring and recovery of System z environment
- ▶ Supports System z and Open Systems data

2.1.10 GDPS three site support

As we have discussed in Chapter 1, “Industry Business Continuity Trends and Directions” *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547, three site BC configurations are now starting to become more common. There are two versions of GDPS with three site support:

- ▶ GDPS Metro Mirror and z/OS Global Mirror
- ▶ GDPS Metro Global Mirror

GDPS Metro Mirror and z/OS Global Mirror

The design of these systems is shown in Figure 2-12.

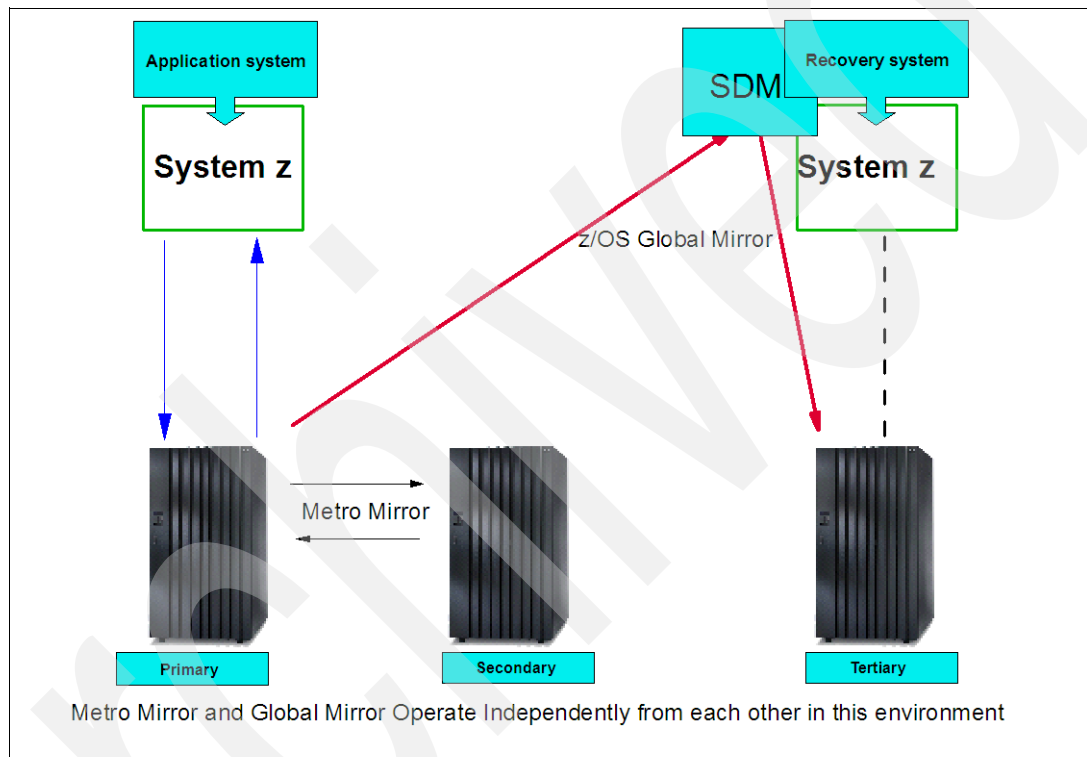


Figure 2-12 GDPS/z/OS Metro Global Mirror - the primary disk system serves as the source disk for both the Metro Mirror and z/OS Global Mirror relationships

Because they are based on fundamentally different disk mirroring technologies (one that is based on a relationship between two disk systems and another that is based on a relationship between a disk system and a z/OS server), it is possible to use Metro Mirror and z/OS Global Mirror from the same volume in a z/OS environment. This also means that GDPS Control Software can be used to enhance the solution.

In this case, GDPS/PPRC or GDPS/PPRC HyperSwap Manager would be used to protect the availability of data (through HyperSwap) on disk systems located within 100km or within the same building. Meanwhile, GDPS/XRC would act as a control point from disaster recovery.

GDPS/Metro Global Mirror

The other form of three site mirroring that is supported by GDPS is based on an enhanced cascading technology. In this case, rather than run both Metro Mirror and z/OS Global Mirror, we use Metro Global Mirror. The data passes from primary to secondary synchronously and asynchronously from the secondary to the tertiary. See Figure 2-13.

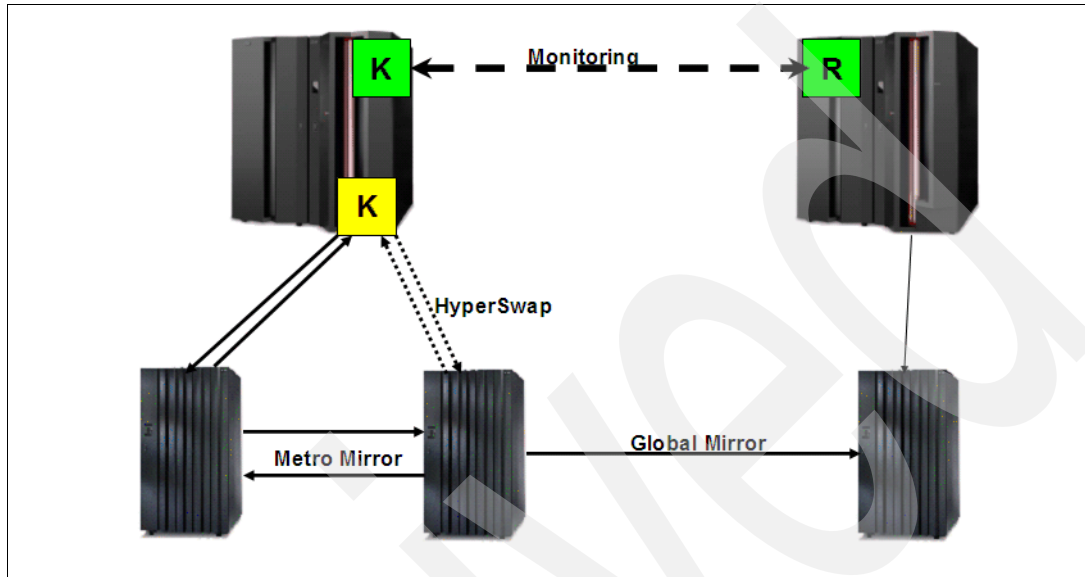


Figure 2-13 GDPS/Metro Global Mirror implementation uses a cascade of data passing from Metro Mirror to Global Mirror

In this case we are, again, able to use two forms of GDPS to support the environment. GDPS/PPRC or GDPS/PPRC HyperSwap Manager provides availability to the data via HyperSwap and metro area failover capabilities while GDPS/Global Mirror provides disaster recovery failover capabilities.

2.1.11 TS7740 Grid Support (GDPS/PPRC and GDPS/XRC)

GDPS also supports TS7740 Grid. By extending GDPS support to data resident on tape, the GDPS solution can provide continuous availability and near transparent business continuity benefit for both disk and tape resident data. As a result, enterprises might no longer be forced to develop and utilize processes that create duplex tapes and maintain the tape copies in alternate sites. For example, previous techniques created two copies of each DBMS image copy and archived log as part of the batch process and manual transportation of each set of tapes to different locations.

Operational data, or data that is used directly by applications supporting users, is normally found on disk. However, there is another category of data that “supports” the operational data, which is typically found on tape systems. Support data typically covers migrated data, point in time backups, archive data, etc. For sustained operation in the recovery site, the support data is indispensable. Furthermore, several enterprises have mission critical data that only resides on tape.

The TS7740 Grid provides a hardware-based duplex tape solution and GDPS can automatically manage the duplexed tapes in the event of a planned site switch or a site failure. Control capability has been added to allow GDPS to *freeze* copy operations, so that tape data consistency can be maintained across GDPS managed sites during a switch between the primary and secondary TS7740 clusters.

2.1.12 FlashCopy support (GDPS/PPRC, GDPS/XRC, GDPS/Global Mirror)

FlashCopy, available on DS6000 and DS8000, provides an instant point-in-time copy of the data for application usage such as backup and recovery operations. FlashCopy can copy or dump data while applications are updating the data. In FlashCopy V2, the source and target volumes can reside in different logical sub system (LSS). Therefore, GDPS also supports a FlashCopy from a source in one LSS to a target in a different LSS within the same disk system.

FlashCopy before resynchronization is automatically invoked (based upon GDPS policy) whenever a resynchronization request is received. This function provides a consistent data image to fall back to, in the rare event that a disaster should occur during resynchronization. FlashCopy can also be user-initiated at any time. The tertiary copy of data can then be used to conduct D/R testing while maintaining D/R readiness, perform test or development work, shorten batch windows, and so on.

There are two FlashCopy modes; COPY mode runs a background copy process and NOCOPY mode suppresses the background copy. GDPS/PPRC and GDPS/XRC support both of these modes.

GDPS also supports:

- ▶ FlashCopy NOCOPY2COPY, which allows changing an existing FlashCopy relationship from NOCOPY to COPY. This gives you the option of always selecting the NOCOPY FlashCopy option and then converting it to the COPY option when you want to create a full copy of the data in the background at a non-peak time.
- ▶ Incremental FlashCopy, with which a volume in a FlashCopy relationship can be refreshed, reducing background copy time when only a subset of the data has changed. With Incremental FlashCopy, the initial relationship between a source and target is maintained after the background copy is complete. When a subsequent FlashCopy establish is initiated, only the data updated on the source since the last FlashCopy is copied to the target. This reduces the time required to create a third copy, so that FlashCopy can be performed more frequently.

2.1.13 IBM Global Technology Services (GTS) offerings for GDPS overview

The following GDPS services and offerings are provided by GTS.

GDPS Technical Consulting Workshop (TCW)

TCW is a two day workshop where Global Technology Services specialists explore the business objectives, service requirements, technological directions, business applications, recovery processes, cross-site and I/O requirements. High-level education on GDPS is provided, along with the service and implementation process. Various remote and local data protection options are evaluated.

Global Technology Services specialists present a number of planned and unplanned GDPS reconfiguration scenarios, with recommendations on how GDPS can assist in achieving business objectives. At the conclusion of the workshop, the following items are developed:

- ▶ Acceptance criteria for both the test and production phases
- ▶ A high level task list
- ▶ A services list
- ▶ Project summary

Remote Copy Management Facility (RCMF)

With this service, RCMF/PPRC or RCMF/XRC automation to manage the remote copy infrastructure are installed, the automation policy is customized, and the automation is verified along with providing operational education for the enterprise.

GDPS/PPRC HyperSwap Manager

IBM Implementation Services for GDPS/PPRC HyperSwap Manager helps simplify implementation by installing and configuring GDPS/PPRC HyperSwap Manager and its prerequisites up and running with limited disruption. On-site planning, configuration, implementation, testing, and education are provided.

GDPS/PPRC, GDPS/XRC, and GDPS/GM

IBM Implementation Services for GDPS assist with planning, configuration, automation code customization, testing, onsite implementation assistance, and training in the IBM GDPS solution.

2.1.14 Summary: Value of GDPS automation

Here we summarize some of the many values of the GDPS Business Continuity automation offerings.

Repeatability, reliability of recovery: Successful, reliable, repeatable business continuity and recovery necessitates a dependency upon automation rather than people. Automation can help reduce risks that critical skilled personnel are not fully available to recover the business, in the event of a site outage.

GDPS automation supports fast, reliable switching all resources from one site to another from a single trigger point, from a single point of control, with minimal operator interaction required.

Single point of Control for disk mirroring: All forms of GDPS act as a single point of control for all disk mirroring related functions. In cases where Open LUN Management is in use, this spans beyond the z/OS environment and manages the disk mirroring related tasks for any attached system. This helps simplify the infrastructure by using one control point, and reduces the complexity that might otherwise reign if control software were not in place to manage the day to day work in the mirroring environment.

Control of Metro Mirror Data Consistency: GDPS/PPRC and GDPS/PPRC HyperSwap Manager automation assures that the secondary copy of the data at the Metro Mirror secondary site is *data consistent*. Data consistency means that, from an application's perspective, the secondary disk's data is time and data consistent. Data consistency in the secondary copy of the data means that databases can be restarted in the secondary location without having to go through a lengthy and time consuming data recovery process, avoiding a time-consuming process of restoring database tape backups and applying logs.

Support of data consistency for tape and disk storage: GDPS/PPRC and GDPS/XRC also support and manage data consistency in the IBM TS7700 Grid. This helps maintain consistency between disks and tapes, and therefore provides a total business continuity solution for the entire storage.

Integration: GDPS automation integrates many technologies, including (but not limited to) System z Parallel Sysplex, z/OS System Automation, Tivoli NetView, and the storage system's advanced mirroring features to form an integrated, automated business continuity and disaster recovery solution.

Exploitation of System z Capacity Backup: GDPS automation exploits the system z Capacity backup (CBU) feature, which can provide additional processing power at the alternate site and help lower costs. The CBU feature increments capacity temporarily at the time it is required, in the event of a disaster or outage.

CBU adds Central Processors (CP) to the available pool of processors and can be activated only in an emergency. GDPS CBU management can also automate the process of dynamically returning reserved CPs after the emergency period has expired.

2.2 GDPS components in more technical detail

In this section we describe, in more technical detail, the various components of the GDPS implementations:

- ▶ GDPS/PPRC support of Consistency Group FREEZE
- ▶ HyperSwap function
- ▶ GDPS Open LUN Management
- ▶ GDPS/PPRC Multi-Platform Resiliency for System z
- ▶ GDPS single site and multi-site workload
- ▶ GDPS extended distance support between sites
- ▶ GDPS/XRC implementation
- ▶ GDPS automation of System z Capacity Backup (CBU)
- ▶ GDPS and TS7700 Grid support (GDPS/PPRC and GDPS/XRC)
- ▶ GDPS FlashCopy support
- ▶ GDPS prerequisites

2.2.1 GDPS/PPRC support of Consistency Group FREEZE

As stated in “Requirement for data consistency” on page 18, data consistency across all primary and secondary volumes, spread across any number of storage systems, is essential in providing data integrity and the ability to do a normal database restart in the event of a disaster. Time consistent data in the secondary site allows applications to restart without requiring a lengthy and time-consuming data recovery process.

GDPS/PPRC uses a combination of storage system and Parallel Sysplex technology triggers to capture, at the first indication of a potential disaster, a data consistent secondary site (site 2) copy of the data, using the Metro Mirror FREEZE function. The FREEZE function, initiated by automated procedures, freezes the image of the secondary data at the very first sign of a disaster, even before any database managers know about I/O errors. This can prevent the logical contamination of the secondary copy of data that would occur if any storage system mirroring were to continue after a failure that prevents some, but not all secondary volumes from being updated.

The purpose of the Consistency Group functionality, combined with the FREEZE function, is to assure that the remote site data is in data integrity, and to prevent mirroring data integrity problems due to a *rolling disaster* (see Chapter 6 “Planning for Business Continuity in a heterogeneous IT environment” in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547).

Rolling disasters occur when different disk systems fail at different points in time, but processing continues. This happens because disasters rarely affect an entire data center at the exact same moment. A fire, for example, might bring down disk systems milliseconds to minutes apart from each other. If the processors continue to attempt transactions and the other systems are still copying data, the individual Logical Subsystems (LSSs) of a synchronous mirrored disk system might end up out of sync in the remote site. Under these circumstances, a database restart becomes impossible and the database must be manually recovered, a process which can take hours or even days.

GDPS reacts to two types of triggers: FREEZE triggers and HyperSwap triggers. A FREEZE trigger is an indication of a mirroring problem, and a HyperSwap trigger is an indication of a primary disk system failure (some of the HyperSwap triggers are only produced if you have enabled HyperSwap). The policy is specifiable so that STOP or FREEZE is performed as appropriate.

When a GDPS trigger event occurs, GDPS issues a simultaneous Consistency Group FREEZE to all disk system LSSs on all GDPS-managed disk systems as soon as an event is detected. The GDPS automation insures that the data at the remote site is frozen as a single, data consistent consistency group entity. Once the FREEZE is completed, the GDPS/PPRC next steps depend on the GDPS policy that has been implemented.

There are different parameters for the FREEZE and HyperSwap:

- ▶ FREEZE and GO
- ▶ FREEZE and STOP
- ▶ FREEZE and STOP COND
- ▶ SWAP,GO
- ▶ SWAP,STOP

FREEZE and GO

GDPS freezes the secondary copy of data where remote copy processing suspends and the production system then continues to execute, making updates to the primary copy of data.

As the mirroring has been suspended at the point of the FREEZE, these updates are not propagated onto the secondary disk systems if there is a subsequent Site1 failure.

FREEZE and STOP

GDPS freezes the secondary copy of data when remote copy processing suspends and immediately quiesces the production systems, resulting in all the work that is updating the primary Metro Mirror devices being stopped, thereby preventing any data loss.

This option might cause the production systems to be quiesced for transient events (false alarms) that interrupt Metro Mirror processing, thereby adversely impacting application availability.

However, it is the only option that can guarantee no data loss and complete disk data consistency.

FREEZE and STOP CONDITIONAL

GDPS freezes the secondary copy of data. If GDPS then determines that the FREEZE was caused by the storage systems that contain the secondary copy of data (that is, a problem in the secondary site), processing is the same as for FREEZE and GO; otherwise processing is the same as for FREEZE and STOP.

SWAP,STOP or SWAP,GO

The first parameter specifies that if HyperSwap is enabled and all the environmental conditions for HyperSwap are met, a HyperSwap is to be attempted in response to HyperSwap triggers. The HyperSwap function starts with a FREEZE, so you have a consistent set of disks in the other site.

The next part of HyperSwap processing is where the contents of the UCBs in all GDPS systems are swapped (in fact, the HyperSwap is attempted on the Controlling GDPS system first, and only attempted on the other systems if that HyperSwap is successful).

The second parameter (GO or STOP) indicates the action GDPS is to take if a FREEZE trigger, rather than a HyperSwap trigger, is presented. In this case, specifying FREEZE=SWAP,GO causes GDPS to react the same as though you had specified FREEZE=GO. Specifying FREEZE=SWAP,STOP causes GDPS to react the same as though you had specified FREEZE=STOP.

If you want to exploit unplanned HyperSwap support, it is not possible to use FREEZE=COND processing should you get a FREEZE trigger — that is, there is no SWAP,COND option.

Next we give some examples to illustrate the effect of the different options.

FREEZE example 1

You are running GDPS and something happens that triggers a GDPS FREEZE. GDPS performs the FREEZE, performs the action you have requested in the FREEZE policy, and then issues a takeover prompt. You determine as a result of investigating that an operator and a CE were reconfiguring a FICON director and mistakenly blocked the wrong paths, removing the last Metro Mirror path between the primary and secondary disk systems.

If your policy specified the FREEZE and GO option, there is no major impact to the production systems. Production continues with no impact during the time it took to diagnose the problem. At an appropriate later time, you do a FlashCopy-before-resync and then resync the primary and secondary Metro Mirror disks after having fixed the blocked paths on the Director.

If your policy specified the FREEZE and STOP, or FREEZE and STOP CONDITIONAL option, all production systems stop before issuing the takeover prompt and probably stay down for the time it takes you to diagnose the problem, thus impacting your business. However, a policy that indicated a guarantee of no data loss at the remote site would have been honored.

FREEZE example 2

Suppose that a FREEZE occurred at 9:00 a.m. GDPS issued an alert and a takeover message and then it is determined that there is a fire in the computer room.

If you chose the FREEZE and GO option: The secondary Metro Mirror devices are frozen at 9:00 and applications continue for some amount of time, updating the primary volumes. When the fire eventually brings all the systems down, any transactions that completed at the primary site after the 9:00 a.m. FREEZE *have not been copied to the secondary site*. The remote site data has preserved data integrity, and a fast, repeatable database restart can be performed.

As part of the database restart and ensuing transaction integrity roll forward/roll back, any valid data that was created after the 9:00 am FREEZE would have to be recreated.

If you chose the FREEZE and STOP or FREEZE and STOP CONDITIONAL options: GDPS stops all production systems before any further processing can take place, and therefore no data loss occurs. When you fail over to the other site and recover the secondary disks, all secondary site data is identical to the primary site data as of 9:00 a.m., when the FREEZE occurred.

Which of the FREEZE options you select: This is clearly a business decision and can differ, depending on the applications involved. Some applications can tolerate some data loss, but not any interruptions of service unless it is a real disaster, while other applications cannot afford to lose any data and might take the risk of an unnecessary outage in order to guarantee the integrity of the secondary data. Obviously all the applications in the GDPS group are treated in the same way, so your decision has to cater for all the applications that run on the systems in the GDPS group.

FREEZE options summary

FREEZE=STOP and FREEZE=SWAP,STOP are the only options that guarantee no data loss and also data consistency across the Metro Mirror FREEZE devices and Coupling Facilities.

FREEZE=COND (FREEZE and STOP CONDITIONAL) could result in a FREEZE and GO being executed. A subsequent failure would yield the same consequences described for FREEZE and GO. This risk is very small but must be evaluated against the business requirements.

FREEZE=GO and FREEZE=SWAP,GO have issues and considerations associated with Coupling Facility, Coupling Data Set, and disk data consistency. As a part of the GTS implementation, the implication of the FREEZE and GO policy would be extensively reviewed and tested. FREEZE=GO results in data loss for any data created after the FREEZE event of a Site1 disk system or complete Site1 failure following a FREEZE event, however, this might be desirable, because a transient trigger event would not fail the primary site applications. A FREEZE=SWAP,GO that results in a FREEZE and GO being performed has the same issues as FREEZE and GO.

During a typical client implementation, a FREEZE and GO is normally the initial choice for the FREEZE policy. Once GDPS has been running for a period of time and you are confident that it is properly set up, and you are not receiving any false triggers, you should consider changing your policy to FREEZE and STOP and eventually evolve into a HyperSwap-enabled environment with a FREEZE=SWAP,STOP policy.

Tip: The appropriate policy for any business depends on the individual business requirements. The GTS implementation that accompanies all GDPS family of offerings assists you in determining and implementing the appropriate Consistency Group FREEZE choices for your environment.

2.2.2 HyperSwap function

HyperSwap extends System z Parallel Sysplex redundancy to Metro Mirror disk systems. HyperSwap provides the ability to rapidly swap, between sites, a large Metro Mirror configuration, within seconds.

The purpose of HyperSwap is to use the System z server and operating system to point to different Unit Control Blocks (UCBs) associated the secondary set of Metro Mirror disks. Upon successful completion of a HyperSwap, the primary production applications access the secondary volumes.

Figure 2-14 shows a diagram of a HyperSwap configuration within a single site.

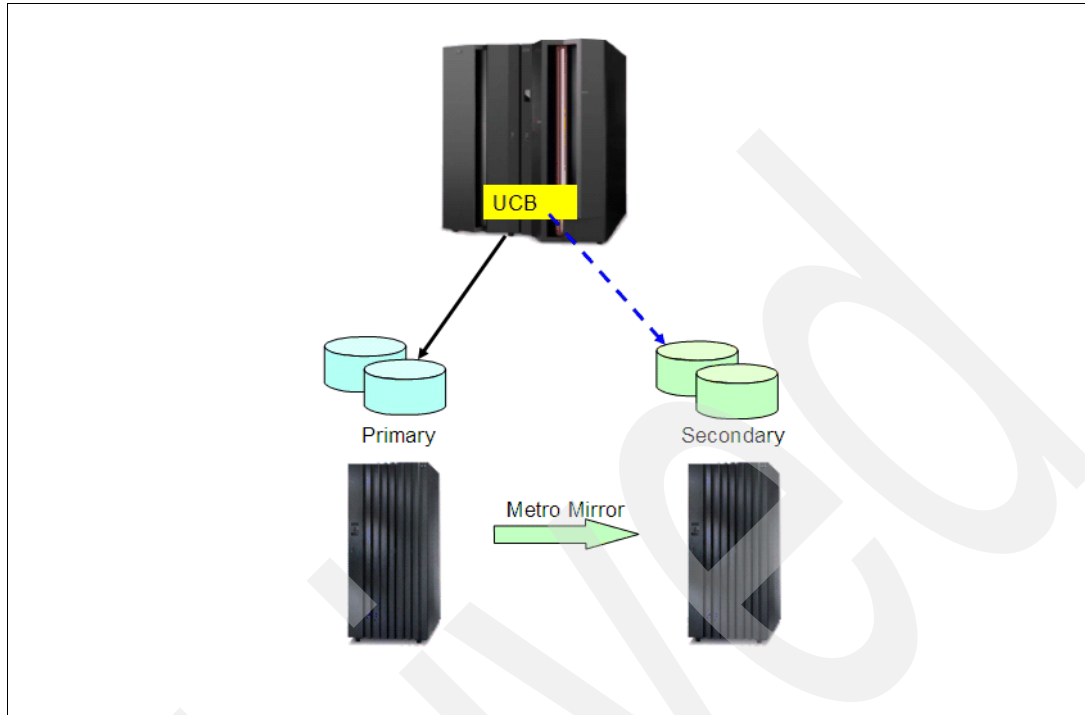


Figure 2-14 The HyperSwap function transparently swaps disks

A *planned* HyperSwap offers the following capabilities:

- ▶ Planned HyperSwap can transparently switch all primary Metro Mirror disk systems with the secondary Metro Mirror disk systems for a planned reconfiguration.
- ▶ Planned HyperSwap support of nondisruptive disk system switching, or site switching, can allow important planned outages such as performing disk configuration maintenance and planned site, or maintenance, to be performed without requiring quiescence of applications.

The *unplanned* HyperSwap function transparently switches to the secondary Metro Mirror disk systems in a matter of seconds if an unplanned outage occurs. As a result:

- ▶ Production systems might be able to remain active during a disk system failure. Disk system failures can be masked from being a single point of failure for an entire Parallel Sysplex.
- ▶ Production servers can remain active during a failure of the site containing the primary Metro Mirror disk systems if applications are cloned and they are exploiting data sharing across the two sites. Depending on the configuration of the servers, the FREEZE policy, and where the workloads were running, the workload in the second site might or might not have to be restarted.

Requirement: The HyperSwap function cannot move any Reserves that might exist for the primary devices at the time of the swap.

For this reason, HyperSwap requires that all Reserves are converted to Global ENQs. This requirement is enforced through GDPS monitoring and GDPS disables HyperSwap if the requirement is found to be violated.

Requirement: The GDPS HyperSwap function requires that all your production system and application data, including operating system volumes (with the exception of couple data set and utility devices) be mirrored in the Metro Mirror configuration.

Figure 2-15 shows a GDPS/PPRC HyperSwap Manager, single site configuration.

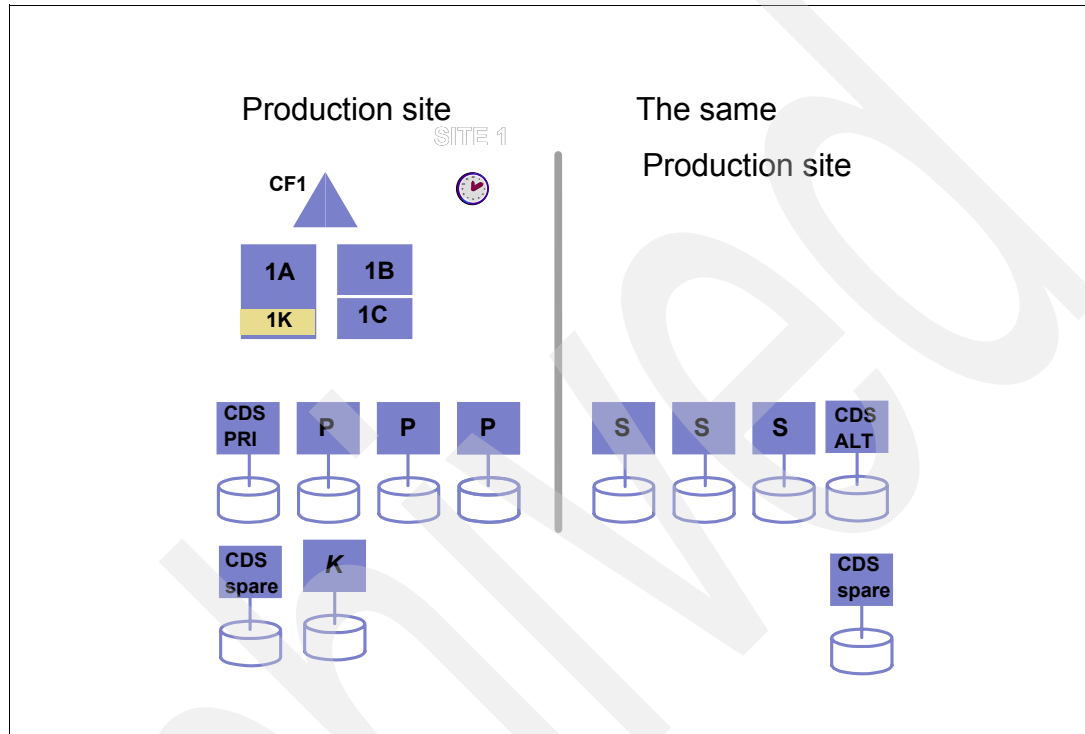


Figure 2-15 GDPS/PPRC HyperSwap Manager configuration - single site configuration

We define the labels in this figure as follows:

- ▶ (P) are the primary disk systems.
- ▶ (S) are the secondary disk systems.
- ▶ (K) are GDPS controlling system disks, are isolated from the production system disks.
- ▶ (1K) is the GDPS controlling system.
- ▶ (CDS) are couple data set.
- ▶ (1A-1B-1C) are production LPARs.

HyperSwap can be done in a planned site switch; for example, to take the primary disks offline for maintenance without taking an outage, as shown in Figure 2-16.

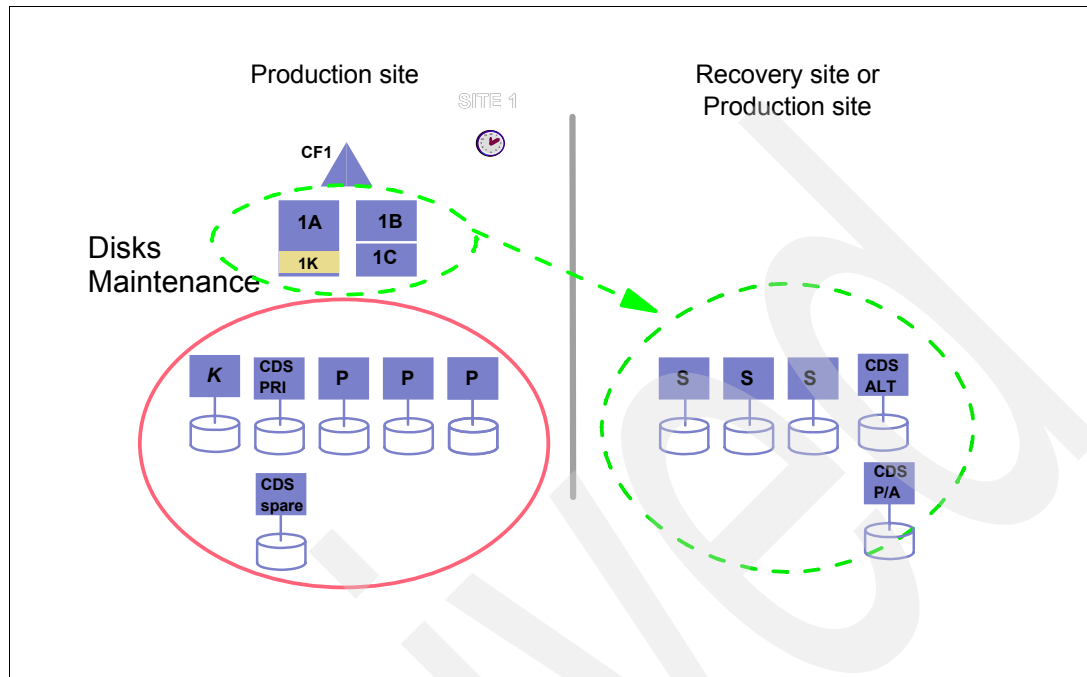


Figure 2-16 Planned swap under GDPS/PPRC HyperSwap Manager

The HyperSwap function also can be invoked in the event of an unplanned disk system failure, as shown in Figure 2-17.

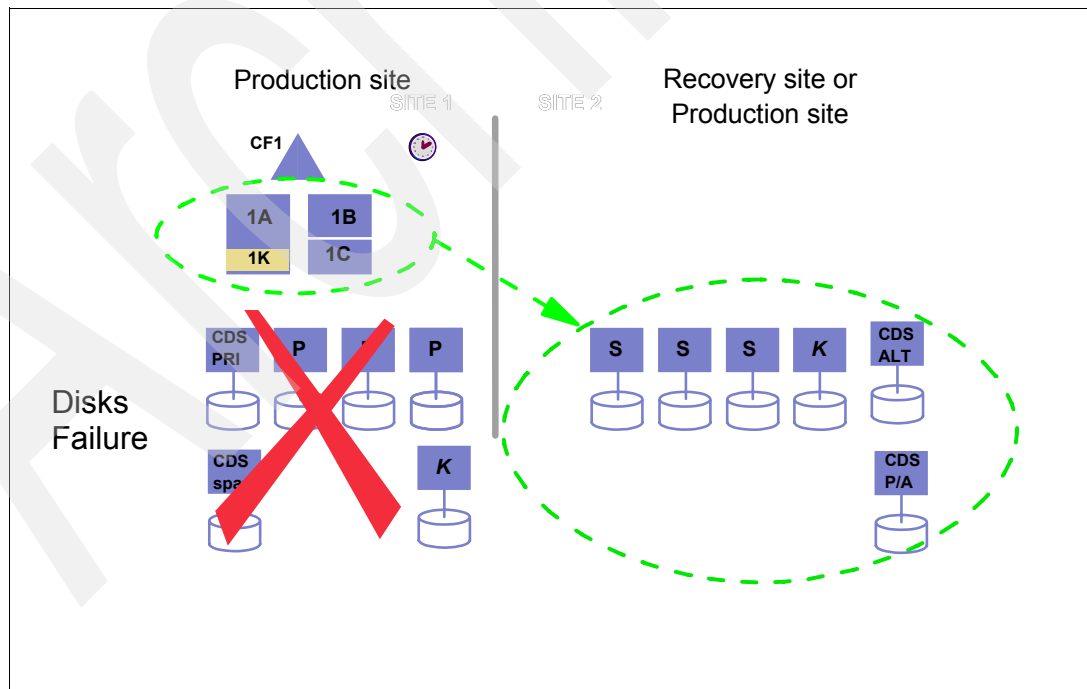


Figure 2-17 GDPS/PPRC HyperSwap unplanned outage disk system failure recovery

In either case, the general steps for a unplanned outage Hyperswap support are:

1. Unplanned outage affects primary disk system.
2. Swap Metro Mirror primary and secondary disks.
3. GDPS reverses the direction of the disk mirroring paths, and via exploitation of the Metro Mirror Failover/Failback functions. This enables bitmap change tracking of incremental updates at the remote site — to prepare for the eventual return back to the original disk systems.
4. Upon successful completion of the HyperSwap, the production workload is running on the secondary set of disks.

Note: GDPS supports Metro Mirror's Failover/Failback function:

- ▶ It offers a faster, more efficient functionality to reverse the direction of a Metro Mirror pair.
- ▶ It automatically enables change recording bitmap to be started on secondary Metro Mirror disks. When we fail back to the source disks, only the incrementally changed tracks have to be copied.
- ▶ *Failover* makes the former secondary be a suspended primary, reverses the Metro Mirror link direction, and starts change recording on secondary disks.
- ▶ *Failback* resynchronizes the secondary to the primary according to Change Recording (CR) bitmap.

HyperSwap symmetrical disk configuration requirement

The HyperSwap function requires that primary and secondary SubSystem ID (SSID) are symmetrically paired (see Figure 2-18), with a one-to-one relationship between the primary and secondary disks and Logical SubSystems (LSS). If primary and secondary SSIDs are not configured symmetrically, GDPS disables the HyperSwap function.

Furthermore, the HyperSwap function requires all the mirrored disks to be in the FREEZE group. GDPS validation checks for this requirement and if not met, disables the HyperSwap function.

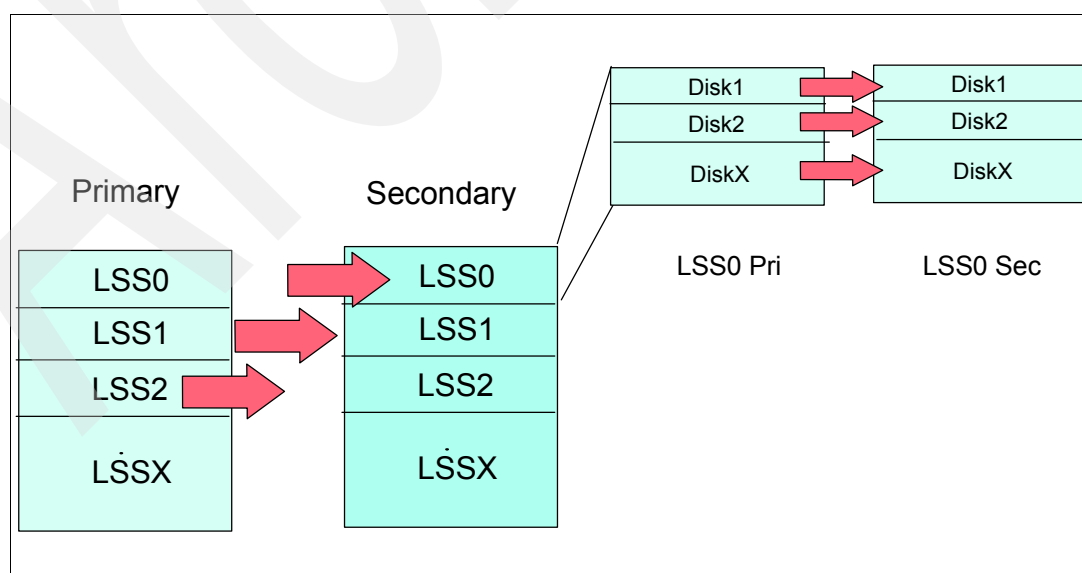


Figure 2-18 Symmetric configuration required by HyperSwap between primary and secondary

2.2.3 GDPS Open LUN Management

As data centers and networks expand, there is an increasing dependency between mainframes and open systems. This results in a requirement for consistency groups that not only span the System z, but for Open Systems data also to be mirrored and included in the consistency group with the System z data.

This requirement is satisfied through the GDPS Open LUN Management capability, available on both GDPS/PPRC and GDPS/PPRC HyperSwap Manager. GDPS Open LUN Management extends the Metro Mirror Consistency Group and FREEZE functions to Open Systems LUNs that are present on the disk systems. This function assures data consistency at the remote site across all of the System z and Open Systems volumes and LUNs defined to be in the same Consistency Group.

With GDPS Open LUN Management, the Consistency Group for Metro Mirror can now span System z and Open data. In the event of either a planned or unplanned outage, the GDPS/PPRC or GDPS/PPRC HyperSwap Manager code is able to control Open Systems LUNs using the same Consistency Group FREEZE functionality that was described in 2.2.1, “GDPS/PPRC support of Consistency Group FREEZE” on page 38.

Figure 2-19 shows a diagram of this function.

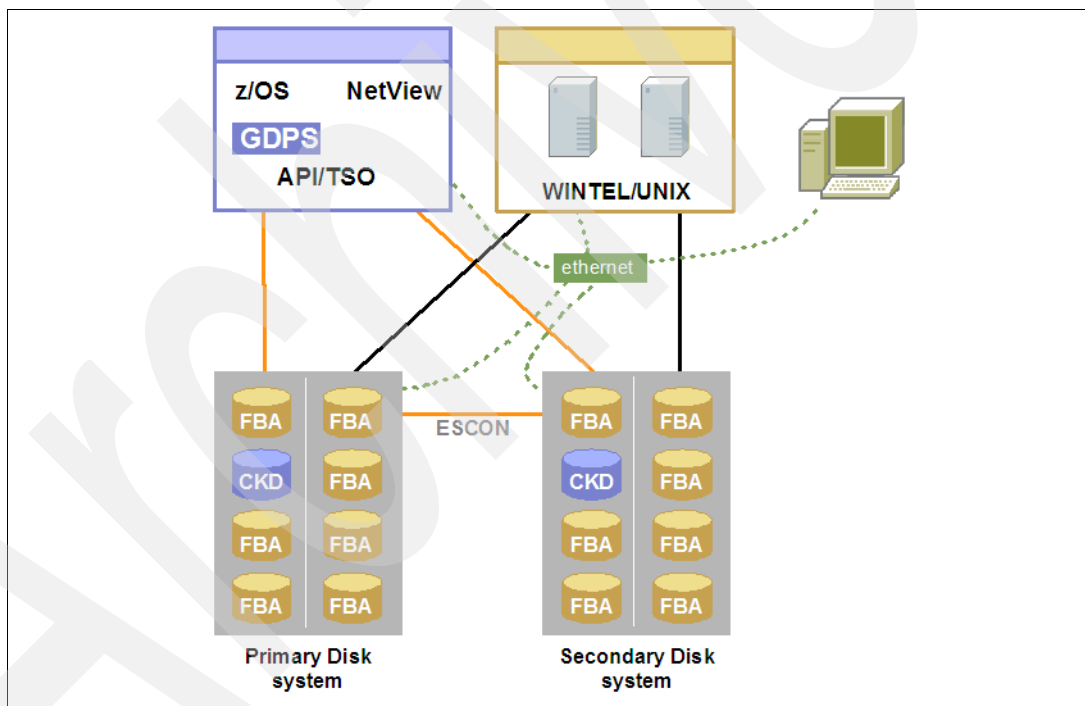


Figure 2-19 GDPS/PPRC Open LUN storage configuration

To make use of GDPS Open LUN Management, at least one array in the Metro Mirror disk system must be assigned to a System z server. This is shown in Figure 2-19, where each Metro Mirror disk system has one array of System z-attached Count Key Data (CKD), while the rest is allocated in Fixed Block (FBA) data used by open systems.

This one System z array allows GDPS to monitor activities in the Metro Mirror disk system (including SNMP data and triggers from the open systems volumes), to manage Metro Mirror for both System z and Open, and trigger and manage the FREEZE function (thus insuring data consistency across all LSSs connected to the GDPS system, whether they are System z data or Open data).

No GDPS code has to be installed on any open systems server. SUSPEND and FREEZE events are triggered and managed by GDPS Open LUN Management, after a trigger SNMP alert.

Note: GDPS/PPRC or GDPS/PPRC HyperSwap Manager does not automate the restart of open systems, nor can it automate open systems FlashCopy. The consistency and restartability of the open systems data is protected, but bringing up systems and applications on the open servers must be a manual process.

The GDPS Open LUN Management functionality is available as part of the GDPS/PPRC and GDPS/PPRC HyperSwap Manager services offering from GTS.

In summary, here are some considerations of the GDPS Open LUN Management function:

- ▶ Some CKD capacity in the Metro Mirror disk system is required.
- ▶ The disk system must be at Metro Mirror architecture level 4.
- ▶ Cross platform or platform level Consistency Group and FREEZE are supported.
- ▶ Manual restart of open systems is required.

2.2.4 GDPS/PPRC Multi-Platform Resiliency for System z

As of GDPS 3.2, GDPS/PPRC fully supports Linux² on System z, running on z/VM partitions, as full participants in the GDPS solution. GDPS/PPRC can now provide a coordinated disaster solution for both workloads.

Note: The z/VM partitions must run Storage Automation for MultiPlatform; see Figure 2-20 on page 48.

² Please check with your IBM representative to assure that your software distribution of Linux on System z has the necessary prerequisite Linux code to support this function.

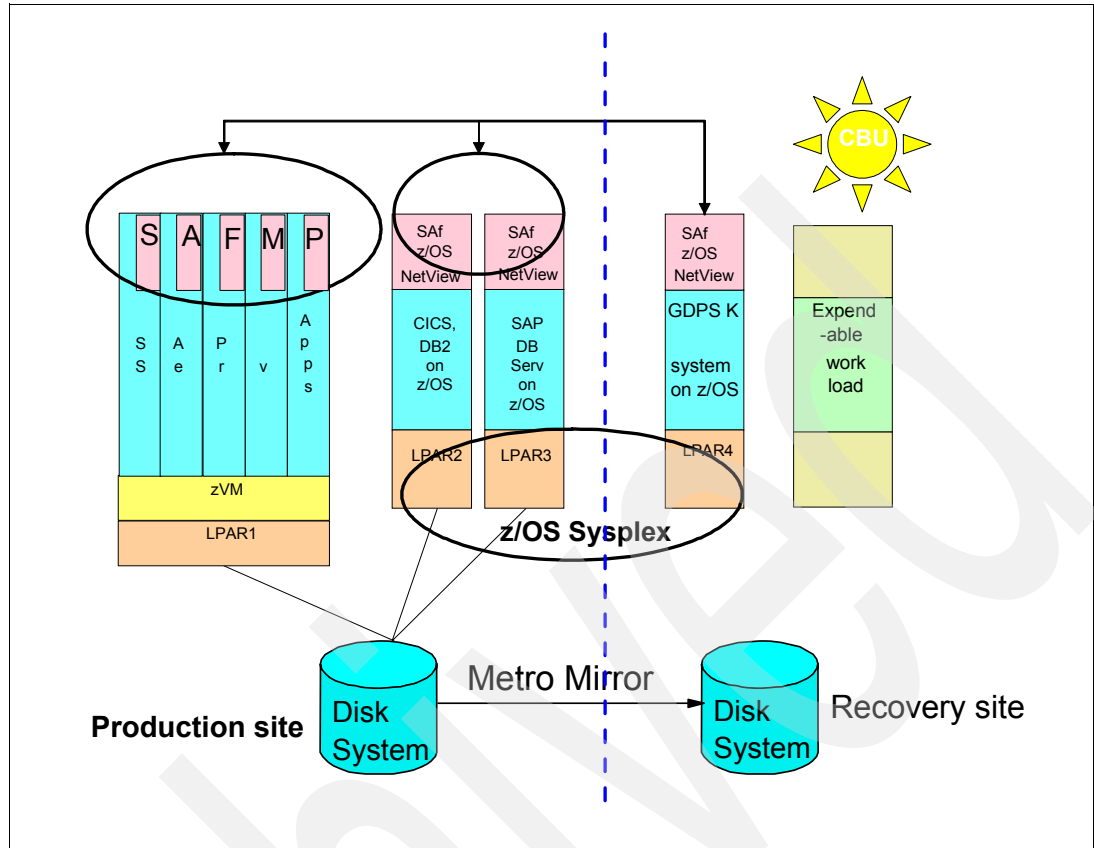


Figure 2-20 GDPS/PPRC Multiplatform resiliency for Linux

z/VM 5.1 and higher provides support for the HyperSwap function, enabling the virtual device associated with one real disk to be swapped transparently to another disk. HyperSwap can be used to switch to secondary disk storage systems mirrored by Metro Mirror.

HyperSwap can help in data migration scenarios to allow applications to move to new disk volumes without requiring them to be quiesced. GDPS/PPRC provides the reconfiguration capabilities for Linux on System z servers and data in the same manner it handles z/OS systems and data. In order to support planned and unplanned outages, GDPS provides the following recovery actions:

- Re-IPL in place of failing operating system images
- Site takeover/failover of a complete production site
- Coordinated planned and unplanned HyperSwap of disk systems, transparent to the operating system images and applications using the disks

Linux in System z environments running on z/VM partitions, participating in a GDPS environment, can be recovered at a remote location in addition to the z/OS-based environments operating in the same GDPS environment.

2.2.5 GDPS single site and multi-site workload

GDPS can be configured to run either single site or multi-site workloads.

Single site workloads are used when all production workload is executed on the primary site, and the recovery site is used for expendable workload (for example, test).

Multi-site workloads are used when the workload is balanced between the production and the recovery sites. Both sites can have the CBU feature. Typically, this is a System z Parallel Sysplex environment.

GDPS/PPRC single site workload examples

A single site workload configuration is shown in Figure 2-21.

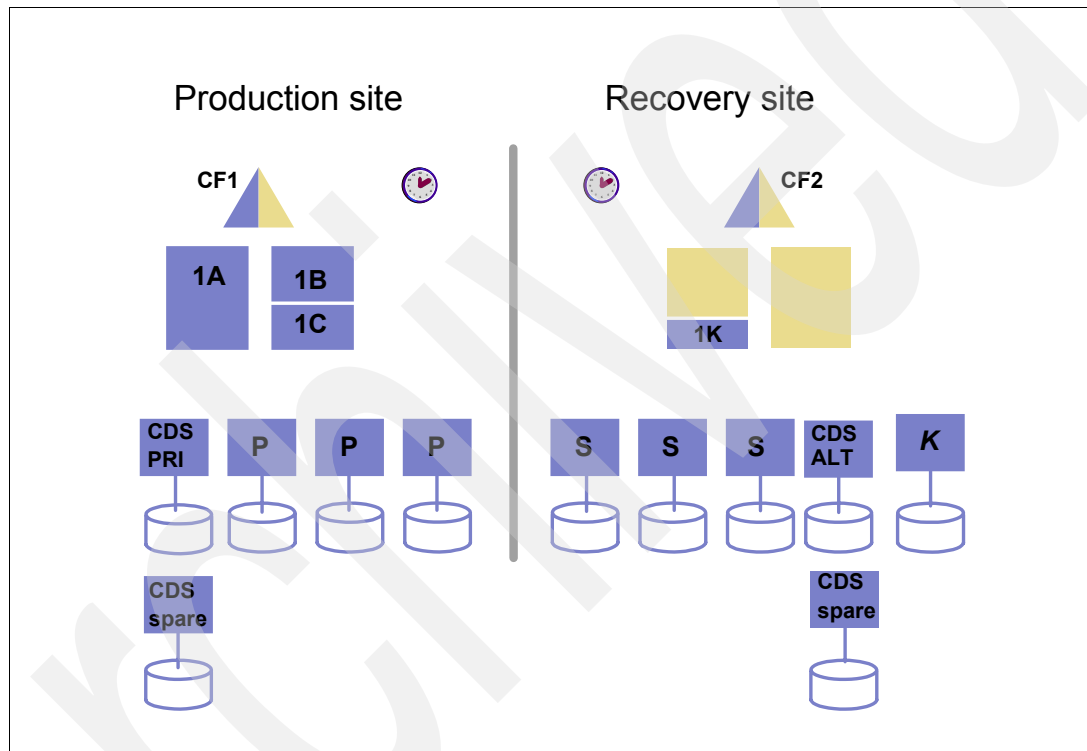


Figure 2-21 GDPS/PPRC in a single site workload cross-site Sysplex configuration

In Figure 2-21, we define the labels as follows:

- ▶ (P) indicates the primary disk systems.
- ▶ (S) indicates the secondary disk systems (mirrored in duplex status).
- ▶ (K) indicates the controlling system disks, which are isolated from the production system disks.
- ▶ (CDS) indicates the couple data sets.
- ▶ (1K) indicates the controlling system.
- ▶ (1A-1B-1C) indicate the production LPARs.

Planned disks swap in single site workload environment

Figure 2-22 shows the steps for a planned site switch in a single site workload environment. In our solution, we have sufficient cross-site connectivity, and we use HyperSwap to swap to the secondary disk and leave the z/OS workload and z/OS processors running where they are.

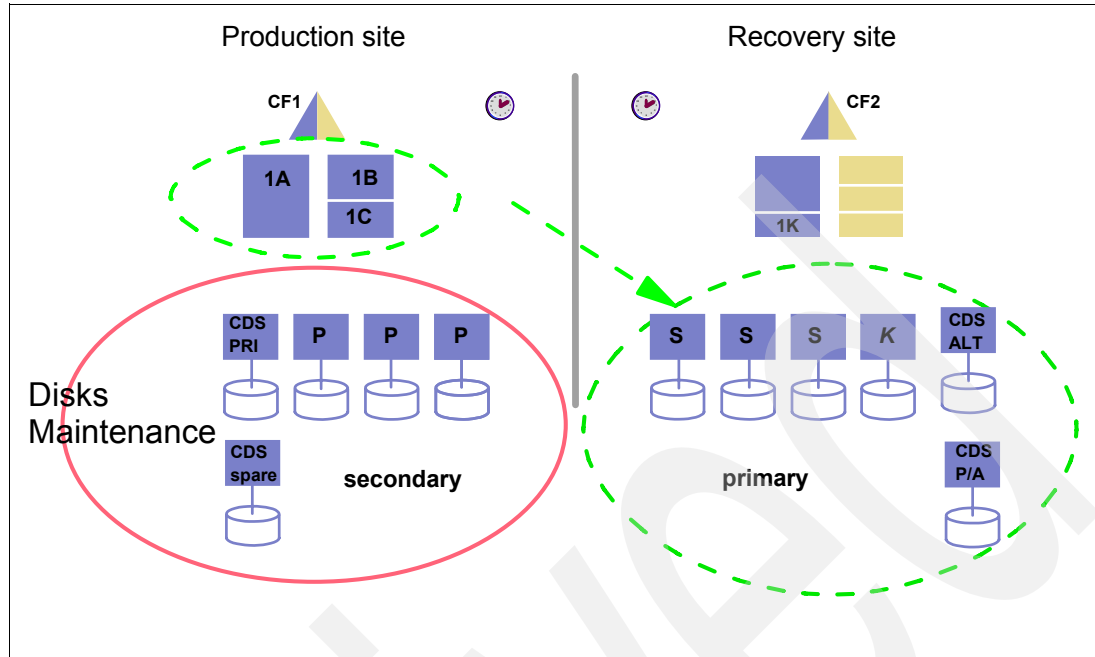


Figure 2-22 Single site workload cross-site Sysplex planned disks reconfiguration

To switch the disks, GDPS performs the following operations according to a predefined user script:

1. Move couple data sets to the recovery site.
2. Swap the Metro Mirror primary and secondary disks.
3. Perform the Metro Mirror Failover/Failback functions to enable changed recording and eventual return to the primary site. Upon successful completion, the primary disks are at the recovery site.
4. Optionally change the IPL table so that subsequent IPL is at the recovery site.

To return to the normal disk configuration, GDPS performs the following operations:

1. Swap Metro Mirror primary and secondary disks and resync the Metro Mirror session, using the Failover/Failback Metro Mirror function. After the successful completion, primary disks are at the production site and are in duplex with the secondary disks at the recovery site.
2. Move couple data set at the production site.
3. Optionally change the IPL table so that the subsequent IPL is at the production site.

Unplanned disk swap in single site workload environment

In the next example, we see the benefits of the HyperSwap in unplanned disk failure (Figure 2-23); the failure is masked to the applications and they can continue on the secondary disks. We have sufficient cross-site connectivity to continue running the applications on the production systems.

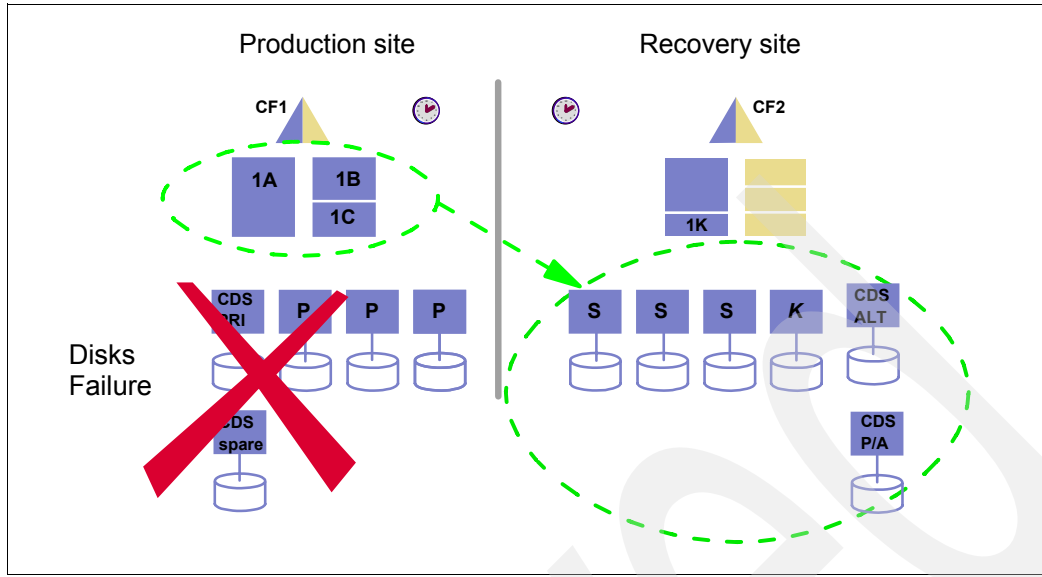


Figure 2-23 Single site workload cross-site Sysplex disks outage

When the unplanned outage GDPS trigger event occurs, GDPS performs the following operations:

1. Swap Metro Mirror primary and secondary disks.
2. Perform the Metro Mirror Failover/Failback functions to enable changed recording and eventual return to the primary site. Upon successful completion, the primary disks are set in the recovery site.
3. Move the couple data set to the recovery site.
4. Optionally change the IPL table so that the subsequent IPL is at the recovery site.

Unplanned site outage in single site workload environment

Because the status of the production systems is not known during an unplanned outage, GDPS does not attempt to shut down systems at the production site, but they are reset (see Figure 2-24).

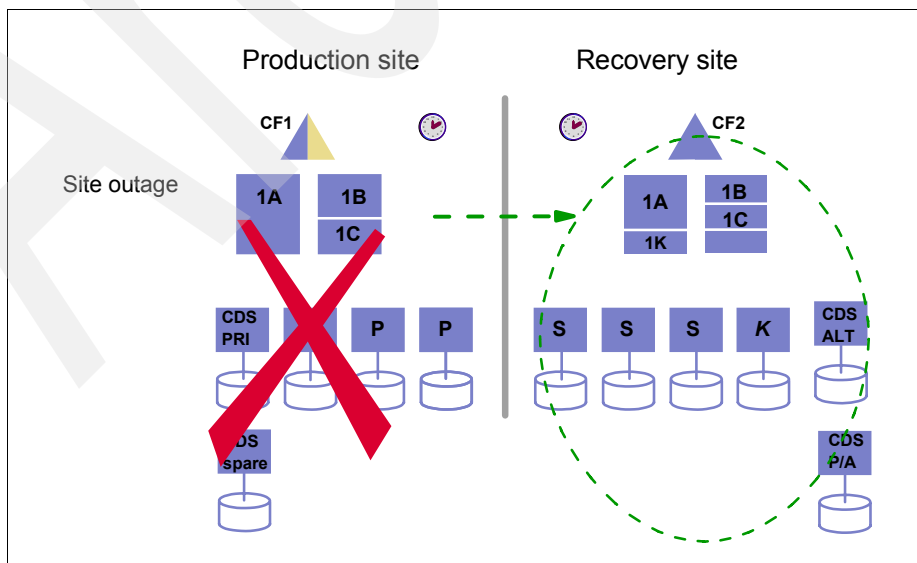


Figure 2-24 Single site workload cross-site Sysplex site outage

Upon receiving the GDPS trigger event, GDPS performs the following operations:

1. Reset all production systems.
2. Swap Metro Mirror primary and secondary disks.
3. Perform the Metro Mirror Failover/Failback functions to enable changed recording and eventual return to the primary site. Upon successful completion, the primary disks are set at the recovery site.
4. Configure couple data set at the recovery site only.
5. Optionally activate CBU capacity.
6. Recover CFRM and start using CF at the recovery site (clean up CF).
7. Change the IPL table so that subsequent IPL is at the recovery site.
8. IPL systems at the recovery site.

To return to the normal disk configuration (after the problem has been fixed), we use a planned site switch in the opposite direction.

GDPS/PPRC multi-site workload

In a multi-site workload configuration, because the application systems are already up and running at both sites, GDPS/PPRC is able to provide Near Continuous Availability for disks and site outage or maintenance.

Disk planned or unplanned swap

In this configuration, because we are running our workload in both sites; we have a GDPS K-System at both sites (to protect both systems from outage).

In Figure 2-25, we define the labels follows:

- ▶ (P) indicates the primary disk systems.
- ▶ (S) indicates the secondary disk systems (mirrored in duplex status).
- ▶ (K) indicates the controlling system disks, which are isolated from the production ones.
- ▶ (CDS) indicates the couple data sets.
- ▶ (1K-2K) indicate the controlling systems.
- ▶ (1A-1B-1C-1D) indicate the production LPARs.

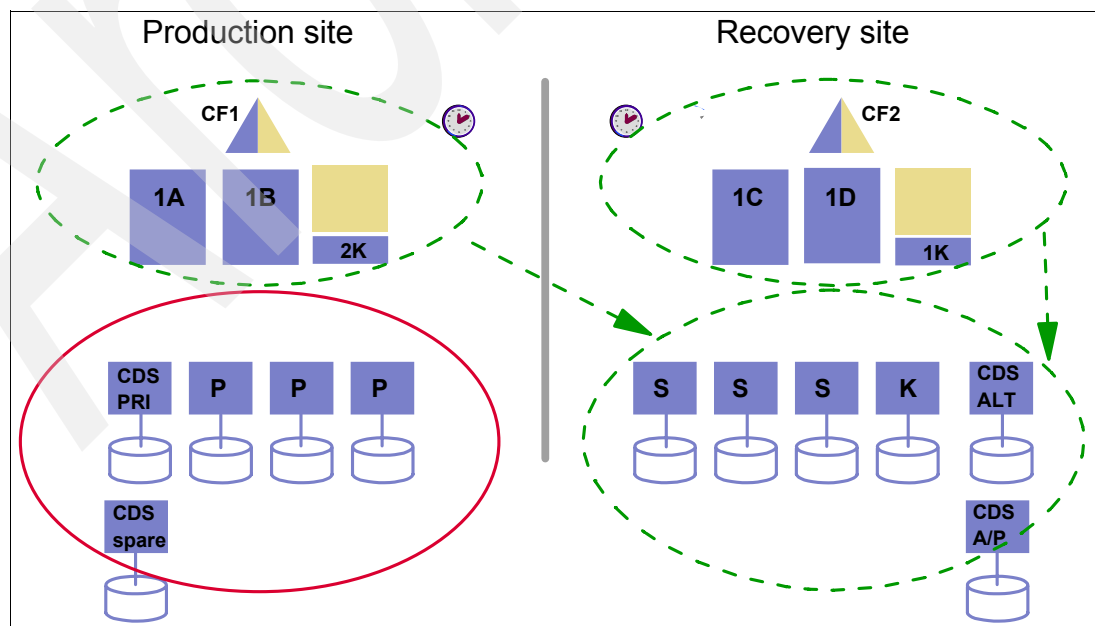


Figure 2-25 Multi-site workload cross-site Sysplex planned or unplanned disks swap

To manage a disks swap in a multi-site workload environment, GDPS performs the following operations:

1. Move the couple data set to the recovery site.
2. Swap Metro Mirror primary and secondary disks.
3. Perform the Metro Mirror Failover/Failback functions to enable changed recording and eventual return to the primary site. Upon successful completion, the primary disks are at the recovery site.
4. Optionally change the IPL table so that a subsequent IPL is at the recovery site.

Site planned or unplanned outage for multi-site workload environment

With this kind of configuration you can have Continuous Availability for production site planned or unplanned outage, not only for disks (see Figure 2-26).

While the GDPS behavior for planned and unplanned site outage is similar to that in a single-site workload Sysplex, the difference is that production images span across the sites. Production images running at the recovery site can leverage Parallel Sysplex Workload Manager to continue to run many applications almost unaffected, thus providing near Continuous Availability.

Note: GDPS cannot insure that the Coupling Facilities are time consistent with the DB2 data on frozen disks (log and database), so you have to shut down DB2 and restart it.

Upon receiving the GDPS trigger event, GDPS performs the following operations as shown in Figure 2-26.

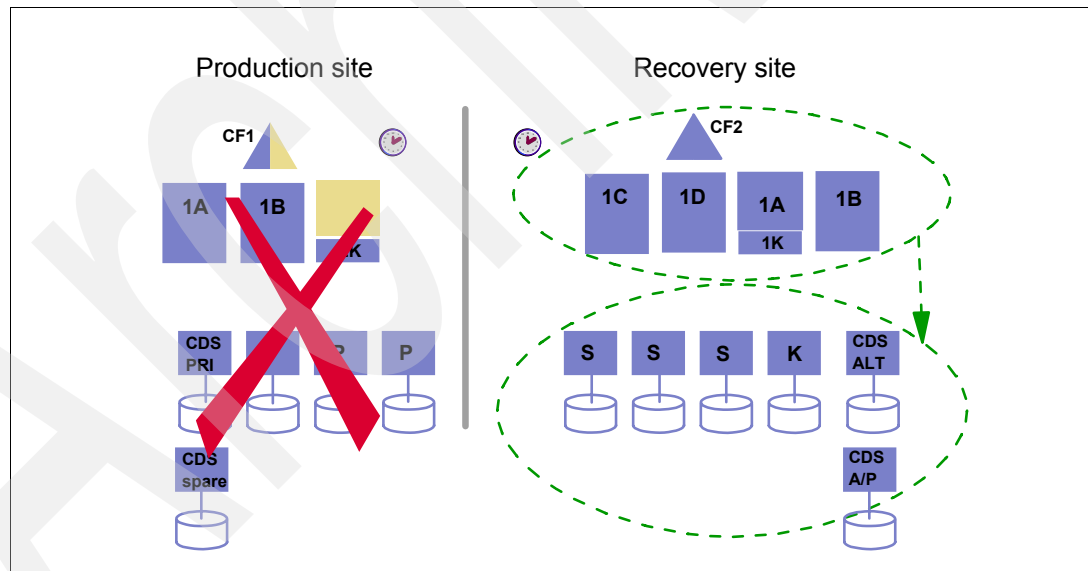


Figure 2-26 Multi-site workload cross-site Sysplex planned or unplanned outage

1. Reset all systems at the production site.
2. Swap Metro Mirror primary and secondary disks.
3. Perform the Metro Mirror Failover/Failback functions to enable changed recording and eventual return to the primary site. Upon successful completion, the primary disks are set at the recovery site.
4. Configure couple data set at the recovery site only.

5. Optionally activate CBU capacity.
6. Recover CFRM and start using CF at the recovery site (clean up CF).
7. Change the IPL table so that subsequent IPL is at the recovery site.
8. IPL systems at the recovery site.

2.2.6 GDPS extended distance support between sites

It is possible to configure GDPS/PPRC or multisite Parallel Sysplex with up to 100 km of fiber between the two sites, as shown in Figure 2-27. This feature, available via RPQ, can potentially decrease the risk of the same disaster affecting both sites, thus permitting recovery of the production applications at another site. Support for ETR (External Time Reference) links and ISC-3 (InterSystem Channel) links operating in Peer mode has been extended from the original capability of 50 km to the extended capability of 100 km.

The ETR links attach the STP to System z servers. The ISC-3 links, operating in Peer mode, and supported on all System z servers, connect z/OS systems to a Coupling Facility. The extended distance support for ETR and ISC-3 links is consistent with other cross-site link technologies that already support 100 km, such as FICON, Metro Mirror, and TS7700 Grid.

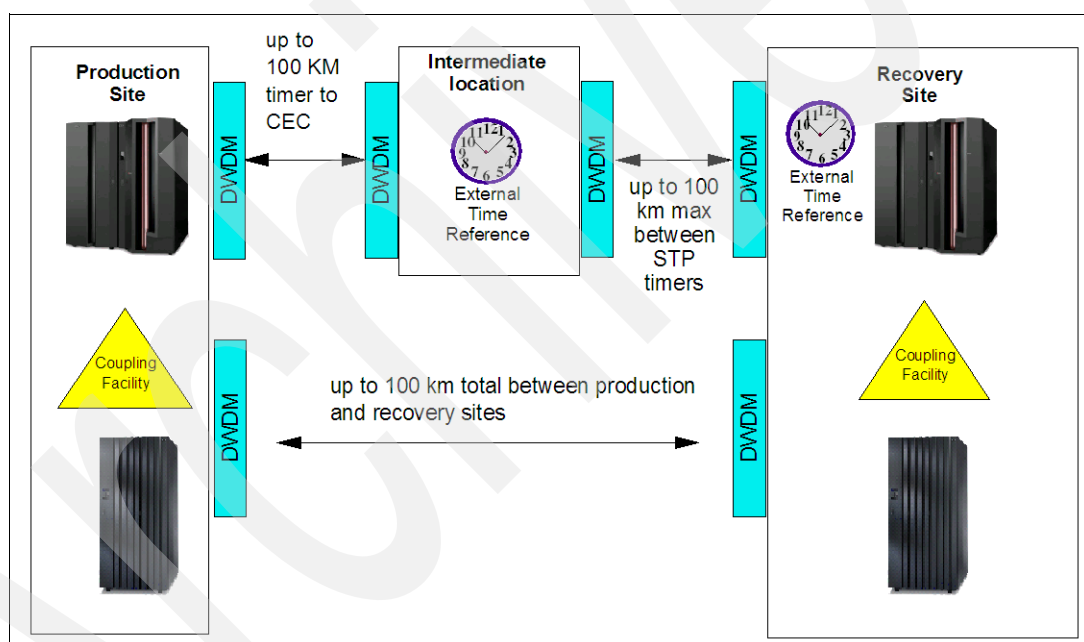


Figure 2-27 Extended distance support for GDPS/PPRC and Multisite Parallel Sysplex

The Server Time Protocol (STP) can connect at a native distance of 100 km without any requirement to place hardware at intermediate sites.

STP is detailed further in 2.1.2, “Server Time Protocol (STP)”.

GDPS/PPRC where remote site is outside the Parallel Sysplex

In this configuration, the GDPS/PPRC production systems and the controlling system are in the same production site (see Figure 2-28), and the GDPS/PPRC recovery site is located outside the Parallel Sysplex. This is typically used when the GDPS/PPRC recovery site is hosted at an off-site Business Recovery Services (BRS) center, which can be an IBM Global Technology Services center, or another client data center that is outside the Parallel Sysplex, but still within the reach of GDPS/PPRC.

Note: In this configuration, the recovery site is typically another site the client owns, or a third party disaster recovery facility such as an IBM Global Technology Services Business Recovery Center.

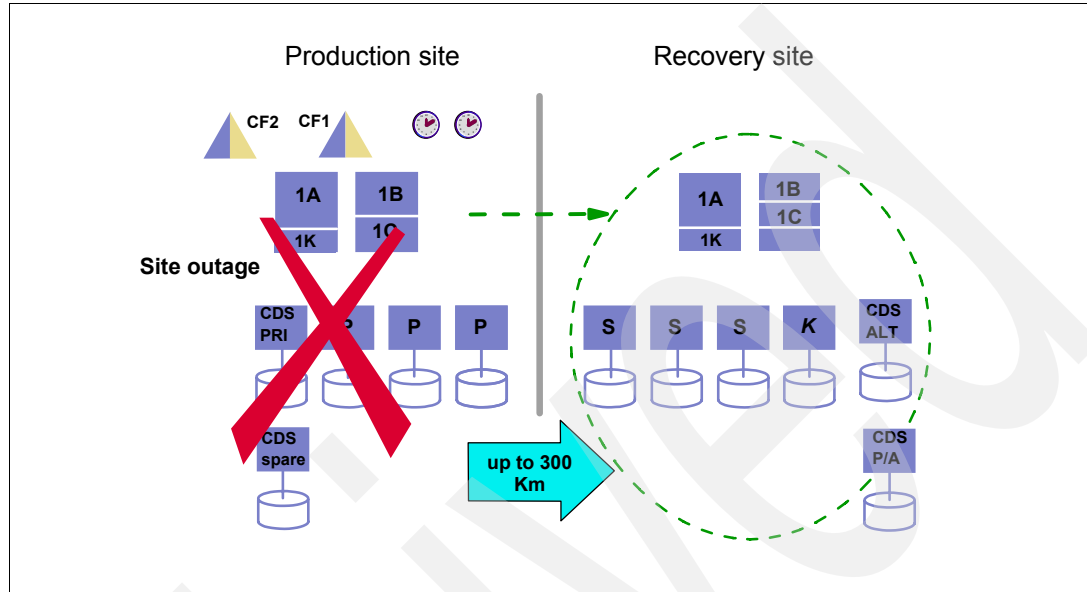


Figure 2-28 GDPS/PPRC in an IBM GTS BRS configuration

When a GDPS trigger event causes a site switch to be declared, GDPS performs the following operations:

1. Do a Consistency Group FREEZE of secondary disks.
2. Perform the Metro Mirror Failover/Failback functions to enable change recording at the remote site, thus enabling eventual return to the primary site with incremental changes only copied back.
3. IPL the controlling system at the recovery site.
4. From GDPS panels:
 - a. Recover secondary disks.
 - b. Optionally invoke CBU.
 - c. Restart systems.
5. Start applications.

The RTO for this configuration takes longer, because we have to IPL the GDPS controlling system at the recovery site before attempting to recover from disaster.

Note: Because this configuration does not use the Coupling Facility or a Sysplex Timer link to the recovery site, the distance can be extended to the Metro Mirror maximum distance of 300 km, when using Metro Mirror. Greater distances are supported on special request.

2.2.7 GDPS/XRC implementation

Now, we proceed on to discussing the technical details of the GDPS/XRC offering.

z/OS Global Mirror (XRC) review

z/OS Global Mirror (zGM) is an IBM asynchronous remote copy technology. It is a combined server and storage disk data mirroring solution, with disk microcode being driven by a System z address space running the System Data Mover (SDM), a standard function in DFSMSdfp™.

The important point is that z/OS Global Mirror assures data consistency while GDPS/XRC is providing remote site automation.

In some cases an asynchronous disaster recovery solution is more desirable than one that uses synchronous technology. Sometimes applications are too sensitive to accept the additional latency.

Note: Although GDPS/XRC is a z/OS only solution, it also supports Linux running on System z.

If your System z Linux distribution supports timestamping of writes, GDPS can manage the zGM of Linux data. In the event of a primary site disaster (or planned site switch), GDPS can automate the recovery of zGM Linux data and can restart Linux systems at the recovery site by booting them from the recovered zGM data.

Depending upon whether Linux is running natively or under VM, there are a number of different options for how the Linux system might be configured in order to handle the possibility of booting from either the XRC primary or secondary devices.

GDPS/XRC

One of the biggest differences between GDPS/PPRC and GDPS/XRC is that GDPS/XRC is not based on a multisite Parallel Sysplex. The production site is a completely independent environment from the recovery site. There are no cross site links for either coupling facilities (represented by CF) or Sysplex timers (represented as clocks).

The mission critical workload is completely based in the production site. The only cross site communication is between the primary systems in the zGM relationship and associated FICON and channel extender configuration.

Figure 2-29 shows an example of how GDPS/XRC would be configured if the protected workload was based completely within the primary configuration.

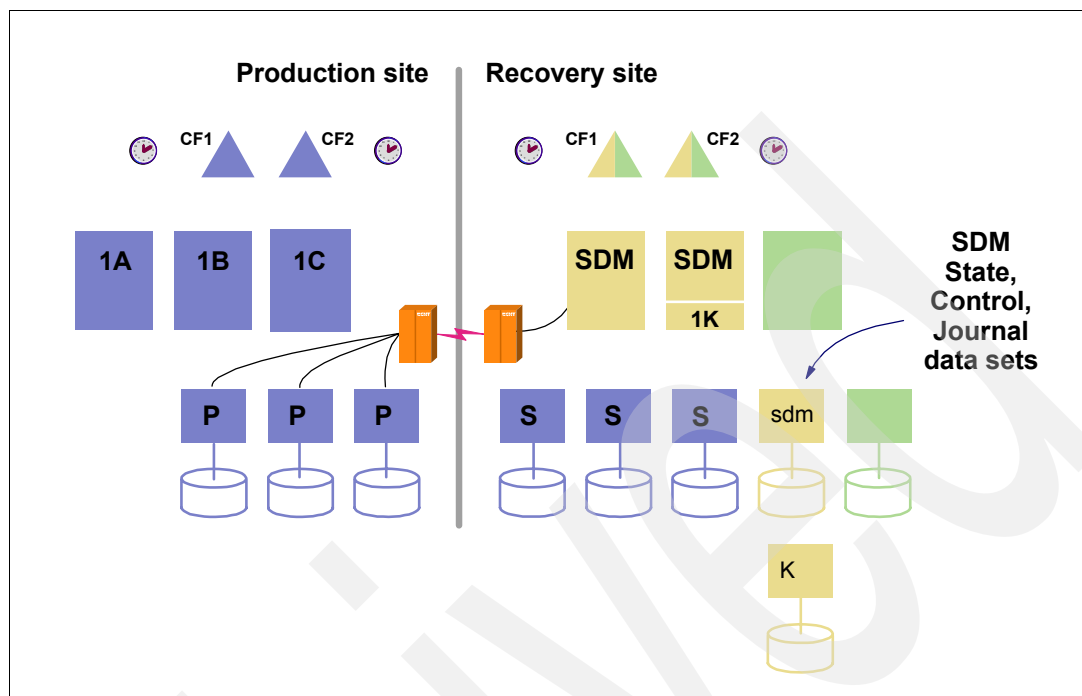


Figure 2-29 GDPS/XRC configuration

For Figure 2-29, we define the labels as follows:

- ▶ (P) indicates the primary disk systems.
- ▶ (S) indicates the secondary disk systems.
- ▶ (1) indicates the controlling system.
- ▶ (K) indicates the controlling system disks, which are isolated from the production system disks.
- ▶ (CDS) indicates the couple data sets.
- ▶ (SDM) indicates the System Data Mover disks.
- ▶ (1A-1B-1C) indicate the production LPARs.

Because the data is mirrored by using zGM (which provides data integrity), there is no requirement for a FREEZE function.

zGM itself manages consistency through Consistency Groups during the movement of data. GDPS/XRC helps to build upon this, however, by automating the recovery process and providing an improved interface for managing both zGM and FlashCopy processes. Additionally, GDPS automation can enable all necessary actions, such as re-IPLing the processors in the recovery site and restarting the production workloads.

GDPS/XRC fully supports automation of Capacity Backup, see 2.2.9, “GDPS automation of System z Capacity Backup (CBU)” on page 61. When a disaster is declared, GDPS/XRC automatically activates the additional CBU processors and memory available within that System z server. As a result, when disaster strikes, the System z processor becomes larger, capable of handling the full workload that had been processed in the production site (see Figure 2-30).

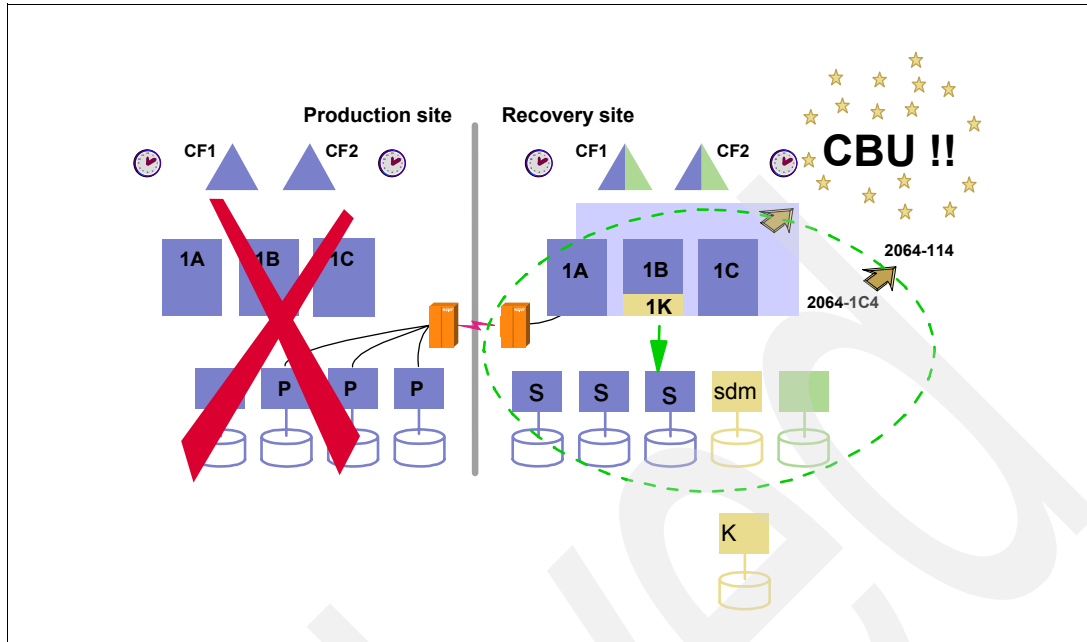


Figure 2-30 GDPS/XRC Failover with CBU

On receiving the GDPS trigger event, GDPS/XRC performs the following operations:

1. Stop zGM session.
2. Recover zGM secondary devices.
3. Deactivate SDM LPAR.
4. Optionally activate CBU.
5. Activate LPAR at the recovery site.
6. IPL systems at the recovery site.

GDPS/XRC is appropriate for businesses that require high levels of data currency in their z/OS environment, but require that data be mirrored outside of the region and cannot tolerate the application impact associated with synchronous disk mirroring.

GDPS/XRC has been proven in the very largest of z/OS mission-critical clients.

2.2.8 GDPS/Global Mirror in more detail

In the following sections, we discuss the technical details of the GDPS/Global Mirror offering.

Global Mirror review

Global Mirror is an asynchronous remote copy technology based on Global Copy and FlashCopy.

GDPS/Global Mirror

GDPS/Global Mirror provides control software to completely automate a z/OS centered Global Mirror environment. As with GDPS/XRC, GDPS/GM does not rely on parallel Sysplex technology and does not require it in either site. Moreover, as with GDPS/XRC, there are no cross site coupling links or timer links.

A sample GDPS/Global Mirror environment is shown in Figure 2-31.

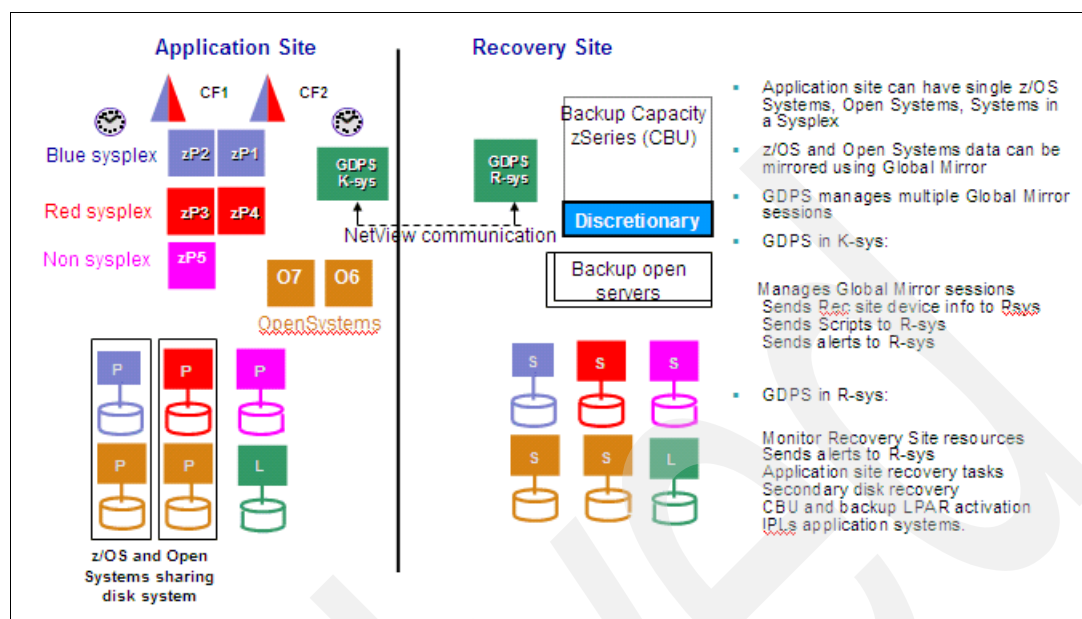


Figure 2-31 An example of a GDPS/Global Mirror environment

The mission critical workload is based in the production location, while the disposable workload can, optionally, be based in the recovery CECs.

One major difference between GDPS/Global Mirror and all other forms of GDPS is that other forms of GDPS have only a K-System and possibly a backup K-System. In the GDPS/Global Mirror environment there is both a K-System in the production site, and a R-System (R for Recovery) in the Recovery Site. In order to make sure that recovery site application costs are kept low, the R-System is not required to be part of the recovery plex; however it must be a dedicated LPAR.

Under normal operations, the K-System and the R-System monitor each others' states by sending NetView messages. GDPS/Global Mirror does not perform a site takeover without being given the command. However, it sends notification of a loss of communication and allow the environment to use any of its usual problem determination steps prior to issuing a takeover command.

Importantly, if it meets the requirements of your environment, GDPS/Global Mirror is able to serve as a single point of control for multiple Global Mirror sessions. That is to say, that although a given global mirror session can span up to 17 disk systems and provide consistency to all of those, if you want, you can have more than one global mirror session operating under the control of GDPS.

Failover

As shown in Figure 2-32, step 1, Global Mirror, like zGM, manages its own data consistency in order to provide data that can be used in a database restart, versus being forced into a database recovery process, so GDPS/Global Mirror bears no responsibility for the consistency. (This concept is also explained in Chapter 6 "Planning for Business Continuity in a heterogeneous IT environment" in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.) However, in a site switch, whether planned or unplanned, GDPS/Global Mirror Control Software takes several actions.

As shown in step 2 in Figure 2-32, the first part of any failover process involving Global Mirror is to verify the state of the FlashCopy. If the FlashCopy did not complete across the entire consistency group, then consistency cannot be assured. In order to verify this, the first action that GDPS/GM takes is to verify the state of the FlashCopy.

GDPS/Global Mirror issues the commands required to query the status of all FlashCopies in the session. If they did not form properly across all volumes (for example, if a site failure occurred during Consistency Group formation but before all FlashCopies could be established), then GDPS/Global Mirror uses the Revertible FlashCopy feature to go back to the last good FlashCopy on all volumes in the Global Mirror session.

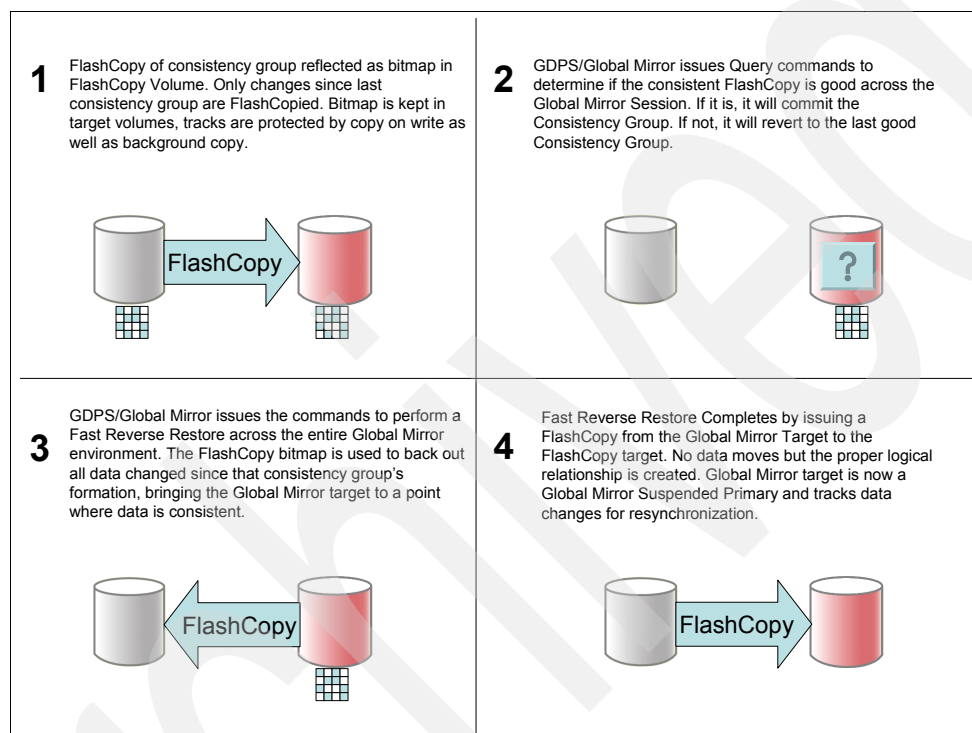


Figure 2-32 Maintaining consistency in GDPS/Global Mirror Fast Reverse Restore

Next, we have to do a Fast Reverse Restore. This is the process by which the Global Mirror target volumes become usable copies of data. As shown in Step 3 of Figure 2-32, the Global Mirror FlashCopy FlashCopies itself to the Global Mirror target volumes. As we have already determined that the FlashCopy is the most recent consistent copy of data, it is our recovery data and is required on the target volumes in order to be able to track changes and resynchronize eventually with the production site volumes when they become available again.

The process of issuing a FlashCopy from the Global Mirror FlashCopy target to its source is unique in that in order to rapidly return to the state of the FlashCopy, we keep a bitmap in the FlashCopy target that allows us to rapidly back out the changes made since the last consistent copy. The FlashCopy action applies this bitmap and restores our position to one that we know to be consistent.

In order to establish the proper logical relationship with the FlashCopy volumes, we then Flash from the Global Mirror target volumes back to the FlashCopy volumes (Step 4 of Figure 2-32). Since the volumes are identical, no data actually passes between the two, but it creates the proper logical relationship of source to target.

GDPS/Global Mirror then performs the tasks associated with reconfiguring and restarting the LPARs in the recovery site. This includes exploitation of CBU with a target recovery time of 2 hours or less. This varies based on specific implementations and, in many cases, is less. This process is detailed in Figure 2-33.

Note: Although GDPS/Global Mirror manages the mirror of Open Systems data through the Open LUN Management feature, the restart of open systems servers and applications is not a GDPS/Global Mirror function.

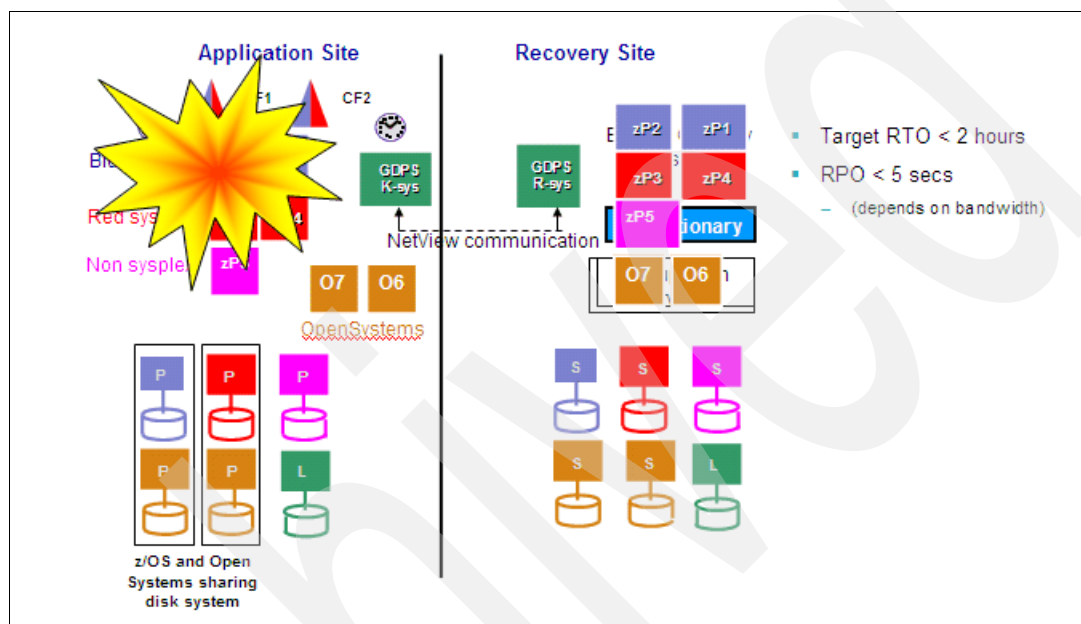


Figure 2-33 GDPS/Global Mirror reacts to a failure in the production data center

2.2.9 GDPS automation of System z Capacity Backup (CBU)

Capacity Backup (CBU) is a System z feature where an additional System z server with extra processors and memory is available. These processors and memory are normally not activated, but if there is a requirement for additional processing power, such as in a Disaster Recovery, they can be turned on.

GDPS further enhances the CBU feature by automating it. When a disaster is declared, the additional processors and memory can be activated immediately through GDPS automation, as shown in Figure 2-34.

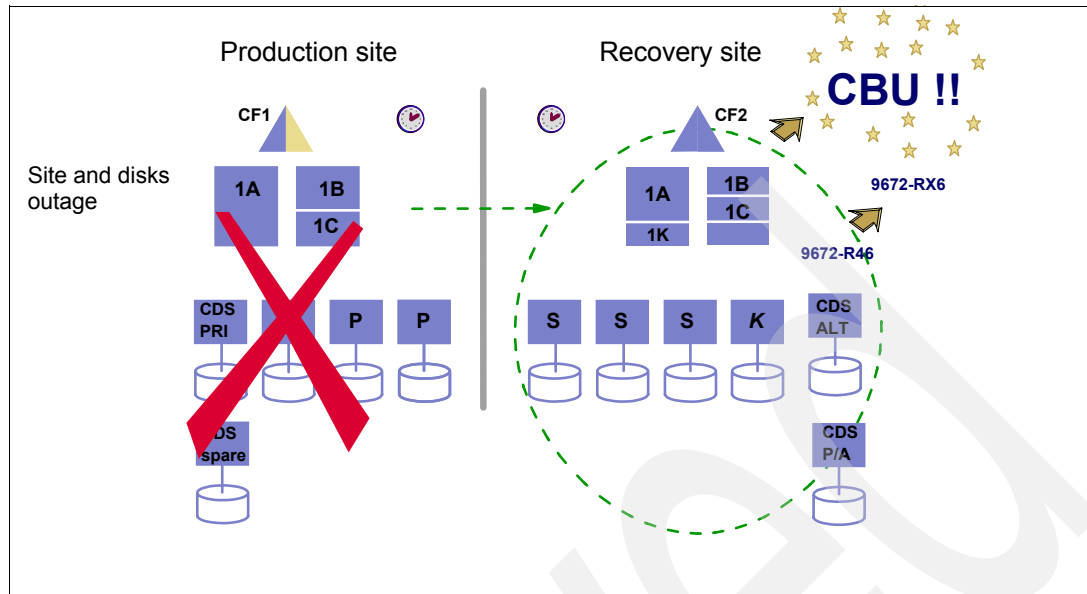


Figure 2-34 Takeover in a single site workload GDPS/PPRC environment with CBU activation

This can represent a significant cost savings because it allows a smaller server to be purchased for the recovery site. In normal operation, the System z server requires only the processing power necessary to run GDPS code and possibly SDM. Should a disaster occur, however, the additional power is immediately available to be turned on.

2.2.10 GDPS and TS7700 Grid support (GDPS/PPRC and GDPS/XRC)

The IBM TS7700 Grid is an extension to the existing IBM TS7700 Virtualization Engine™. It is designed to improve data availability by providing a dual copy function for tape virtual volumes and reducing single points of failure.

Background

Both GDPS/XRC and GDPS/PPRC support the IBM TS7700 Grid tape library; they provide control over tape mirroring with an automatic library switch. Tape control data sets have to be mirrored to the secondary site to have a coordinated failover between disks and tapes.

Figure 2-35 shows a diagram of a GDPS TS7740 Grid implementation.

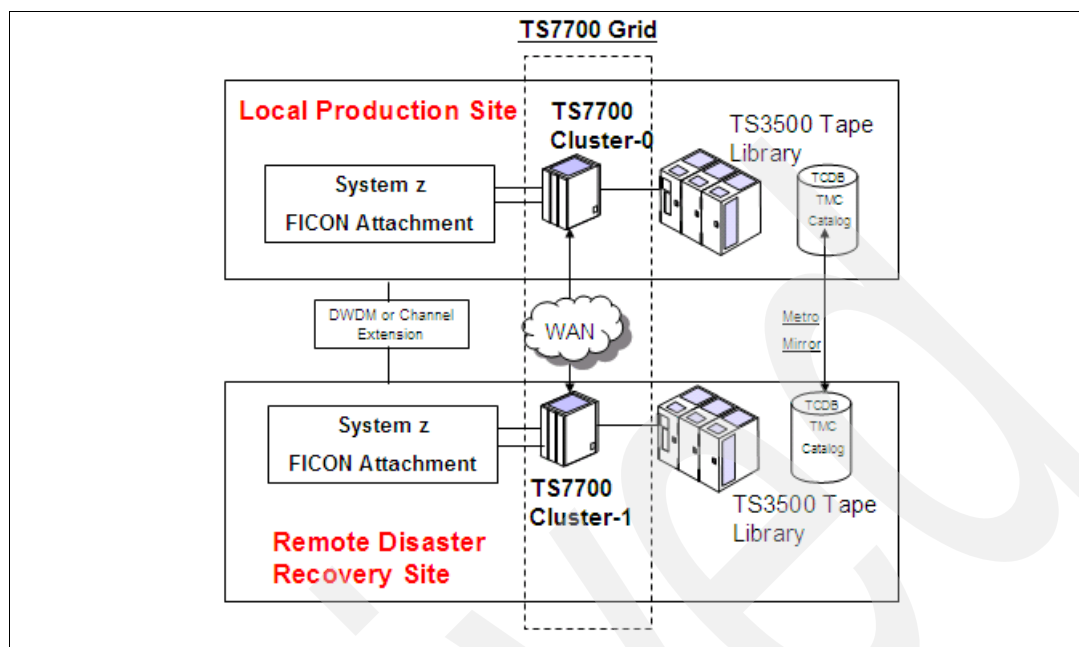


Figure 2-35 TS7740 Grid with GDPS Physical view

The highlights of the GDPS support of TS7740 Grid are as follows:

- ▶ New workload mode to support GDPS
- ▶ Consistency between tape and tape control data set if placed on mirrored disks
- ▶ Planned and unplanned tape libraries switch
- ▶ Coordinated disks and tape failback and failover in single GDPS scripts
- ▶ GDPS lists virtual volume *in-progress* dual copies at time of switch to facilitate manual adjustment of tape control data set

How GDPS TS7700 Grid support works

The TS7700 Grid currently makes dual copies for all virtual volumes based on its Consistency Policy Options for the respective cluster:

- ▶ **Run (R)**
This cluster has a valid replication of the logical volume before we provide Device End to the Rewind Unload (RUN) command from the host (this is a direct parallel to current PtP Immediate mode copy setting).
- ▶ **Deferred (D)**
This cluster gets a valid replication of the logical volume at some point in time after the job completes (same as the deferred mode in the PTP)
- ▶ **No Copy (N)**
This cluster does not receive a copy for volumes in this management class.

Note: The Consistency Policy setting of a GDPS/PPRC would have both Clusters set as RUN (R) to support the GDPS environment. In this environment, all primary data must reside in one site and all secondary copies must reside in another site to ensure recoverability in case of a disaster.

For details on IBM TS7700 Grid and GDPS operation, you can refer to the IBM Redbook, *IBM System Storage Virtualization Engine TS7700: Tape Virtualization for System z Servers*, SG24-7312.

2.2.11 GDPS FlashCopy support

GDPS supports use of FlashCopy to make tertiary and safety copies in a GDPS environment (see Figure 2-36). These FlashCopies can be GDPS-initiated or user-initiated:

- ▶ **GDPS-initiated**
 - Before re-synchronization
- ▶ **User-initiated**
 - Requested from GDPS panels
 - Facilitates disaster recovery testing
 - Enables parallel processing on tertiary copy, like backup and data mining

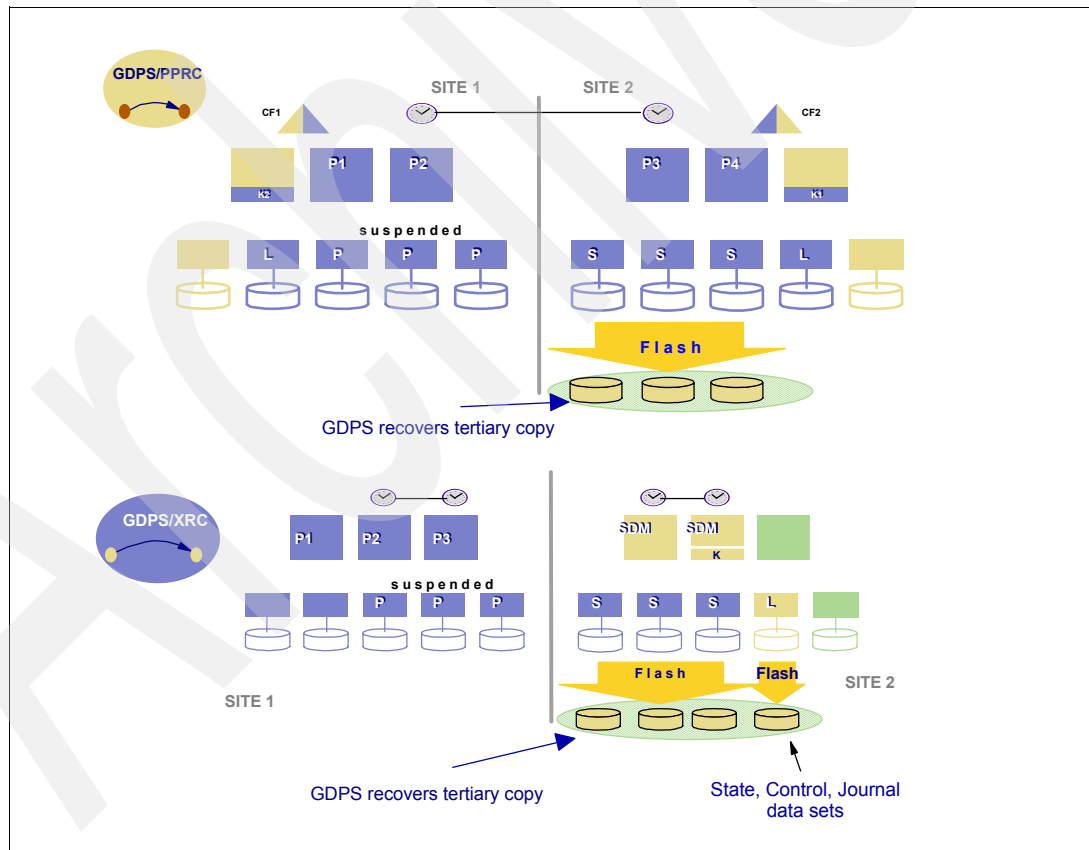


Figure 2-36 GDPS FlashCopy support

As shown in Figure 2-36, use of FlashCopy in the secondary Sysplex allows you to create instant tertiary point-in-time copies of the data as many times as necessary. This data can be accessed and modified as required but does not affect the integrity of the secondary copy or the mirroring process.

Additionally, when you want to resynchronize with the primary disk at the production site after a disaster or maintenance, having a point-in-time copy of the data available at the recovery site secures you from the unexpected, should something occur that compromises the integrity of the data on the secondary disks.

Note: During a resynch, once the resynch is started and until the volumes reach duplex state, the secondary volumes are in a *duplex pending* state, which means that they are fuzzy and not yet able to be used for recovery.

To keep a consistent copy of volumes until the resynch is complete, GDPS implements support for making a tertiary, backup copy, using a function known as *FlashCopy before Resynch*.

To test your disaster recovery procedures, you require a copy of your data that is independent of the on-going mirroring. Any tests that change data on your secondary copy could compromise its integrity should you be unfortunate enough to be the victim of a disaster during testing or the creation of a new secondary copy after the completion of a test.

Note: GDPS/PPRC HyperSwap Manager only automates the creation of FlashCopy in NOCOPY mode. RCMF does not support FlashCopy.

GDPS helps to facilitate taking copies by automating the FlashCopy commands via operator panels or scripts. The GDPS interface enables you to issue commands in the primary configuration and create point-in-time copies in the secondary configuration or automate the system to take periodic FlashCopies. As a result, it is possible to test more frequently and maintain a secure tertiary copy.

2.2.12 GDPS prerequisites

For IBM to perform GDPS services, Table 2-2 shows an overview of the prerequisites. The prerequisites listed might not contain all of the requirements for this service. For a complete list of prerequisites, consult your IBM sales representative.

Table 2-2 GDPS prerequisites

Prerequisites	RCMF	GDPS
Supported version of z/OS or z/OS, z/VM V5.1 or higher (note 1)	X	X
IBM Tivoli System Automation for Multiplatforms V1.2 or higher (note 1)		X
IBM Tivoli System Automation for z/OS V2.2 or higher (note 2)		X
IBM Tivoli NetView V5.1 or higher (note 2)	X	X
Storage system with Metro Mirror Freeze function (CGROUP Freeze/RUN) (note 3 + note 4)	X	X
zGM support with Unplanned Outage support (note 4)	X	X
Multisite Base or Parallel Sysplex (GDPS/PPRC)		X
Common Timer Reference (Sysplex Timer) for zGM	X	X

Notes on Table 2-2:

Note 1: z/VM is a prerequisite if GDPS/PPRC Multi-platform Resiliency is required.

Note 2: For GDPS/PPRC HyperSwap Manager, the following software products are required:

- ▶ IBM Tivoli System Automation for GDPS/PPRC HyperSwap Manager with NetView, V1.1 or higher, or
- ▶ IBM Tivoli NetView for z/OS V5.1 or higher, together with one of the following features:
 - IBM Tivoli System Automation for GDPS/PPRC HyperSwap Manager, V1.1 or higher
 - IBM Tivoli System Automation for z/OS V2.2 or higher

Note 3: GDPS/PPRC HyperSwap requires Metro Mirror support for Extended CQuery. GDPS/PPRC Management of Open Systems LUNs requires support for Open LUN, management of Open LUN via CKD device addresses, and Open LUN SNMP alerts.

Note 4: GDPS FlashCopy support requires FlashCopy V2 capable disk systems.

In addition, for a listing of all recommended maintenance that should be applied, as well as to review Informational APAR II12161, visit this site:

<http://www.ibm.com/servers/storage/support/solutions/bc/index.html>

2.2.13 GDPS summary

GDPS can allow a business to achieve its own continuous availability and disaster recovery goals. Through proper planning and exploitation of the IBM GDPS technology, enterprises can help protect their critical business applications from an unplanned or planned outage event.

GDPS is application independent and, therefore, can cover the client's comprehensive application environment. Note that specific software system solutions such as IMS Remote Site Recovery are very effective, but applicable to IMS applications only. When comparing GDPS with other near continuous availability and D/R solutions, you might want to ask the following questions:

- ▶ What is your desired level of improvement for your application availability?
- ▶ How does the solution handle both planned and unplanned outages?
- ▶ Which solution meets the business RTO?
- ▶ Do you want to minimize the cost of taking repetitive volume dumps, transporting the cartridges to a safe place, and keeping track of which cartridges should be moved to which location and at what time?
- ▶ What is the cost of disaster recovery drills?

The ease of planned system, disk, Remote Copy, and site reconfigurations offered by GDPS can allow your business to reduce on-site manpower and skill required for these functions. GDPS can enable a business to control its own near continuous availability and disaster recovery goals.

2.2.14 Additional GDPS information

For additional information about GDPS solutions or GDPS solution components, refer to the following Web sites and publications:

GDPS home page:

<http://www.ibm.com/systems/z/gdps/>

System z Business Resiliency Web site:

<http://www.ibm.com/systems/z/resiliency>

For an overview of how planning, implementation, and proper exploitation of System z Parallel Sysplex clustering technology can enable your System z IT infrastructure to achieve near Continuous Availability, refer to “Five Nines / Five Minutes — Achieving Near Continuous Availability” at:

<http://www.ibm.com/servers/eserver/zseries/psa/>

Additional GDPS information can be found at:

<http://www.ibm.com/services/us/index.wss/so/its/a1000189>

<http://www.ibm.com/common/ssi/fcgi-bin/ssialias?infotype=an&subtype=ca&appname=Demonstration&htmlfid=897/ENUS205-035>

<http://www.ibm.com/common/ssi/fcgi-bin/ssialias?infotype=an&subtype=ca&appname=Demonstration&htmlfid=897/ENUS305-015>

<http://www.ibm.com/servers/storage/software/sms/sdm/index.html>

You can also refer to the following IBM Redbooks:

- ▶ *IBM System Storage DS8000 Series: Copy Services in Open Environments*, SG24-6788
- ▶ *IBM System Storage DS8000 Series: Copy Services with IBM System z*, SG24-6787
- ▶ *IBM System Storage DS6000 Series: Copy Services in Open Environments*, SG24-6783
- ▶ *IBM System Storage DS6000 Series: Copy Services with IBM System z*, SG24-6782

2.3 Geographically Dispersed Open Clusters (GDOC)

Geographically Dispersed Open Clusters (GDOC) is a multivendor solution designed to protect the availability of critical applications that run on UNIX®, Windows, or Linux servers. GDOC is based on an Open Systems Cluster architecture spread across two or more sites with data mirrored between sites to provide high availability and Disaster Recovery.

2.3.1 GDOC overview

In this section we provide a concise overview of GDOC function, benefits, and applicability.

Solution description

GDOC (Figure 2-37) provides a single point of control for Business Continuity in open systems environments. It is based on the VERITAS Cluster Server. Because of this basis, it is able to provide an architecture which is capable of delivering an IBM solution, although it might use parts of the infrastructure from other vendors.

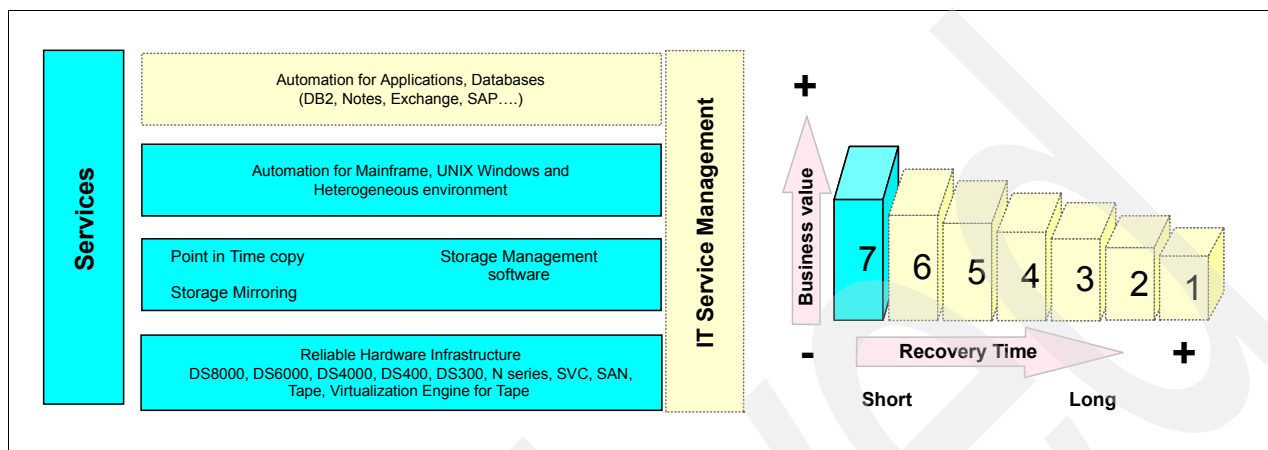


Figure 2-37 GDOC diagram

For purposes of mirroring data, GDOC supports the use of either storage based block level mirroring or server based logical volume mirroring. The hardware block level mirroring is based on the specific disk system that it is attached although it is possible to use server based mirroring based on VERITAS Volume Replicator software. In either case, it supports both synchronous and asynchronous mirroring.

Platform support

GDOC supports the following operating systems, which are supported by VERITAS software:

- ▶ IBM AIX
- ▶ Linux
- ▶ Microsoft® Windows
- ▶ SUN Solaris™
- ▶ HP UX

Additional information about required operating system levels can be found on the VERITAS Web site at:

<http://www.veritas.com>

Solution components

A GDOC solution is based on a combination of Global Technology Services services together with VERITAS software. The services are split between a consulting and planning phase followed by an implementation and deployment phase.

GDOC consulting and planning

The first phase consists of assessing business availability and recovery requirements. During this stage, the desired solution is either defined or validated. It determines the steps required to implement the desired solution.

This phase can be performed through a GDOC Technology Consulting Workshop engagement. It results in a high-level assessment and a roadmap for solution implementation and deployment.

GDOC implementation and deployment

The second phase consists of:

- ▶ GDOC conceptual, logical, and physical design
- ▶ GDOC solution build and test
- ▶ GDOC prototype build, if required
- ▶ Pilot deployment, if required
- ▶ GDOC solution roll-out and acceptance by the client

VERITAS software components

There is no predefined set of VERITAS software components because the solution can be adapted to specific client requirements. These are the main components used:

- ▶ VERITAS Cluster Server (VCS):
 - VERITAS Cluster Server Hardware Replication Agent for IBM PPRC
 - VERITAS Cluster Server Global Cluster Option
 - VERITAS CommandCentral Availability, which was VERITAS Global Cluster Manager
- ▶ VERITAS Storage Foundation (optional):
 - VERITAS Volume Manager
 - VERITAS Volume Replicator
 - VERITAS File System (UNIX)

Additional software modules can be added to these basic components to support specific environments such as SAP or DB2.

2.3.2 GDOC in greater detail

A GDOC solution controls resource and application availability and initiates application failover to alternative servers when the necessity arises. Replication can be performed either at the server level using VERITAS Volume Replicator for software replication, or at the hardware level using storage system remote replication functions.

Application availability and failover are controlled using VERITAS Cluster Server (VCS) and specialized modules for applications such as SAP and DB2 and others. Additional cluster modules are available. An example is VERITAS Cluster Server Hardware Replication Agent for PPRC that is available to control DS6000 and DS8000 Metro Mirror based solutions. Site failover is controlled by VERITAS Cluster Server. VERITAS CommandCentral Availability provides cluster centralization, monitoring, and reporting capabilities.

Figure 2-38 shows a diagram of GDOC functionality.

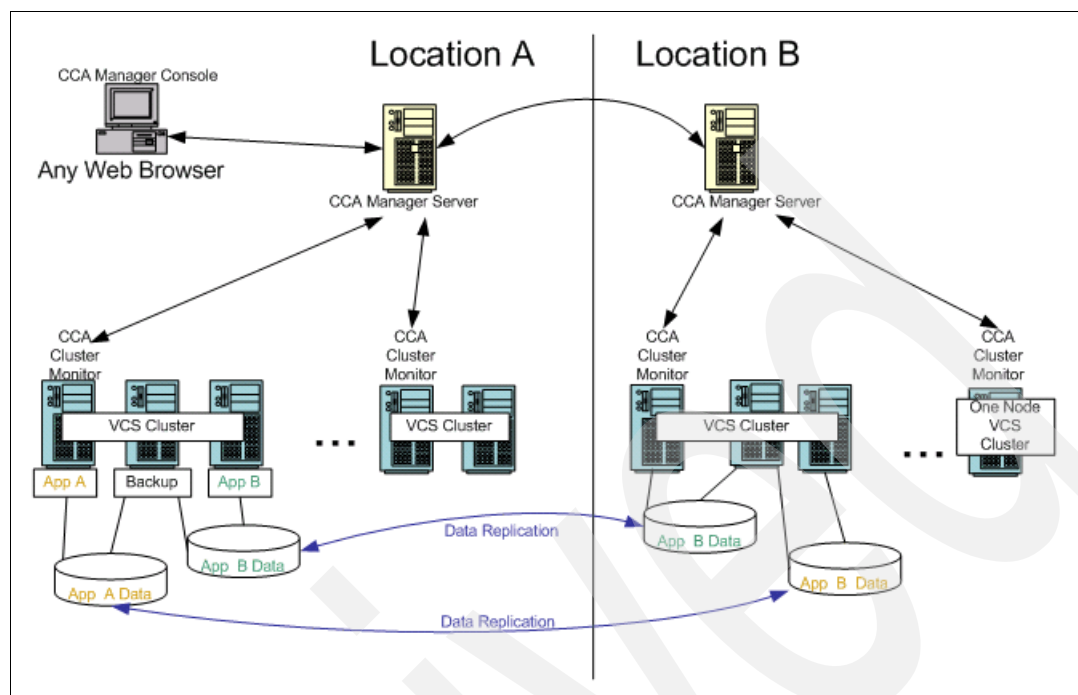


Figure 2-38 GDOC functional overview

Normally applications run on the primary site. Application data is contained in an external storage system. This allows the application to fail over to a secondary server at the primary site under VCS control. The application accesses the data in the primary site's storage system.

Data is replicated continuously between the storage systems located at the primary and secondary sites using data replication functions such as DS6000/DS8000 Metro Mirror or VERITAS Volume Replicator. VERITAS Volume Replicator maintains dependent write consistency, write order fidelity using VERITAS terminology, at the individual server level. If you have a requirement to maintain dependent write consistency across multiple servers, then you must consolidate the storage in one or more external storage systems and use Metro Mirror or Global Mirror solutions. These solutions can guarantee dependent write consistency across multiple servers and storage systems.

The VERITAS CommandCentral Availability, the follow-on product to VERITAS Global Cluster Manager, controls failover of applications to the secondary site. This is shown in the picture as a site migration. A site migration typically requires operator intervention, all that is required is an operator confirmation to perform the site failover.

This solution extends the local High Availability (HA) model to many sites. Dispersed clusters and sites are linked by public carrier over a wide area network and SAN. Each site is aware of the configuration and state of all of the sites (global cluster management). Complete site failover occurs in the event of a catastrophe, and the basis for this failover is the replicated data. This solution provides global availability management from a single console.

Additional information

This offering is delivered by Global Technology Services (GTS). For more information, contact your IBM sales representative.

2.4 HACMP/XD

HACMP/XD H is a software solution for AIX to mirror a logical disk volume over a dedicated IP Network. This disk volume is called a Geo MirrorDevice (GMD).

2.4.1 HACMP/XD overview

In this section we provide an overview of HACMP/XD, its function, and its benefits. Figure 2-39 shows a diagram of HACMP/XD.

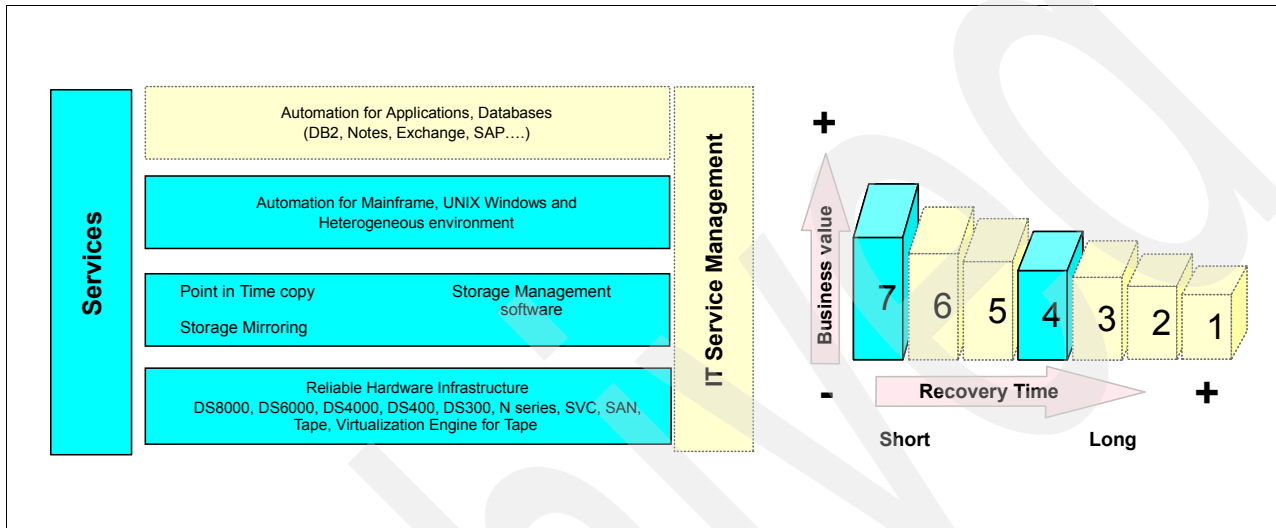


Figure 2-39 HACMP/XD

Solution description

HACMP/XD is for AIX/System p environments that have business requirements to maintain continuous availability of applications in a multi-site environment within metropolitan distances. This is accomplished by combining High Availability Cluster Multiprocessors (HACMP) and HACMP/XD (Extended Distance), IBM System Storage Disk Systems (Including DS6000, DS8000, and SAN Volume Controller), and IBM Metro Mirror.

Each of these components serves a specific purpose:

- ▶ **HACMP** is the high availability clustering software for AIX environments. In Chapter 9 “High Availability Clusters and Database Applications in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547, it is referred to as a “Shared Nothing” cluster. In a typical HACMP environment, the nodes are all attached to a common disk system and can be active/active or active/passive. In either case, a failure on an active node triggers a failover of processors and applications to the surviving node. HACMP is typically used to protect availability within a site, while the HACMP/XD component deals with failing to an alternate recovery site.
- ▶ **HACMP/XD** is an extension to HACMP which enables the process of failing over to an alternate site. This can be done through storage hardware based mirroring or through server based IP or GLVM Mirroring
- ▶ **IBM Disk Systems** provide a choice of highly reliable and scalable storage including the DS6000, DS8000, and SAN Volume Controller (SVC). More information about these products is given in Chapter 7, “IBM System Storage DS6000, DS8000, and ESS” on page 253 and 11.2, “IBM System Storage SAN Volume Controller” on page 365.

- ▶ **Metro Mirror** is the IBM name for Synchronous Mirroring technologies. In the DS6000 and DS8000, we support a maximum distance of 300 km for mirror links. (Greater distances are supported on special request.) In the SVC, we support a maximum distance of 100km. This function is more thoroughly detailed in 7.7.3, “Remote Mirror and Copy (Peer-to-Peer Remote Copy)” and 11.2, “IBM System Storage SAN Volume Controller” on page 365.

The specific requirements for HACMP/XD are:

- ▶ System p Server
- ▶ AIX
- ▶ HACMP/XD and HACMP
- ▶ Supported disk storage system - currently IBM System Storage DS8000, DS6000, or SVC

Additional information

You can find further information about these topics on the IBM Web site:

<http://www.ibm.com/systems/p/ha>

For further details on HACMP/XD with Metro Mirror, see *HACMP/XD for Metro Mirror: Planning and Administration Guide*, SC23-4863.

For more information, contact your IBM Sales Representative.

2.4.2 HACMP/XD in greater detail

HACMP/XD for Metro Mirror provides an automated disaster recovery solution, and it supports application-level failover/fallback. Additionally, for planned software or hardware maintenance activities, this solution can provide for failover procedures to move applications and data to the other site (secondary cluster node) as maintenance is performed on one cluster node at a time; the applications can remain operative on the other node.

HACMP/XD uses Metro Mirror to replicate AIX volume groups to a remote location up to 300 km away. The nodes in each site use the same volume groups, but each one accesses them only from their local disk systems.

The definition for a Metro Mirror replicated resource contains the volume identifier and the name of the disk system. HACMP recognizes which volumes mirror each other for each Metro Mirror replicated resource.

Note: Metro Mirror copies volume information, including the PVID, from one volume in a Metro Mirror pair to the other. The volumes at both sites contain the same logical volumes and must therefore be imported with the same volume group name. This also allows single-name entries in a resource group definition.

Resource groups that include Metro Mirror replicated resources

The definition for a Metro Mirror replicated resource contains the volume identifier and the name of the disk system. HACMP recognizes which volumes mirror each other for each Metro Mirror replicated resource.

An HACMP resource group is a collection of resources that comprise the operating environment for an application. Applications, as resources in a resource group, are made highly available. Resource group management policies direct which node hosts the resource group during normal operation and when the host node fails or goes offline. With HACMP/XD for Metro Mirror, resource group configuration is the same as for other resource groups.

In addition, the resource group includes:

- ▶ A shared volume group and Metro Mirror replicated resources associated with the individual volumes in the volume group
- ▶ Nodes that all have access to the HMCplex
- ▶ An intersite management policy to handle a resource group during site recovery

HACMP/XD sites

HACMP/XD supports two sites: a primary (active) site and a secondary (standby) site. The Inter-Site Management Policy for a resource group directs how a resource group and its resources failover in response to an outage and how they fallback if configured to do so. For each resource group, one site is an active production site and the other a backup site. If the nodes at the active production site become unavailable, the backup site becomes the active production site. For HACMP/XD, each site contains at least one IBM disk system and the nodes attached to it.

Resource group management policies

Resource groups have two types of management policies:

- ▶ Resource group management policies determine failover behavior if a node becomes unavailable.
- ▶ Site management policies determine failover behavior if all of the nodes at a site are not available.

HACMP/XD requires two HACMP sites for use within a resource group to control which volume in a Metro Mirror pair a node can access. Although nodes at both sites can access a volume group, access is only permitted the source volume in a Metro Mirror pair.

This prevents nodes at different sites from accessing the same volume group at the same time. Within a resource group, the nodes at one site can handle the Metro Mirror replicated resources differently than the nodes at the other site, especially in cases where the states (suspended or full-duplex) of the volumes are different at the two sites.

Failover and fallback

HACMP/XD handles the automation of failover from one site to another in response to an outage at a production site, minimizing recovery time. When a site fails, the resource group configuration determines whether source volumes are accessible from the secondary site. HACMP/XD uses the Metro Mirror Failover and Fallback functions to provide automatic failover and recovery of a pair of Metro Mirror volumes. The Metro Mirror Failover and Fallback features help to reduce downtime and recovery time during recovery by simplifying the tasks required to make the disk available.

HACMP/XD automates application recovery by managing:

- ▶ Failover of nodes within a site based on node priority (as identified in the nodelist for a resource group) by utilizing HACMP clusters within that site.
- ▶ Failover between sites (as specified by the site management policy for a resource group) by utilizing HACMP/XD automation between the sites
- ▶ Fallback of a resource group or site as configured.

When an application is running on an active production site:

- ▶ Updates to the application data are made to the disks associated with the active production site.

- Data is mirrored in the secondary volumes through Metro Mirror.

If the node or the disks at the production site become unavailable:

- The application moves to a server at the backup site.
- The application continues operation using the mirrored copy of the data.

When the initial production site becomes active again, resource group and site management policies determine whether or not the application moves back to the previous site:

- The direction of mirroring can be reversed.
- The application can be stopped and restarted on another node.

HACMP/XD management of Metro Mirror pairs

HACMP/XD performs the following functions:

- It manages the failover and re-synchronization of the Metro Mirror pairs.
- It issues commands directly to the disk systems.

If nodes at one site fail, go offline, or have Cluster Services stopped on them, the nodes at the surviving site send commands to the disk systems to execute a Metro Mirror Failover task which results in:

- Allowing the nodes that are taking over to access the target volume
- Putting the Metro Mirror pair into a suspended state

When the nodes at the failed site recover, the nodes at the surviving site send commands to the disk system to execute a Metro Mirror Failback task that re-synchronizes the volumes.

Figure 2-40 shows a sample two-site configuration using Metro Mirror between the two disk systems with HACMP/XD. This example shows a single production site and a single recovery site and is based on the assumption that all the primary volumes are on the production site.

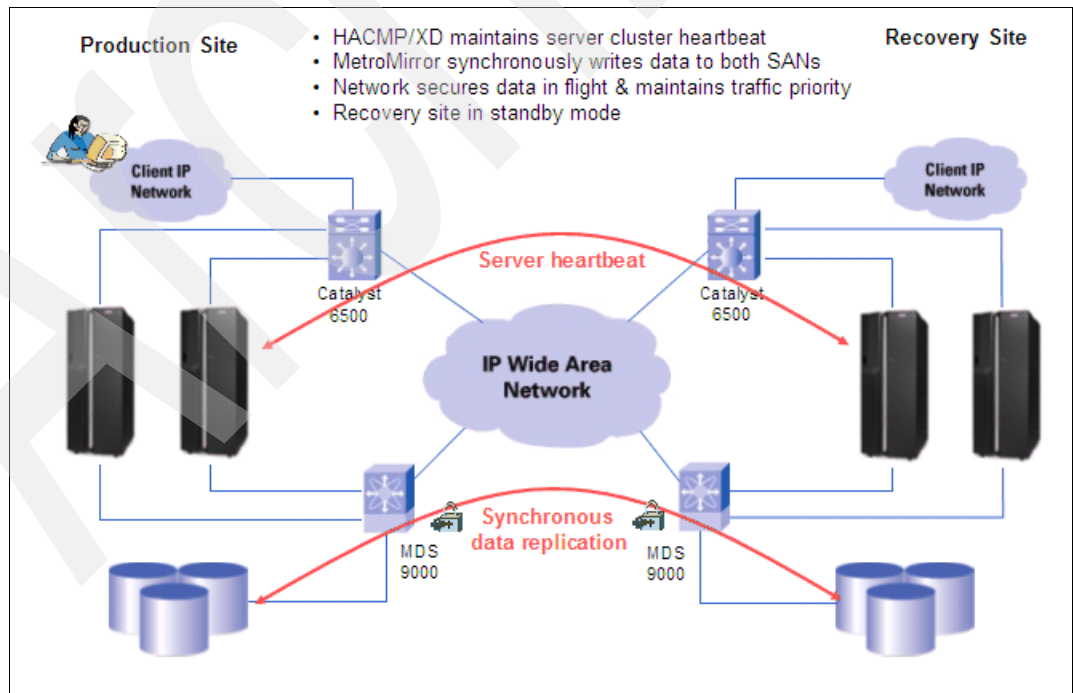


Figure 2-40 A sample configuration using HACMP/XD

Limitations

Disk heartbeat is not supported across sites, they can only be used locally within a site.

Summary

HACMP/XD takes advantage of the following components to reduce downtime and recovery time during Disaster Recovery:

- ▶ Metro Mirror Failover and Failback functions
- ▶ HACMP cluster management

HACMP/XD provides high availability and Disaster Recovery through:

- ▶ Automatic fallover of Metro Mirror protected volume pairs between nodes within a site
- ▶ Automatic fallover of Metro Mirror protected volume pairs between sites
- ▶ Automatic recovery/reintegration of Metro Mirror protected volume pairs between sites
- ▶ Support for user-defined policy-based resource groups

Note: A replicated resource is an HACMP resource that has a primary and secondary instance that is copied from one site to another.

- ▶ Support for the following Inter-Site Management Policies for resource groups:
 - Prefer Primary Site (*Recommended*). In a two-site configuration, replicated resources at startup are on the site with the higher priority, fall over to the other site and then fall back to the site with the higher priority.
 - Online on Either Site. Replicated resources are on either site at startup, fall over to the other site and remain on that site after fallover.

Note: An HACMP resource group is a collection of resources that comprise the operating environment for an application.

- ▶ Support for VERBOSE_LOGGING and EVENT_EMULATION
- ▶ Support for the IBM Subsystem Device Driver (SDD) multipathing software
- ▶ Support for cluster verification and synchronization
- ▶ Support for C-SPOC facility. C-SPOC makes cluster management easier, as it allows you to make changes to shared volume groups, users and groups across the cluster from a single node. The changes are propagated transparently to other cluster nodes.

2.4.3 High Availability Cluster Multi-Processing (HACMP) for AIX

The IBM tool for building AIX-based mission-critical computing platforms is the HACMP software. The HACMP software ensures that critical resources, such as applications, are available for processing. HACMP has two major components: high availability (HA) and cluster multi-processing (CMP).

The primary reason to create HACMP clusters is to provide a highly available environment for mission-critical applications. For example, an HACMP cluster could run a database server program which serves client applications. The clients send queries to the server program which responds to their requests by accessing a database stored on a shared external disk. In an HACMP cluster, to ensure the availability of these applications, the applications are put under HACMP control.

HACMP takes measures to ensure that the applications remain available to client processes, even if a component in a cluster fails. To ensure availability in case of a component failure, HACMP moves the application (along with the resources that ensure access to the application) to another node in the cluster.

High availability and hardware availability

High availability is sometimes confused with simple hardware availability. Fault tolerant, redundant systems (such as RAID) and dynamic switching technologies (such as Dynamic LPAR) provide recovery of certain hardware failures, but do not provide the full scope of error detection and recovery required to keep a complex application highly available.

Modern applications require access to all of these elements:

- ▶ Nodes (CPU, memory)
- ▶ Network interfaces (including external devices in the network topology)
- ▶ Disk or storage devices

Recent surveys of the causes of downtime show that actual hardware failures account for only a small percentage of unplanned outages. Other contributing factors include:

- ▶ Operator errors
- ▶ Environmental problems
- ▶ Application and operating system errors

Reliable and recoverable hardware simply cannot protect against failures of all these different aspects of the configuration. Keeping these varied elements and therefore the application, highly available requires:

- ▶ Thorough and complete planning of the physical and logical procedures for access and operation of the resources on which the application depends. These procedures help to avoid failures in the first place.
- ▶ A monitoring and recovery package which automates the detection and recovery from errors.
- ▶ A well-controlled process for maintaining the hardware and software aspects of the cluster configuration while keeping the application available.

Role of HACMP

The HACMP planning process and documentation include tips and advice on the best practices for installing and maintaining a highly available HACMP cluster. Once the cluster is operational, HACMP provides the automated monitoring and recovery for all the resources on which the application depends. HACMP provides a full set of tools for maintaining the cluster while keeping the application available to clients.

HACMP lets you:

- ▶ Quickly and easily set up a basic two-node HACMP cluster by using the Two-Node Cluster Configuration Assistant.
- ▶ Set up an HACMP environment using online planning worksheets to simplify the initial planning and setup.
- ▶ Test the HACMP configuration by using the Cluster Test Tool. You can evaluate how a cluster behaves under a set of specified circumstances, such as when a node becomes inaccessible, a network becomes inaccessible, and so forth.
- ▶ Ensure high availability of applications by eliminating single points of failure in an HACMP environment.
- ▶ Leverage the high availability features available in AIX.

- ▶ Manage how a cluster handles component failures.
- ▶ Secure cluster communications.
- ▶ Set up fast disk takeover for volume groups managed by the Logical Volume Manager (LVM).
- ▶ Monitor HACMP components and diagnose problems that might occur.

Cluster multi-processing is a group of loosely coupled machines networked together, sharing disk resources. In a cluster, multiple server machines cooperate to provide a set of services or resources to clients. Clustering two or more servers to back up critical applications is a cost-effective high availability option. You can use more of your site's computing power while ensuring that critical applications resume operations after a minimal interruption caused by a hardware or software failure. Cluster multi-processing also provides a gradual, scalable growth path. It is easy to add a processor to the cluster to share the growing workload. You can also upgrade one or more of the processors in the cluster to a more powerful model.

HACMP provides an environment that ensures that mission-critical applications can recover quickly from hardware and software failures. HACMP combines custom software with industry-standard hardware to minimize downtime by quickly restoring services when a system, component, or application fails. While not instantaneous, the restoration of service is rapid, usually 30 to 300 seconds.

A relevant goal of high availability clustering software is to minimize, or ideally, eliminate, the necessity to take resources out of service during maintenance and reconfiguration activities. HACMP software optimizes availability by allowing running clusters to be dynamically reconfigured. Most routine cluster maintenance tasks, such as adding or removing a node or changing the priority of nodes participating in a resource group, can be applied to an active cluster without stopping and restarting cluster services. In addition, an HACMP cluster can remain online while making configuration changes by using the Cluster Single Point of Control (C-SPOC) facility. C-SPOC makes cluster management easier, by allowing changes to shared volume groups, users and groups across the cluster from a single node. The changes are propagated transparently to other cluster nodes.

Physical components of an HACMP cluster

HACMP high availability works by identifying a set of resources essential to uninterrupted processing and by defining a protocol that nodes use to collaborate to ensure that these resources are available. HACMP defines relationships among cooperating processors where one processor provides the service offered by a peer should the peer be unable to do so.

An HACMP cluster consists of the following physical components:

- ▶ Nodes
- ▶ Shared external disk devices
- ▶ Networks
- ▶ Network interfaces
- ▶ Clients
- ▶ Sites

With HACMP, physical components can be combined into a wide range of cluster configurations, providing flexibility in building a cluster that meets processing requirements. Figure 2-41 shows one example of an HACMP cluster. Other HACMP clusters could look very different, depending on the number of processors, the choice of networking and disk technologies and so on. Note that many of the components in the figure are duplicated. This duplication helps to eliminate single points-of-failure. As mentioned earlier, these failures might be due to hardware failures, but more likely to human error.

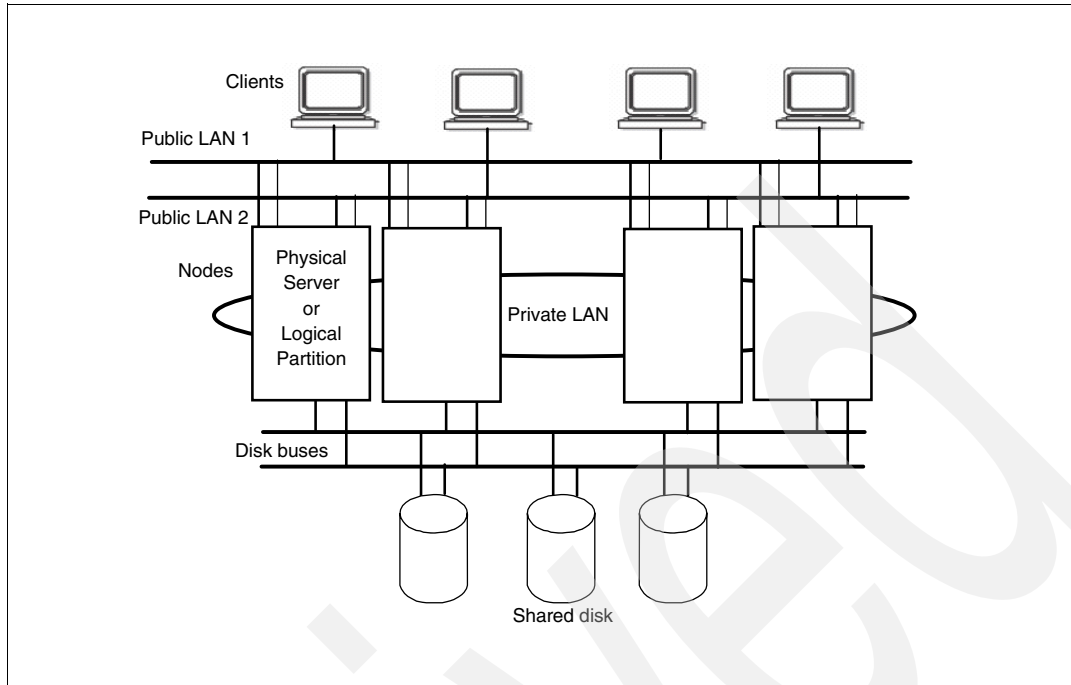


Figure 2-41 Sample HACMP cluster

Nodes

Nodes form the core of an HACMP cluster. A node is a processor that runs AIX, the HACMP software, and the application software. In an HACMP cluster, up to 32 System p servers (using LPARS, or standalone systems, or a combination of these) cooperate to provide a set of services or resources to other entities.

Two types of nodes are defined:

- ▶ *Server nodes* form the core of an HACMP cluster. Server nodes run services or *back-end* applications that access data on the shared external disks.
- ▶ *Client nodes* run *front end* applications that retrieve data from the services provided by the server nodes. Client nodes can run HACMP to monitor the health of the nodes and to react to failures.

Shared external disk devices

Each node has access to one or more shared external disk devices. A shared external disk device is a disk physically connected to multiple nodes. The shared disk stores mission-critical data, typically mirrored or RAID-configured for data redundancy. A node in an HACMP cluster must also have internal disks that store the operating system and application binaries, but these disks are not shared. Depending on the type of disk used, HACMP supports the following types of access to shared external disk devices — non-concurrent access and concurrent access.

- ▶ In non-concurrent access environments, only one connection is active at any given time and the node with the active connection owns the disk. When a node fails, disk takeover occurs; the node that currently owns the disk leaves the cluster and a surviving node assumes ownership of the shared disk.
- ▶ In concurrent access environments, the shared disks are actively connected to more than one node simultaneously. Therefore, when a node fails, disk takeover is not required.

Networks

As an independent, layered component of AIX, HACMP can work with any TCP/IP-based network. Nodes in an HACMP cluster use the network to:

- ▶ Allow clients to access the cluster nodes
- ▶ Enable cluster nodes to exchange heartbeat messages
- ▶ Serialize access to data (in concurrent access environments)

The HACMP software defines two types of communication networks: TCP/IP-based networks, and device-based networks. A TCP/IP-based network is a standard TCP/IP network. A device-based network uses communication through non-TCP/IP systems.

- ▶ *TCP/IP-based network.* Connects two or more server nodes and optionally allows client access to these cluster nodes, using the TCP/IP protocol. Ethernet, Token-Ring, ATM, HP Switch and SP Switch networks are defined as TCP/IP-based networks.
- ▶ *Device-based network.* Provides a point-to-point connection between two cluster nodes for HACMP control messages and heartbeat traffic. Device-based networks do not use the TCP/IP protocol and therefore continue to provide communications between nodes even if the TCP/IP subsystem on a server node fails. Target mode SCSI devices, Target Mode SSA devices, disk heartbeat devices, or RS232 point-to-point devices are defined as device-based networks.

Clients

A client is a processor that can access the nodes in a cluster over a LAN. Clients each run a *front end* or client application that queries the server application running on the cluster node.

Sites

You can define a group of one or more server nodes as belonging to a site. The site becomes a component, like a node or a network, that is known to HACMP. HACMP supports clusters divided into two sites.

Using sites, you can configure the cross-site LVM mirroring. You configure logical volume mirrors between physical volumes in separate storage arrays and specify to HACMP which physical volumes are located at each site. Later when you use C-SPOC to create new logical volumes, HACMP automatically displays the site location of each defined physical volume, making it easier to select volumes from different sites for LVM mirrors.

Obviously this concept is also used by the HACMP/XD feature. Each site can be a backup data center for the other, maintaining an updated copy of essential data and running key applications. If a disaster disables one site, the data is available within minutes at the other site. The HACMP/XD solution thus increases the level of availability provided by HACMP by enabling it to recognize and handle a site failure, to continue processing even though one of the sites has failed and to reintegrate the failed site back into the cluster. If sites are configured in the cluster, the Resource Group Management utility can be used to bring a resource online, take it offline, or move it to another node only within the boundaries of a site.

Elimination of single points of failure

A single point of failure exists when a critical cluster function is provided by a single component. If that component fails, the cluster has no other way to provide that function and essential services become unavailable.

To be highly available, a cluster must have no single point of failure. While the goal is to eliminate all single points of failure, compromises might have to be made. There is usually a cost associated with eliminating a single point of failure. For example, redundant hardware increases cost. The cost of eliminating a single point of failure should be compared to the cost of losing services should that component fail.

HACMP helps in eliminating each of these resources as a single point-of-failure:

- ▶ Nodes
- ▶ Applications
- ▶ Networks and network interfaces
- ▶ Disks and disk adapters

Although it might be easier to understand that adding duplicate hardware, such as a node with the appropriate software, eliminates a single point-of-failure, what is not so obvious is how an application is eliminated as a single point-of-failure. You can make an application highly available by using:

- ▶ An application server
- ▶ Cluster control
- ▶ Application monitors
- ▶ Application Availability Analysis Tool

Putting the application under HACMP control requires user-written scripts to start and stop the application. Then, an application server resource is defined that includes these scripts and the application server resource, typically along with a highly available IP address and shared disk on which the application data resides, into a resource group. To start the resource group, HACMP ensures that the proper node has the IP address and control of the disk, then starts the application using the application server start script. To move a resource group, it first stops the application using the stop script (if the node is running) and releases the resources, then starts the resource group on another node. By defining an application server, HACMP can start the application on the takeover node when a fallover occurs.

Note: Application takeover is usually associated with IP address takeover. If the node restarting the application also acquires the IP service address on the failed node, the clients only have to reconnect to the same server IP address. If the IP address was not taken over, the client has to connect to the new server IP address to continue accessing the application.

Additionally, you can use the AIX System Resource Controller (SRC) to monitor for the presence or absence of an application daemon and to respond accordingly.

Application monitors

You can also configure an application monitor to check for process failure or other application failures and automatically take action to restart the application.

Application Availability Analysis

The Application Availability Analysis tool measures the exact amount of time that any of your applications has been available. HACMP collects, time-stamps, and logs extensive information about the applications you choose to monitor with this tool. You can select a time period and the tool displays uptime and downtime statistics for a specific application during that period.

Additional information

For further information about HACMP architecture and concepts, see these publications:

- ▶ *HACMP for AIX: Concepts and Facilities*, SC23-4864
- ▶ *HACMP for AIX: Administration and Troubleshooting Guide*, SC23-4862

2.4.4 HACMP/XD for HAGEO

By its IP-based replication functions, *HACMP/XD for HAGEO* can function as either a synchronous or asynchronous copy providing unlimited distance data mirroring. Data entered at one site is sent across a private TCP/IP network and mirrored at a second, geographically distant location.

HACMP/XD: HAGEO is a logical extension to the standard HACMP. While standard HACMP ensures that the computing environment within a site remains highly available, HACMP/XD: HAGEO ensures that one or more applications remain highly available even if an entire site fails or is destroyed by a disaster.

A site is a data center that is running HACMP/XD: HAGEO. Cluster nodes belong to one of two sites. All cluster nodes, regardless of the physical location, form a single HACMP/XD cluster. The site becomes a component, like a node or a network, that is known to standard HACMP. A site can have from one to seven nodes at the primary geographic location and as few as one at the secondary location for a total of eight nodes in an HACMP/XD: HAGEO cluster.

Data entered on nodes at one site is written locally and mirrored to nodes at the other site. If a site fails, the HACMP/XD: HAGEO process automatically notifies the system administrator and makes the geo-mirrored data available at the remote site.

When the failed site has recovered, starting the HACMP/XD: HAGEO process on a node at the reintegrating site automatically reintegrates the site into the cluster. The geo-mirrored data is synchronized between the sites during the reintegration.

The HACMP/XD cluster thus increases the level of availability provided by HACMP by enabling it to recognize and handle a site failure, to continue processing even though one site has failed and to reintegrate the failed site back into the cluster. To do that, it contributes additional cluster components and additional process facilities.

2.4.5 HAGEO cluster components

HAGEO components include:

- ▶ Geographic networks
- ▶ Geographic mirror devices

Figure 2-42 illustrates the components. The Dial Back Fail Safe (DBFS) connection between sites is explained in “Geographic networks” on page 82. Routers are included in the illustration, but they are not monitored by HACMP/XD.

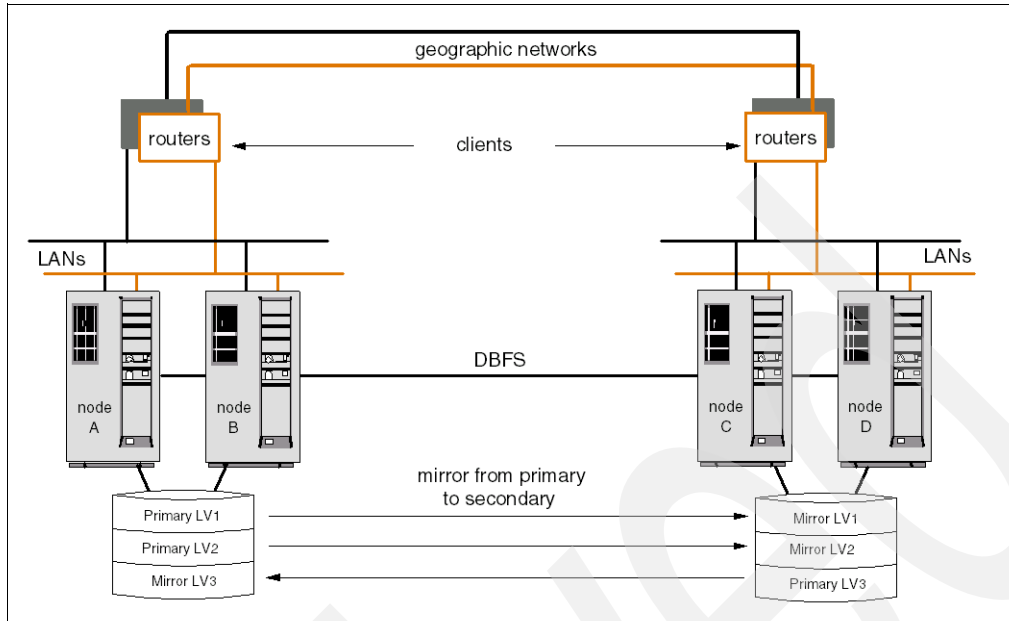


Figure 2-42 Sample HAGEO cluster

Nodes in an HAGEO cluster are located in two different geographic locations, called Site 1 and Site 2 in this example. The nodes in each location are defined as belonging to one of two sites, as well as to the cluster as a whole. The two sites together form a single HACMP/XD cluster. The site is an HAGEO cluster component. The HACMP Cluster Manager recognizes and handles site-related events.

Geographic networks

Geographic networks are point-to-point TCP/IP networks dedicated for mirroring data between the two sites within the HAGEO cluster. They are also used for HACMP heartbeat traffic. Clients do not have access to the geographic networks.

You establish at least two networks that follow diverse routes, besides setting up a method of backup communications for heartbeat traffic only. The backup system is used to determine whether a communications failure between sites on all geographic networks is due to network problems or to a site failure. With only one heartbeat network it is not possible to tell if the remote site or the network has failed. As shown in the example above, the software provides a DBFS routine that uses a telephone line as a possible method of backup communications.

These geographic networks are dedicated for HACMP/XD and HACMP traffic only, in order to have predictable response time and data transfer rate.

Geographic mirrors

HACMP/XD: HAGEO adds a geographic mirror component to the cluster nodes that is used to mirror data between sites. This is the same software used in GeoRM (2.4.6, "IBM Geographic Remote Mirror for AIX (GeoRM)" on page 86). Geographic data mirroring ensures that even if a site fails, the data is available and up-to-date at the other site. The cluster and the clients can keep on functioning.

Data mirroring with HACMP/XD: HAGEO can happen in these modes:

- *Synchronous mode.* During normal operation in synchronous mode, data is written first on the remote peer, then on the local node where the transaction was initiated. Synchronous mode offers the mirror image in the least amount of time between the writes on each site.

It ensures that the same data exists on the disks at both sites at the completion of every write. The synchronous process ensures that the local and remote disks have exactly the same data at all times. However, performance at the local site can be slowed somewhat by the mirroring process.

- ▶ *Mirror Write Consistency mode.* Mirror Write Consistency (MWC) mode is a variation of the synchronous mirror. Data is written to the local logical volume concurrently with the copy to the remote site. This improves the performance of the mirror. As with the synchronous mode, the write request does not return to the application until both the local and the remote writes are completed. The GeoMirror device keeps a detailed record in its state map of all requests that have not completed both local and remote writes. This guarantees that both sites have identical copies of the data, even in the event of a site failure during a transaction.
- ▶ *Asynchronous mode.* In asynchronous mode, the GeoMirror device does not wait for the write to the remote disk to complete before it returns to the application. The writes to the remote disk lag behind the writes to the local disk, up to a point set by the user. This process optimizes local response time. Performance at the site where the data is entered is minimally affected. The HACMP/XD: HAGEO process tracks the acknowledgments received from the remote site so that remote writes are not allowed to lag behind by more than a set number of bytes. GeoRM monitors a high water mark attribute to control how many 1KB blocks of data the secondary asynchronous device might lag behind the primary. The default is 128 KB. If that number is reached, messages are sent synchronously until enough acknowledgments are received to allow asynchronous mode again.

The entire geo-mirroring process is transparent to the application. It proceeds as though it simply wrote to a local disk.

The HACMP facilities monitor the state of nodes, networks and adapters, including the geographic networks and adapters and the geographic mirrors in the HACMP/XD: HAGEO cluster. The HACMP facilities combined with the HACMP/XD: HAGEO facilities detect and handle the following types of failures:

- ▶ *Local or intrasite failure* is the failure of a specific system component within a site.
- ▶ *Site isolation* is the failure of all geographic networks in a cluster. Both sides are still up, but they cannot communicate through the geographic network. The HACMP Cluster Manager uses a secondary HACMP/XD: HAGEO network, a network defined to HACMP for heartbeat traffic only, to verify whether site isolation exists, just as it uses an RS232 serial connection in a regular HACMP cluster to check on node isolation. If heartbeat communication over the secondary network is still possible, the problem is site isolation.
- ▶ *Site failure* is the failure of all nodes at a site. No heartbeat communication between sites exists.

A local failure is the failure of a specific system component within a site. The HACMP facilities handle local failures, assuming more than one node exists at a site, by following normal procedures. The HACMP Cluster Manager continues to compensate when a local area network, adapter, or node fails by switching service to another component of the same type, at the same site. In an HAGEO cluster with two nodes per site, for example, the HACMP Cluster Manager can handle the failure of one node within the site by having the second node at the same site take over those resources. When the failed node rejoins, its geo-mirroring information is updated and disks are resynchronized if necessary.

If the Cluster Manager determines that site isolation exists, it checks to see if the site has been configured as the dominant site. This site continues functioning; the Cluster Manager brings down the other site gracefully. Bringing one site down is necessary to avoid the data divergence that could occur if data continues to be entered at both sites, while geo-mirroring

is not possible. When only one site is functioning as the result of either site failure or bringing down a site, two things happen:

1. HACMP transfers ownership of the resources at the failed site to the viable site. In a concurrent access configuration, no transfer of ownership is necessary. In a cascading or rotating configuration, any resources defined for takeover are taken over by the nodes on the viable site.
2. The nodes at the takeover site mark the last geo-mirroring transaction completed. Then they keep track of all data entered after communication with the other site stopped.

The cluster continues processing even though one of the sites is down.

When a site fails and reintegrates, or after a site isolation is corrected, updates that occurred while only one site was functioning are resynchronized between the sites. HACMP/XD: HAGEO coordinates this process as part of the HACMP reintegration process.

HACMP/XD: HAGEO components

The software has three significant functions:

- ▶ *GeoMirror* consists of a logical device and a pseudo device driver that mirrors, at a second site, the data entered at one site.
- ▶ *GeoMessage* provides reliable delivery of data and messages between GeoMirror devices at the two sites.
- ▶ *Geographic topology* provides the logic for integrating the geo-mirroring facilities with HACMP facilities to provide automatic failure detection and recovery from events that affect entire sites.

A *GeoMirror device* is a logical device that has a local and a remote component. The local site is the site where the data is entered. The remote site is the geographically distant location that receives disk block transmissions from the local site and mirrors them. A GeoMirror device can receive data on either side. That means either side can be the local device where I/O occurs or the remote peer where writes are mirrored. This capability is necessary in case of node failure, or even site failure.

The GeoMirror device is layered above the logical volume manager (LVM) or above a physical disk. The applications to be geo-mirrored are directed to write to the GeoMirror device. GeoMirror devices support file systems, logical volumes, or raw disks. The GeoMirror device behaves like the disk devices it supports, so it is transparent to the application.

An HACMP/XD: HAGEO cluster can have a total of 1024 GeoMirror devices, distributed among two to eight nodes. Thus each node in a HACMP/XD: HAGEO cluster can have many GeoMirror devices. In synchronous mode each device can have up to seven remote peers. An HACMP/XD: HAGEO cluster can have a total of 256 GeoMirrored file systems.

The local and remote components of the GeoMirror devices on each node communicate through the GeoMessage component. GeoMessage sends data written to the local device across the geographic networks to the remote peer. The remote device sends back acknowledgments that the data is written at the remote site.

Each GeoMirror device on a node maintains a state map. The state map is a record of the current state of all data regions written on the GeoMirror device by the node. When a site fails and recovers, the HACMP/XD: HAGEO software reads the state maps on each node in order to reconstruct and update the mirrors on the recovered node, thereby synchronizing the GeoMirror devices. That process is automatic. If the synchronization process is interrupted for any reason, it restarts from the point at which it was interrupted. The state map device is a raw logical volume.

The geographic topology function has the tools that allow the HACMP Cluster Manager to detect and react to HACMP/XD: HAGEO cluster site-related events. The geographic topology function includes the following tools:

- ▶ Scripts and programs that integrate handling GeoMirror and GeoMessage in cluster events such as node and network joins and failures
- ▶ Scripts that integrate the starting and stopping of the GeoMirror and GeoMessage functions into the HACMP start and stop scripts
- ▶ Error-log messages to ensure GeoMirror and GeoMessage activities are logged

Figure 2-43 describes the links between the different components.

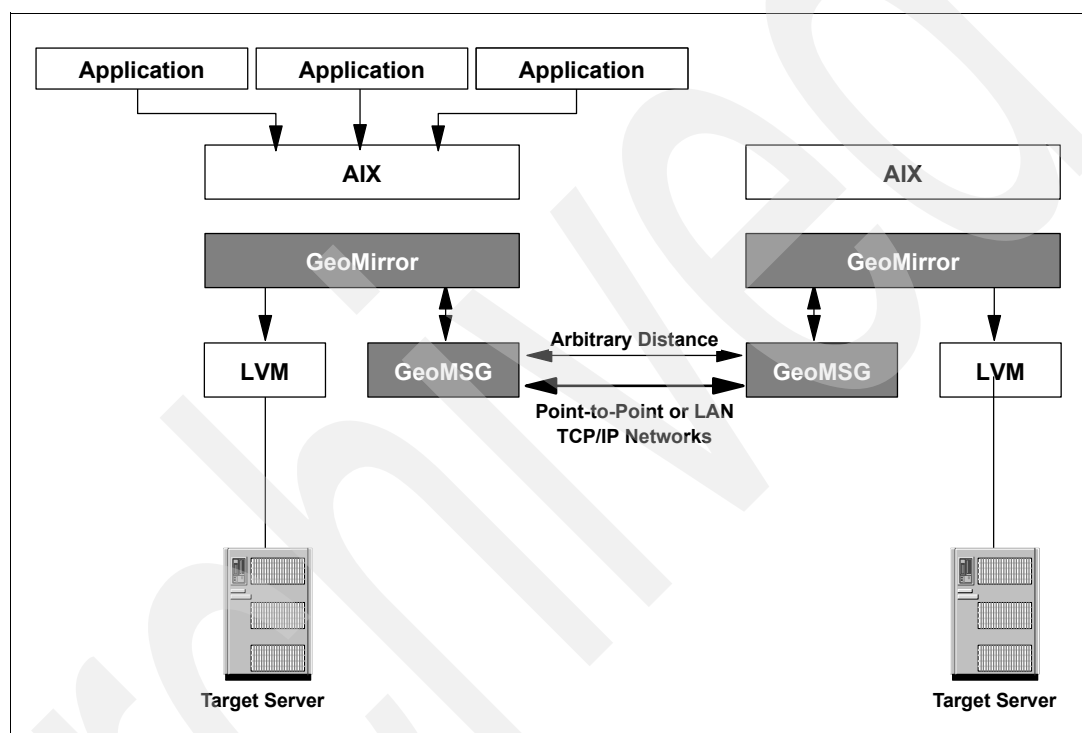


Figure 2-43 Components

Criteria for the selection of HACMP/XD:HAGEO

The following criteria can influence your choice of this solution:

- ▶ The distance between sites is limited only by an acceptable latency, and it could normally be more than 100 km. Usually, this is a solution for providing long-distance mirrors.
- ▶ It requires HACMP/XD to provide GeoMirror functionality.
- ▶ In this configuration, the active server writes to its disk, then waits for acknowledgement from the remote server where the mirrored write was applied. (However, note that this sequence depends slightly on which of the three mirroring modes you are using – synchronous, asynchronous, or synchronous with mirror write consistency.) Therefore, the disk systems on the SAN do not have to be visible from servers at both sites. In fact, the inability to interconnect SANs would normally be a good reason for choosing this solution.
- ▶ As a disaster recovery solution, it applies only to AIX servers with HACMP/XD and GeoMirror.
- ▶ It uses *Host based* or *Software mirroring*, which imposes a small extra workload on the server.

- ▶ Where distance limitations do not allow disk-based heartbeats, we recommend independent WAN links, a non-IP heartbeat if possible, and modem dialup to check heartbeat if other methods are lost.
- ▶ It is the most latency-dependent mirroring mechanism because it is IP-based.

Additional information

For further details on HACMP/XD: HAGEO, see:

- ▶ *HACMP/XD for AIX 5L Concepts and Facilities for HAGEO Technology SA22-7955*
- ▶ *HACMP/XD for AIX 5L Planning and Administration Guide for HAGEO Technology SA22-7956*
- ▶ *Disaster Recovery Using HAGEO and GeoRM, SG24-2018*

2.4.6 IBM Geographic Remote Mirror for AIX (GeoRM)

IBM Geographic Remote Mirror for AIX (GeoRM) helps ensure business reliability and provides for Disaster Recovery by supplying mirroring of data over unlimited distances.

GeoRM ensures that critical data is continuously being copied and stored at a geographical location that is not prone to the same planned or unplanned outage as the site where the application normally runs.

GeoRM provides wide-area mirroring of client data between systems connected by IP-based networks. Data stored on a disk at one site is sent across a TCP/IP network and mirrored on a disk at a second, geographically distant location. Data is mirrored at the logical volume (LV) level in a primary/secondary relationship. The application has access to the primary LV only, while the secondary LV, located at the remote site, is used for mirroring only.

This capability of the GeoRM software is called *GeoMirroring Devices* (GMDs). The distance between sites is unlimited. The integrity of the data is guaranteed. Data mirroring with GeoRM can be *synchronous*, ensuring real-time data consistency, or *asynchronous*, for maximum performance while maintaining sequential data consistency. When running in asynchronous mode, GeoRM monitors a high water mark attribute. This attribute controls how many 1-KB blocks of data the secondary asynchronous device might lag behind the primary.

A third mode, called *Mirror Write Consistency* (MWC), is a variation of the synchronous mirror. Data is written to the local logical volume concurrently with the copy to the remote site. This increases the performance of the mirror. As with the synchronous mode, the write request does not return to the application until both the local and the remote writes are completed. The GMD state map of all requests that have not yet completed both local and remote writes guarantees that both sites have identical copies of the data, even in the event of a site failure during a transaction.

GeoRM provides a solution for these system operations that can affect Business Continuity and reliability:

- ▶ **Data availability and security.** GeoRM provides a way to back up data to disk at a geographically distant site.
- ▶ **Scheduled maintenance and distributed operations.** GeoRM provides a way to ensure data synchronization and coordination across geographically distant sites.
- ▶ **Business Continuity and Disaster Recovery.** GeoRM provides a way to have a hot site with current data, which can be part of a complete Disaster Recovery Plan.

GeoRM provides a System p computing environment which ensures that business operations are minimally affected by planned maintenance operations or by a disaster such as a flood, fire, or bombing that wipes out an entire site.

GeoRM configurations

There are the two basic GeoRM configurations:

- *GeoRM active-backup.* Applications run on GeoMirror devices at the active site and are mirrored to corresponding GeoMirror devices at the backup site.
- *GeoRM mutual backup.* Applications run on GeoMirror devices at each site and are mirrored on GeoMirror devices at the other site.

Figure 2-44 shows a three-node GeoRM configuration. The configuration has two sites with two nodes in Site 1 and one node in Site 2. The nodes in Site 1 can be geographically separated, even though GeoRM considers them as one logical site. The three nodes have access to the geographic networks through the routers. Data entered in Site 1 is sent over a geographic network and mirrored on a corresponding disk at Site 2. Each node in Site 1 mirrors the disks it owns to the node C located in Site 2.

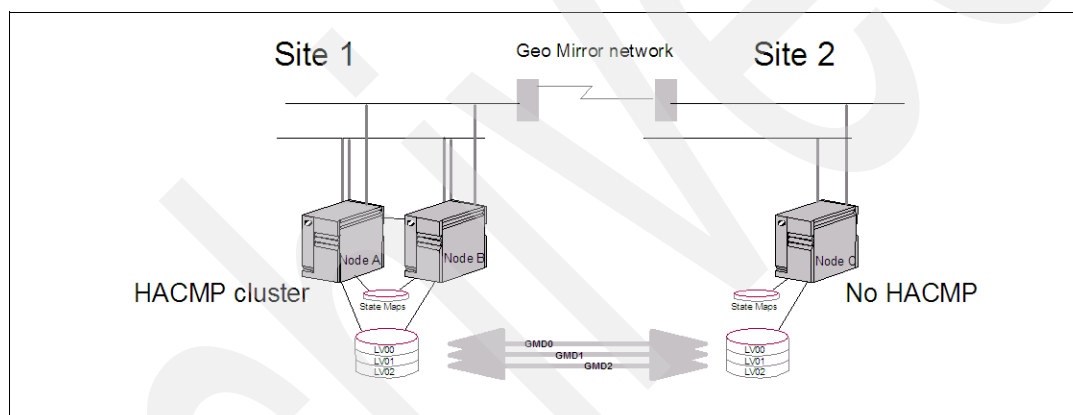


Figure 2-44 Sample GeoRM configuration

GeoRM combines the levels of data readiness, site inter-connectivity and recovery site readiness. GeoRM supplies a reliable replication system (geo-mirroring) that ensures business operations are minimally affected by a disaster or scheduled downtime. The GeoRM configurations require some manual intervention to handle recovery procedures. The geo-mirrored devices must have their roles switched from secondary to primary to allow access of the data. You can automate error notification and have planned procedures established for various contingencies, to minimize down time in case of any emergency. So GeoRM can be classified as a Tier 6 disaster recovery solution.

GeoRM with HACMP

Many clients use HACMP to protect their mission-critical systems against hardware and software failures. The cluster nodes in an HACMP cluster are typically located in the same room or building and thus do not provide the ability to preserve access to data should a disaster strike. GeoRM can be used to mirror data to a remote site. Combining the two products provides the capability to automatically recover from local hardware failures and the capability to manually recover from total site disasters.

Do not confuse using HACMP and GeoRM together in this fashion with the capabilities offered by HACMP/XD for AIX. HACMP/XD provides automatic failure detection, failover processing and recovery from site failures and disasters. The HACMP and GeoRM interaction provides the ability to geographically mirror data from an HACMP cluster to a remote site

using GeoRM. It does not provide automatic failover between sites. If you require automatic failover between sites, consider implementing the HACMP/XD product.

Additional information

For further details on GeoRM and its integration with HACMP, see:

- ▶ *Geographic Remote Mirror for AIX: Concepts and Facilities*, SC23-4307
- ▶ *Geographic Remote Mirror for AIX: Planning and Administration Guide*, SC23-4308
- ▶ *Disaster Recovery Using HAGEO and GeoRM*, SG24-2018

2.4.7 Cross-site LVM mirroring

AIX Logical Volume Manager (LVM) provides software-based disk mirroring, which has similar characteristics to hardware-based disk mirroring. However, without HACMP, deploying AIX split mirrors in a clustered, highly available environment is a manual task, and therefore, many administrators have preferred hardware mirroring.

Using HACMP AIX split mirrors can be automated in cluster failover scenarios. This disaster recovery solution is known as cross-site LVM mirroring. This allows storage devices and servers to be located at two different sites, and uses AIX LVM mirroring to replicate data at geographically separated sites interconnected via a SAN, for example.

The disks can be combined into a volume group via the AIX LVM, and this volume group can be imported to the nodes located at different sites. The logical volumes in this volume group can have up to three mirrors. Thus, you can set up at least one mirror at each site.

The disk information is replicated from a local site to a remote site. The speed of this data transfer depends on the physical characteristics of the channel, the distance and LVM mirroring performance. The LVM groups keep all data consistent at the moment of transaction since writes are applied in the same order on both disk systems. The information stored on this logical volume is kept highly available and, in case of certain failures, the remote mirror at another site still has the latest information, so the operations can be continued on the other site.

HACMP automatically synchronizes mirrors after a disk or node failure and subsequent reintegration. HACMP handles the automatic mirror synchronization even if one of the disks is in the PVREMOVED or PVMISSING state. Automatic synchronization is not possible for all cases, but you can use C-SPOC to manually synchronize the data from the surviving mirrors to stale mirrors after a disk or site failure and subsequent reintegration.

Failover

In the following sections we describe the circumstances under which failover occurs.

Single-Site Data Availability

When a failure occurs on the Disk System level, there is no failover procedure. The operating system merely sees itself as accessing one volume instead of two and continues processing under that condition rather than, for example, swapping the disk that it has access to, as is the case in the GDPS/PPRC HyperSwap feature in the System z environment.

Cross-Site LVM

In cases where the environment is configured so that disk systems are in physically separate locations, there are two failure conditions that have to be addressed. If the failure is at the disk system level, then operations proceed as they would in a single site data availability configuration. There is no failover, and AIX proceeds, reducing the number of volumes used per write by 1.

In cases where the failure is at the site level, there is a recovery required. In this case, the site with the remaining disk is no longer viewed as a full volume group by AIX. If, for example, there were two locations in the configuration, then it would be viewed as half of a volume group. For this reason, when performing a recovery off of such data, it is necessary to force **varyonvg**. This causes the volume group to vary on and be usable for recovery purposes. AIX has already seen to it that the data is consistent, so the recovery should proceed as normal.

Criteria for the selection of this solution

The following criteria can influence your choice of this solution:

- ▶ Distance between sites would normally be within metropolitan area distances. Using direct connections as opposed to a switched Fibre Channel network would shorten the distance.
- ▶ This solution requires only HACMP, which has a slight cost advantage over HACMP/XD.
- ▶ In AIX only environments LVM can be a very simple and effective method of maintaining data availability and short distance disaster recovery.
- ▶ In this configuration the active server performs all writes to disk on both sites, the passive server simply monitors the heartbeat of the active server. SAN disk must therefore be *visible* from servers at both sites, but the disk can be of any type as long as one server can be attached to both sets of disk at the same time. This might rule out a multivendor disk solution, but keep in mind that additional software, for example, different multipathing software or drivers, could avoid this.
- ▶ It uses *Host based* or *Software mirroring*, which imposes a small extra workload on the server.
- ▶ This solution can use a disk-based heartbeat.

Additional information

For further information about HACMP architecture and concepts, see:

- ▶ *HACMP for AIX: Concepts and Facilities*, SC23-4864
- ▶ *HACMP for AIX: Administration and Troubleshooting Guide*, SC23-4862

2.4.8 Global Logical Volume Mirroring (GLVM)

GLVM is intended to serve as a long term replacement for GeoRM and, when combined with HAGEO, provides similar function, but in a simplified method when compared to management of the Geo-Mirror Devices under GeoRM.

As shown in Figure 2-45, GLVM uses LVM and builds on it to extend to longer distances. It performs its mirroring over TCP/IP networks versus the FCP networks that are used for LVM connections.

The mirroring component of GLVM comes in the form of the RPV client device driver. This appears to be a driver for a disk system and sits underneath the level of AIX LVM. Rather than focusing on creation of GMDs (GeoMirror device), as is the case in GeoRM, we simply run LVM and then use the RPV (remote physical volume) driver to trick it into thinking that the second copy of disk is local when, in fact, it is not.

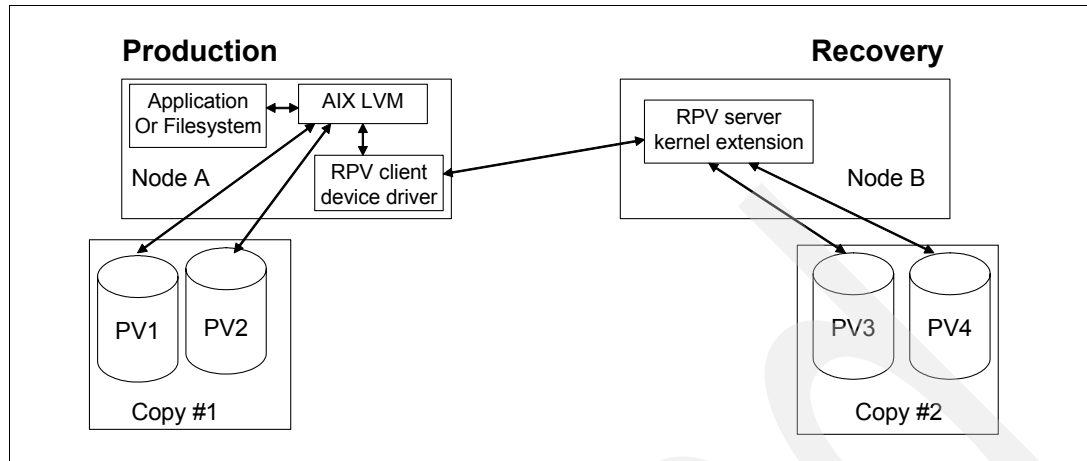


Figure 2-45 A sample GLVM configuration

In the diagram, Node A performs its writes through LVM. LVM then writes the data to copy #1 and simultaneously believes that it is writing to copy #2 locally by way of the RPV client device driver. In truth, the RPV client device driver communicates with an extension of its kernel in the recovery site known as the *RPV Server Kernel Extension*, which then passes the identical write to the second set of physical volumes.

When Node B becomes active, the operation is reversed. The RPV server kernel extension becomes active in Node A and writes are mirrored from A to B. This is possible because under GLVM it is a requirement to set up the reverse configuration as well, specifically to support failover and failback operations (Figure 2-46).

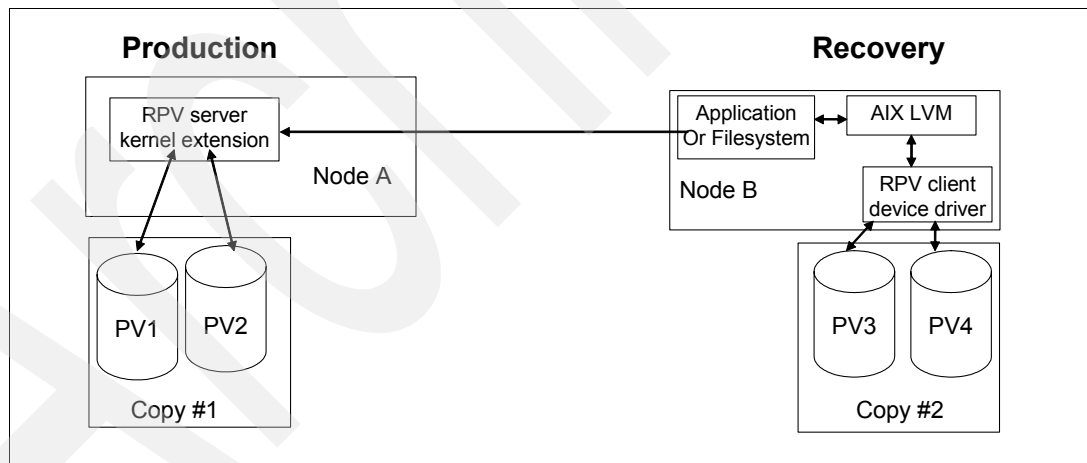


Figure 2-46 GLVM failover

Combining GLVM with HACMP/XD

As is the case with Metro Mirror (HACMP/XD) and GeoRM (HACMP/XD for HAGEO), GLVM can be enhanced with the automation provided in the HACMP/XD code. In this case, GLVM continues to maintain control of the mirroring environment, but HACMP/XD manages the configuration and reconfiguration of the server nodes.

Further information

For further information about how to implement HACMP/XD and GLVM, refer to *Implementing High Availability Cluster Multi-Processing (HACMP) Cookbook*, SG24-6769.

2.4.9 AIX micro-partitions and virtualization

In this section we discuss AIX micro-partitions and virtualization.

Introduction

AIX provides virtualization for processing resources. This allows System p environments to make more efficient use of processor resources when planning a business continuity solution.

The following components are involved:

- ▶ **Micro-partitions** allow a single processor engine to be split it into up to 10 individual partitions.
- ▶ **Shared processor pool** is a group of processors that can be used by any process as its resources are required.
- ▶ **Virtual IO (VIO)** provides a shared pool of Ethernet and FC adapters to be shared among processes.

In the following sections, we provide more details on each of these entities.

Micro-partitions

Power5 Micro-partitioning splits the processing capacity of a single processor into *processing units*, where 1% of a processor is a single processing unit. Under Power5 Micro-partitions, the minimum processor size is 10 processing units and the theoretical maximum is 10, although overhead in a loaded system reduces this.

2.5 Continuous Availability for MaxDB and SAP liveCache

This solution addresses the failover to a standby server and fast restart of SAP liveCache landscapes.

2.5.1 MaxDB and SAP liveCache hot standby overview

In this section we have a high level overview of the MaxDB and SAP liveCache hot standby solution.

This is a Continuous Availability technology, as shown in Figure 2-47. It uses AIX HACMP and FlashCopy to protect SAP SCM 4.1 and above environments. It gives very rapid application failover to a standby server.

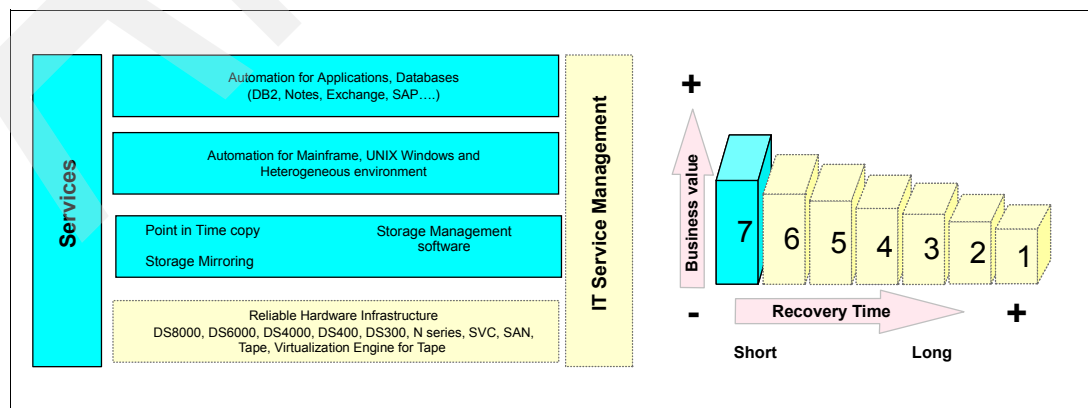


Figure 2-47 MaxDB and SAP liveCache hot standby solution positioning

Solution description

In an SAP SCM environment, SAP liveCache is becoming an increasingly critical component. SAP liveCache is an SAP database that is preloaded into memory for very fast access. The benefits and performance advantages of liveCache result from its ability to build a very large memory cache and perform specially tailored functions against in-memory data structures.

LiveCache can be provided with a failover solution using HACMP. In a traditional failover solution the database disks are taken over by the backup server and the database recovered and restarted. This approach has several disadvantages in a large liveCache implementation.

The memory cache for large implementations can exceed many gigabytes in size. A traditional failover solution must restart the database and rebuild this memory cache. This might represent a large time delay, before liveCache is ready to resume production. Before this action can even begin, a failover solution must acquire and activate the disks of the failed system and perform database recovery. All of these activities increase with the size of the liveCache both on disk and in memory. The larger the liveCache, the more likely is its importance in the SAP landscape; and the longer its failover and recovery time.

SAP has introduced hot-standby functionality with liveCache 7.5, available with SCM 4.1, to provide the fastest possible means of recovery. While this functionality has been implemented for both the MaxDB and liveCache, we focus on the liveCache for SCM.

This solution design for the liveCache hot standby requires specific functionality on behalf of the supporting I/O subsystem (split mirror and concurrent volume access) and is closely integrated with the control software of the subsystem via an API. The integration of this solution requires an I/O subsystem specific shared library, mapping the SAP requests to the subsystem, and a cluster solution on the part of the System p AIX server platform to manage the failover control and IP access. IBM offers the I/O subsystem solution for AIX for the SVC, DS8000, and ESS. HACMP provides the cluster functionality.

Solution highlights

The benefits of this implementation are the following:

- ▶ Speed of recovery and return to production
- ▶ Coverage of server outage
- ▶ Coverage of database outage
- ▶ Coverage of data disk failures
- ▶ Automated failover and fallback
- ▶ Designed to result in minimal or no performance impact on production system
- ▶ Ease of management for DB administrators

Solution components

This solution, which addresses SAP SCM environments that run on AIX, has the following components:

- ▶ System p servers
- ▶ AIX 5.1 or higher
- ▶ Databases:
 - MaxDB 7.5
 - SAP liveCache 7.5 (available with SCM 4.1)
- ▶ Storage systems:
 - IBM System Storage ESS with Advanced Copy Services
 - IBM System Storage DS8000 with Advanced Copy Services
 - IBM System Storage SVC with Advanced Copy Services

- ▶ TCP/IP connections from each server to the storage system
- ▶ High Available Cluster Multiprocessing (HACMP)
- ▶ MaxDB and SAP liveCache hot standby storage dependent library:
 - For SAN Volume Controller, the IBM2145CLI and SecureShell (SSH) are required on all servers.
 - For DS8000, DSCLI is required on all servers and FlashCopy.
 - For ESS, the IBM2105CLI is required on all servers and FlashCopy.

2.5.2 MaxDB and SAP liveCache hot standby in greater detail

In this section we give a more detailed description of the setup and mode of operation of MaxDB and SAP liveCache hot standby solution.

The high level operating landscape is shown in Figure 2-48. We have two AIX nodes, node A and B, sharing an external SAN-connected disk system. The nodes are connected to the storage system over TCP/IP network and HACMP provides application monitoring and failover between the nodes.

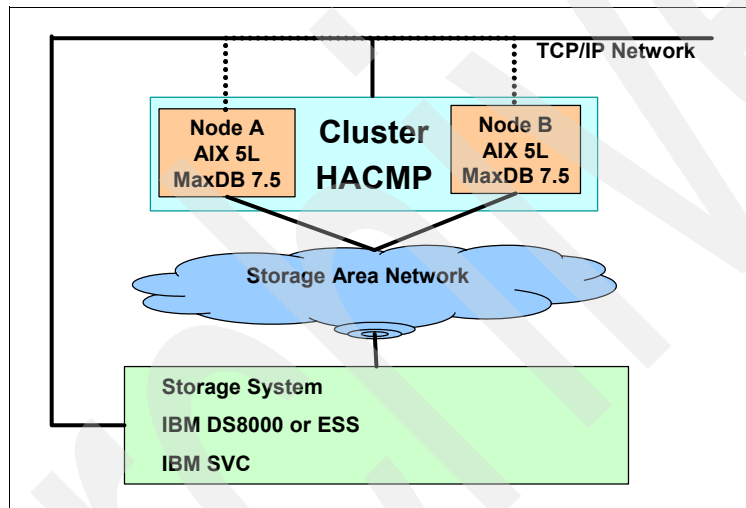


Figure 2-48 Basic setup

In Figure 2-49 we see the initial setup tasks on node A, the production server, and node B, the standby server. You perform the following steps:

1. Connect the first two LUNs (physical storage devices) to node A, run **cfgmgr** to discover the LUNs and then create an AIX volume group and logical volumes.
2. Perform a FlashCopy between the first and third volumes and when the FlashCopy completes go to step 3.
3. Import the volumes on node B.

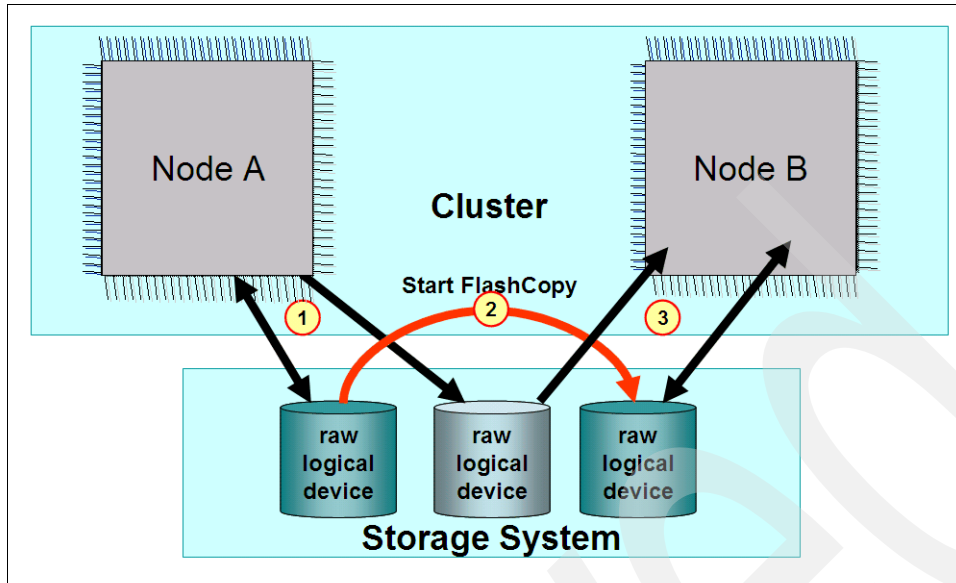


Figure 2-49 Initial host setup

The third step, shown in Figure 2-50, is to set up the liveCache environment. You perform the following steps:

1. Install the first liveCache instance on node A. This populates the first set of data volumes and the shared log volume.
2. Install and configure the libhss interface between SAP liveCache and copy services on the storage system.
3. Configure node A as the master server.

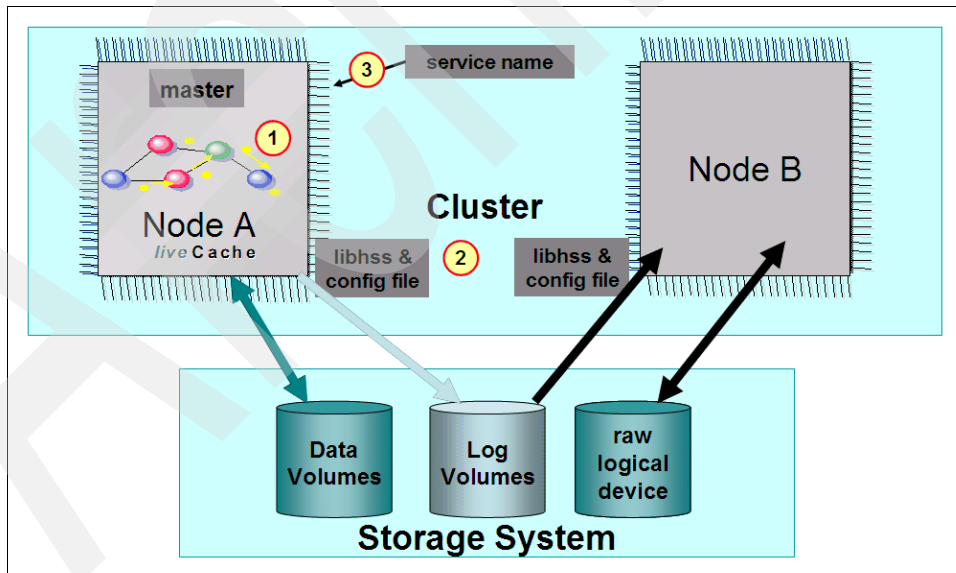


Figure 2-50 Setup liveCache

In the fourth step, shown in Figure 2-51, we add the standby to the hot standby system.

1. We initialize the standby database on node B, using the database management tool from SAP. This copies the database parameter files to the standby, and starts and initializes the standby.

2. Initialization means that the database is requesting a copy of the data volume. This is done using the **libhss** interface, which issues a FlashCopy task to copy an I/O consistent image of the master data volumes on node A to the standby on node B.
3. As soon as the FlashCopy command returns, the database on the standby node turns to the database mode **STANDBY**, continuously reading the shared log volumes and applying the changes to the internal memory and to the data volumes.

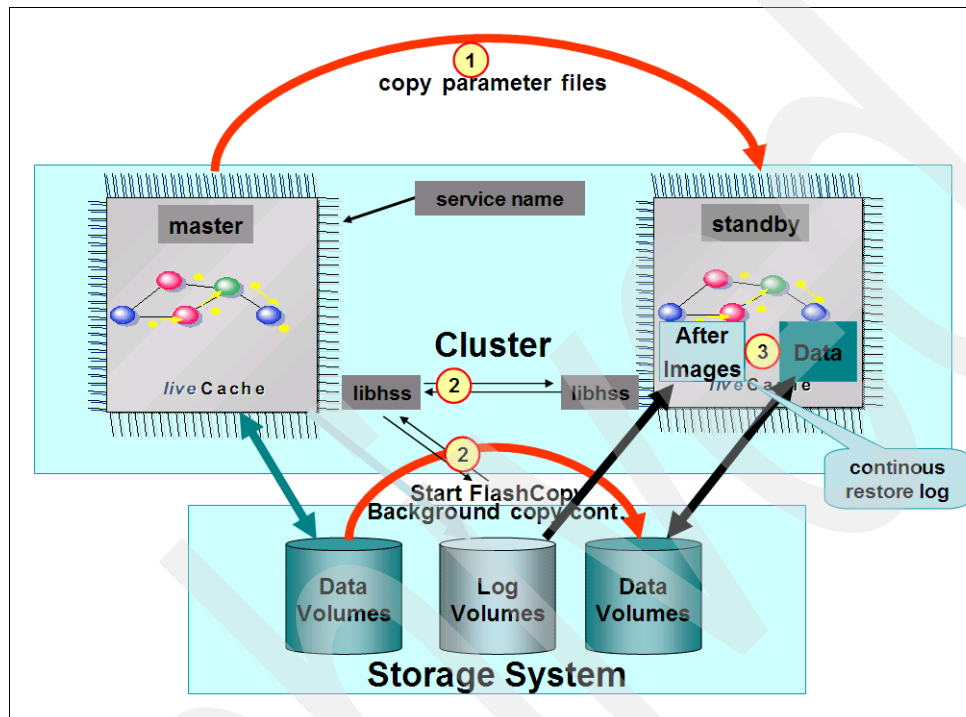


Figure 2-51 Initialize hot standby

The next step, shown in Figure 2-52, is to enable high availability between node A and node B. You must perform the following steps:

1. Install HACMP on both nodes and configure the service for high availability.
2. Start the application so it can access the service managed by HACMP.

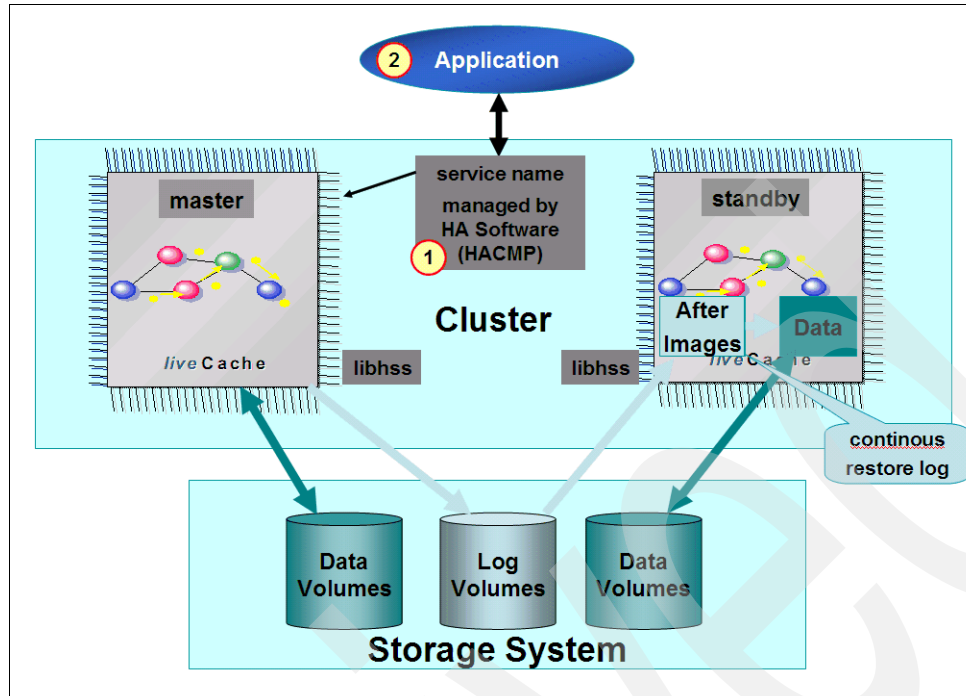


Figure 2-52 Enable high availability

In Figure 2-53 we illustrate how the failover scenario works. HACMP monitors the status of servers and applications. These are the steps in a failover:

1. The master on node A fails either because the server fails or because the server loses access to storage.
2. HACMP initiates a failover of the service to node B, the remaining node.
3. The standby database accesses the log volumes in r/w mode, rereads the log and applies all remaining committed transactions. After this, the database switches to mode MASTER and handles all application requests.

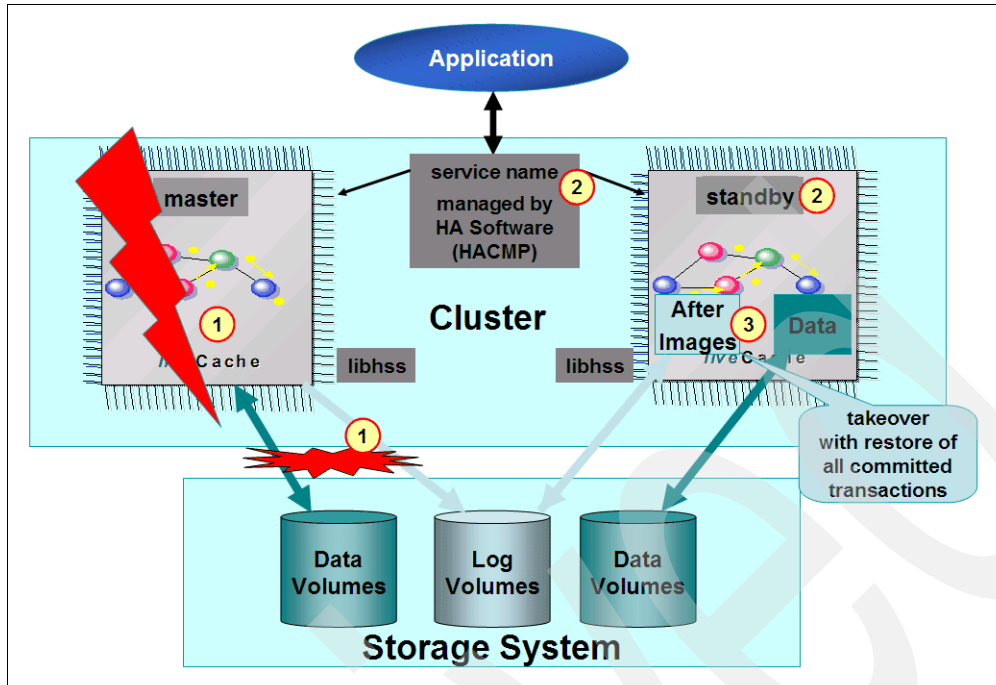


Figure 2-53 Failover scenario

After the failover operation has completed, the service remains on node B. Commands are available to transfer the service back to node A.

Another conceivable failover configuration for MaxDB and SAP liveCache hot standby is using Metro Mirror to replicate data between a local and a remote storage system. This extends support to servers in separate physical locations (Business Continuance configurations), to further help safeguard these critical business processes. See Figure 2-54.

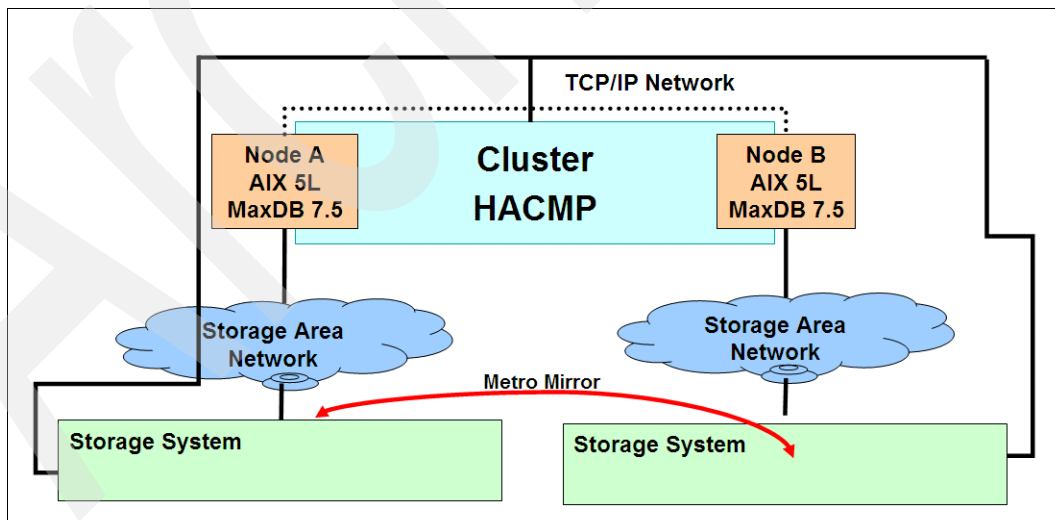


Figure 2-54 PPRC scenario

Additional information

For more information, see the following Web site or contact your IBM sales representative:

<http://www.ibm.com/servers/storage/solutions/sap/index.html>

2.6 Copy Services for System i

In this section we cover the IBM Copy Services for System i.

2.6.1 Answering the question: Why use external disk?

The System i space (since the AS/400® days) is traditionally known for its use of integrated disk. Going forward, external storage, such as the IBM System Storage DS8000 is available and compatible with the System i. These facts beg the question, “If the internal disk is there and works well, then why use external storage?” Although this chapter focuses on continuous availability, this is an important question to clarify, particularly for System i environments.

The answer, in brief, is that by making use of external storage, a System i can now work within a SAN. This opens up alternatives and efficiencies that are not available with System i integrated disk alone, including:

- ▶ A single storage infrastructure for a heterogeneous server environment
- ▶ Tiered storage based on the requirements of applications
- ▶ Point in Time copies and disk mirroring
- ▶ Automated failover to mirror copies

Heterogeneous server environments

These days, mixed server environments are common, including mainframe, Linux, open, and the types of servers. Although an environment using System i servers only might be able to survive quite well using all internal storage, it might represent an inefficiency of Total Cost of Ownership, when other server types are present, since the internal disk cannot be shared across all server types.

By using a SAN with all servers connected, a single storage farm can be used by all servers in the environment. This could represent a considerable savings in terms of the cost of buying disk for each server type and the cost required to manage each type of disk.

A sample layout is shown in Figure 2-55.

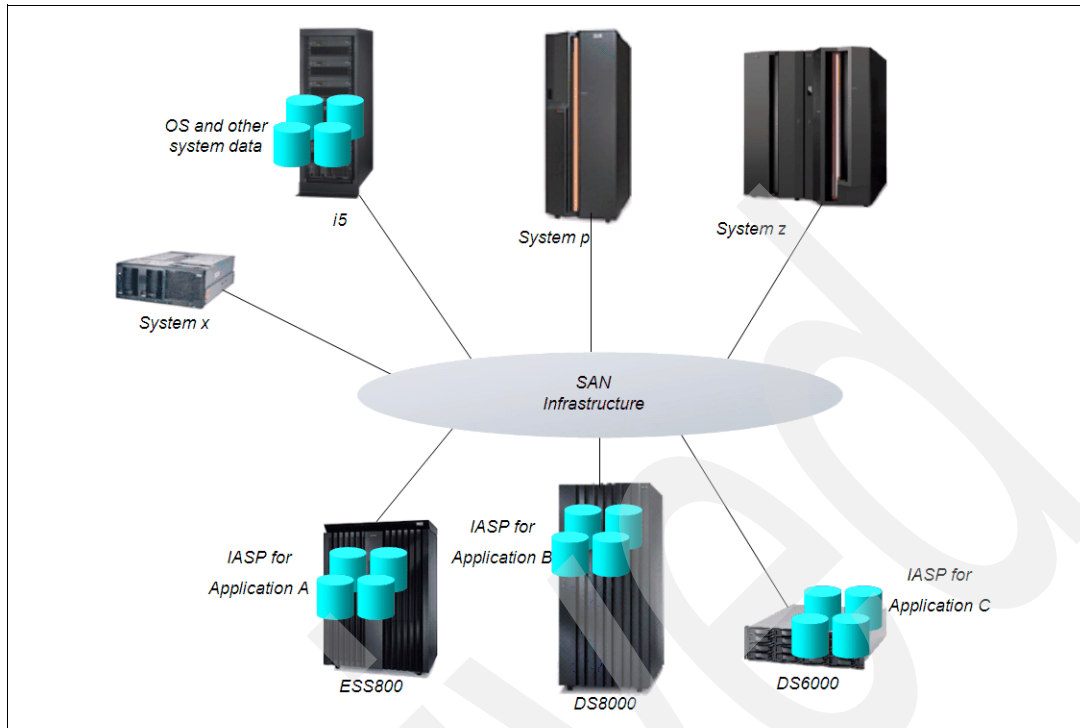


Figure 2-55 Multi-server storage SAN

Tiered storage

Different applications have different performance requirements. Using a SAN gives an enterprise the freedom to choose where they choose to keep the data of any given application, whether they must use a DS8000, DS6000, or internal disk.

A SAN environment, together with the Copy Services for System i offering, makes this possible through the use of Independent Auxiliary Storage Pools (IASPs).

Point in time copy and disk mirroring

Where System i internal storage is concerned, there has never been a direct “block level” replication function. Replication between System i servers for business continuity purposes was always handled at the server level by employing High Availability Business Partner (HABP) software that would assist by mirroring application logs to a recovery site standby server that could apply those logs. While this is sufficient for some environments, the Copy Services for System i offering enables enterprises to take advantage of copy services such as Metro Mirror, Global Mirror, and FlashCopy.

A FlashCopy is a point in time copy of the data in an IASP. Creating this copy makes an additional copy of data available, separate from the production data. This copy can be used for backups without impacting the associated application. Likewise, the copy could be used for test data or for data mining.

Disk Mirroring uses the disk system itself as the mechanism for making an identical copy of data. This can be via Metro Mirror or Global Mirror (7.7, “Copy Services functions” on page 270). These copy services maintain zero data loss in the case of Metro Mirror, or mirror data at unlimited distances with efficient use of bandwidth in the case of Global Mirror. In both cases the mirror is application independent.

Fast and efficient failover

While failover is possible through HABP software, System i is capable of a faster failover with better data currency when using IASPs with System i Clusters in combination with IASPs.

2.6.2 Independent Auxiliary Storage Pools (IASPs)

Typically, System i views all attached storage, both internal and external, as a single level, single device — a single LUN, you could say. This is actually quite an amazing level of storage virtualization given the age of the i5/OS® and its predecessor, OS/400®.

That said, it is also one of the main barriers to effective use of tiered storage and copy services, because there is no way to break out data of different priorities or to determine which data should be mirrored and which should not. IASPs are a function designed to do just that and, in their own way, introduce their own sort of LUN. See Figure 2-56.

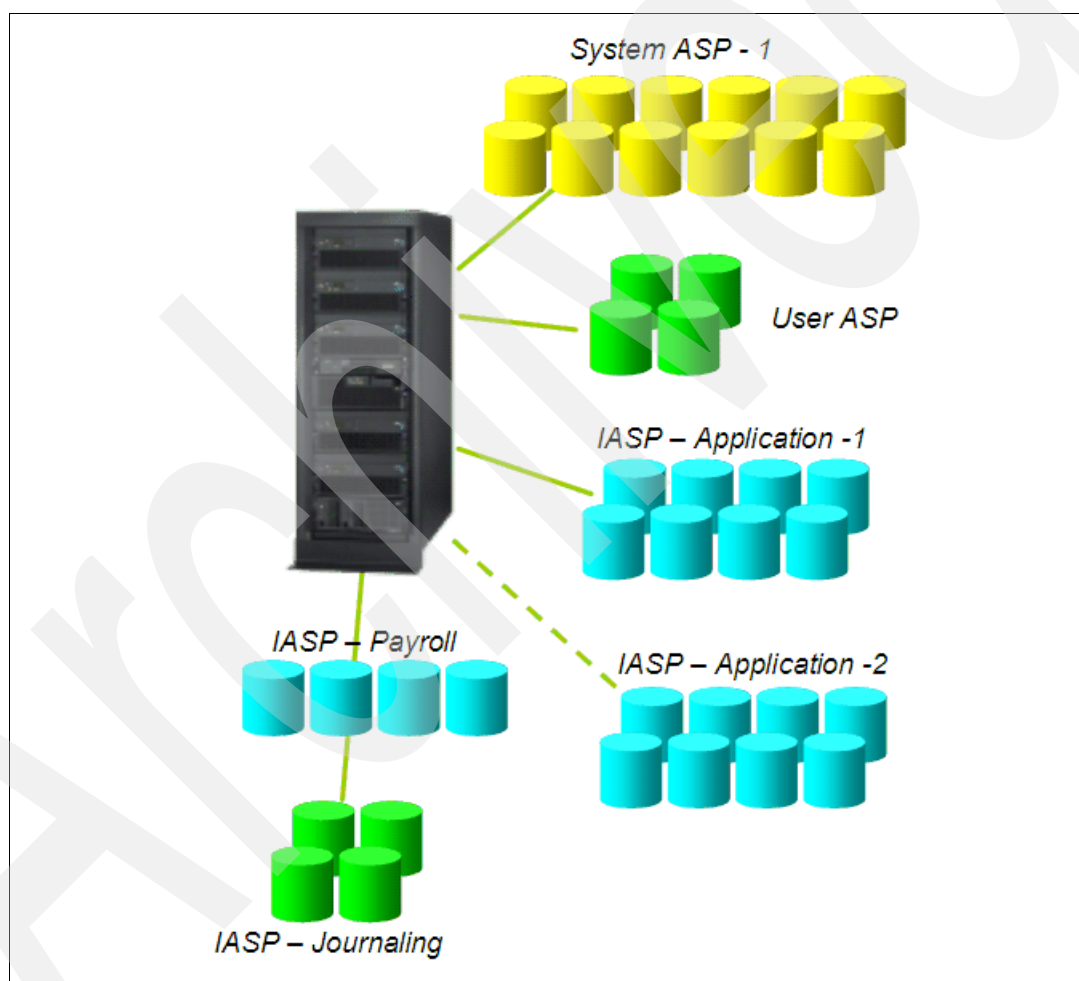


Figure 2-56 Independent Auxiliary Storage Pools (IASPs)

The IASPs allow disk associated with a given application to be put into its own disk pool. This has immediate benefits in terms of isolating maintenance windows. Without IASPs, the entire storage pool would have to go offline in order to perform maintenance on any application using it. With IASPs just the one particular storage pool has to go offline in order to do whatever maintenance is necessary.

This also serves as a basis for Copy Services and Clustering.

While IASPs are not, in and of themselves, a function of the Copy Services for System i toolkit, the involvement of System i and i5/OS resources in the development of the toolkit insures that IASP usage can be accomplished cleanly and effectively within its resources.

Requirements

IASPs require i5/OS V5R2 or higher and IASP spool files support V5R3 and higher. They are only possible as part of the System i Copy Services offering from STG Lab Services.

2.6.3 Copy Services with IASPs

Metro Mirror, Global Mirror, and FlashCopy are available.

Metro Mirror

While it is technically possible to do Metro Mirror as it would be done in other environments, the fact that System i typically sees only a single level of storage creates a challenge. This is because mirroring in a single level environment means that all data is mirrored. In a synchronous environment like Metro Mirror, this can create performance issues because, in addition to application specific data, the copy service also mirrors the *SYSBAS information. The *SYSBAS information is not required in the recovery site and doing so creates more transactions that must be synchronously mirrored outside of the relevant applications. As each transaction must be mirrored in turn, this could negatively impact the performance of all applications in that specific System i environment.

Metro Mirror is explained in more depth in 7.7, “Copy Services functions” on page 270.

Global Mirror

As an asynchronous mirroring technology, Global Mirror does not impact the function of any production applications. However, bandwidth is one of the largest costs involved in any ongoing business continuity effort. As such, it is important to consider whether it is truly necessary to mirror all data.

Through modeling, STG Lab Services has demonstrated a significant cost savings by using IASPs and mirroring only the application data. Again, the *SYSBAS data does not have to be mirrored and doing so can actually complicate recovery efforts. IASPs give the option of separating out that data and only mirroring the application specific data that is required.

Global Mirror is explained in more depth in 7.7, “Copy Services functions” on page 270

FlashCopy

When issuing a FlashCopy in a non-IASP environment, all applications using the source storage must be quiesced in order to produce a consistent, restartable, copy of data in the target storage.

By combining FlashCopy with IASPs, it is possible to eliminate the necessity to quiesce the applications. Specific applications can have their Point in Time copy made at will, rather than all copies being made at once. This copy can then be used to create tape backups without a long application outage, or for data mining, as required.

FlashCopy is explained in more depth in 7.7.1, “Point-In-Time Copy (FlashCopy)” on page 270.

2.6.4 Copy Services for System i

Copy Services for System i is required in order to use Metro Mirror or Global Mirror within a System i environment. This is not a single product or pre-packaged software but, rather, a services offering tailored to the individual environment where it is implemented, and is available only through STG Lab Services.

Its benefits include:

- ▶ The ability to use IASPs
- ▶ Automatic failover
- ▶ A utility to ensure that all parts of the replicated environment function

Next we explain each of these in more detail. Figure 2-57 shows a diagram of server and storage clustering configurations.

Use of IASPs

As noted above, IASPs provide greater flexibility for data management. Without IASPs, all data is viewed in a single level of storage. This means that any time there is a storage outage, all data must take the same outage. Any time data is mirrored or copied, all data must be mirrored or copied.

Using ASPs allows the data to be broken down into more manageable blocks that can take individual outages or be mirrored independently. It also allows for participation within a SAN environment.

Automatic failover

Copy Services for System i is not just a tool to make more effective use of DS6000/DS8000 copy services within a System i environment. It also provides a true high availability offering.

Through the Copy Services offering, it is possible to set up mirror pairs along with a server cluster. Doing so allows the environment to failover some, or all, applications from one location to another. As always, these locations could be within a single physical site or between geographically dispersed sites, depending on the form of disk mirroring in use.

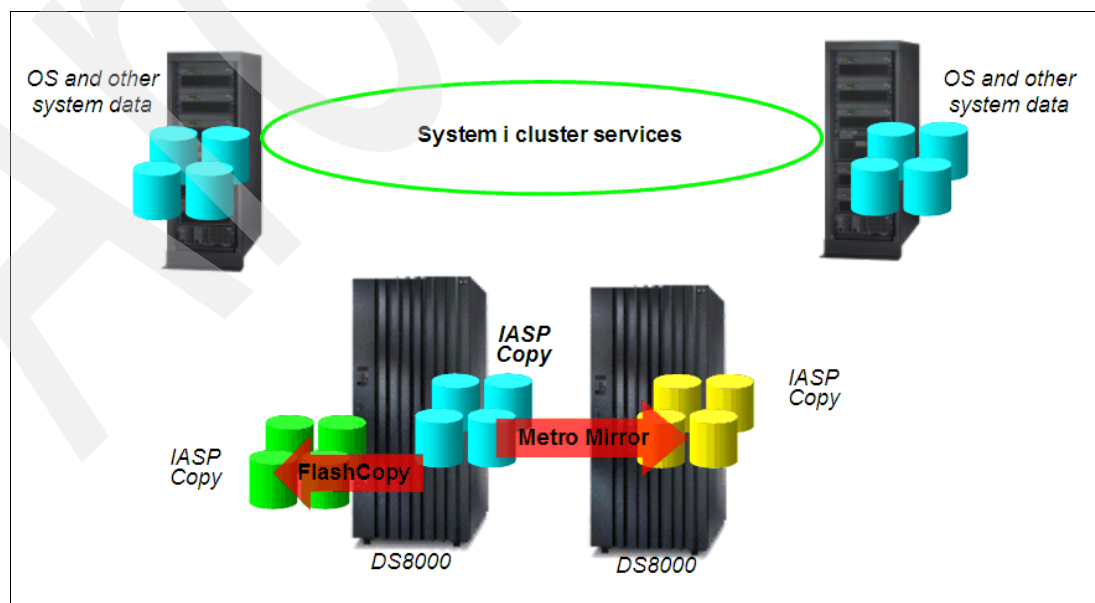


Figure 2-57 Server and storage clustering with Copy Services for System i and Metro Mirror or Global Mirror

Copy Services for System i *can* failover the environment through a single command. The **SWPPRC** command asks for confirmation that a site switch is really what is required. Once that confirmation is given, the command varies off the IASPs in their current location, and performs all commands necessary to bring them up in the alternate location. This process typically occurs within a matter of minutes.

To maintain data consistency, Copy Services for System i requires that all data be sent across all available links. As usual in Business Continuity disk mirroring scenarios, it is recommended to configure alternate fiber routes.

Utility for validation of copy environment

Obviously, validation is critical in a Business Continuity solution, to ensure that the data is being copied correctly, and is recoverable. To ensure that the environment is usable and recoverable, Copy Services for System i includes a program called **CHKPPRC**.

When **CHKPPRC** is executed, it validates that the System i servers, their HMCs, the disk systems, remote copy links, and remote copy technology are properly functioning. Typically this occurs every couple of hours.

2.7 Metro Cluster for N series

For N series, continuous availability is handled through integrated tools. Specifically, N series utilizes clustering functionality known as Metro Cluster, which builds on synchronous mirror technology in order to move data from one N series to another.

SyncMirror for Metro Cluster

SyncMirror® more closely resembles logical volume mirroring than it does other forms of block level hardware mirroring. Typically, it is used within an N series, writing to a pair of disks.

In Metro Cluster, as demonstrated in Figure 2-58, rather than writing to the disk and then allowing the disk system to synchronously mirror the data to the recovery disk system, the disk controller issues two writes at once; One write goes to the local disk while the other goes to the remote disk. Both N series can be active at the same time and would each write to the other. In order to recover in the case of a failure event, each N series carries a dormant Operating System image of their counterpart. During a failure event, that alternate image would activate and access the recovery volumes locally.

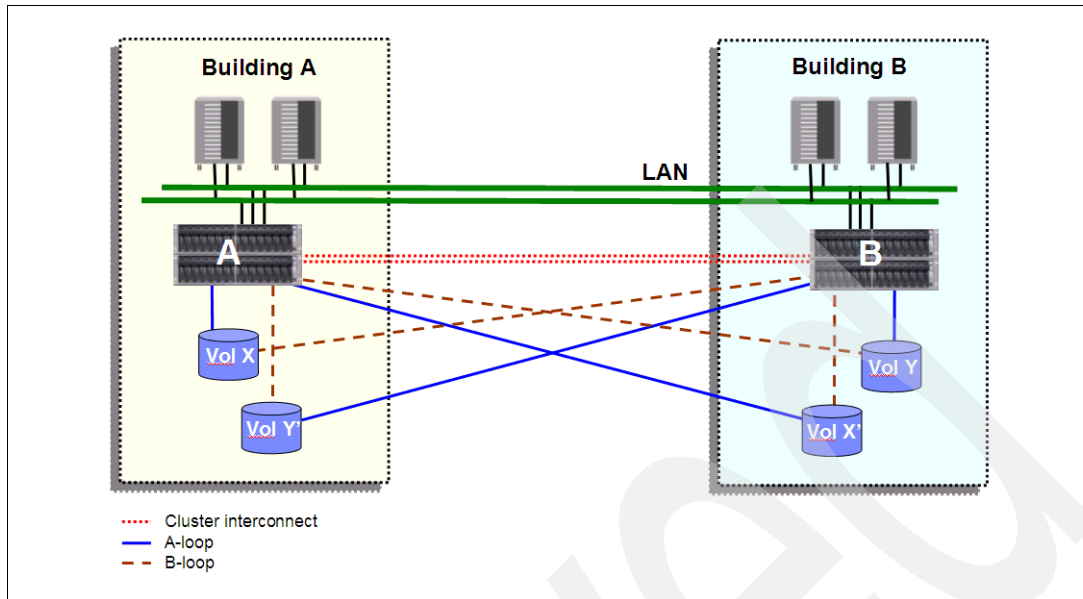


Figure 2-58 Sync Mirror environment for Metro Cluster.

Metro Cluster failover

Should a failure occur, the receiving N series activates the dormant copy of the production system's operating system. Within a matter of minutes, the second N series is available and appears identical to the now disabled N series (Figure 2-59).

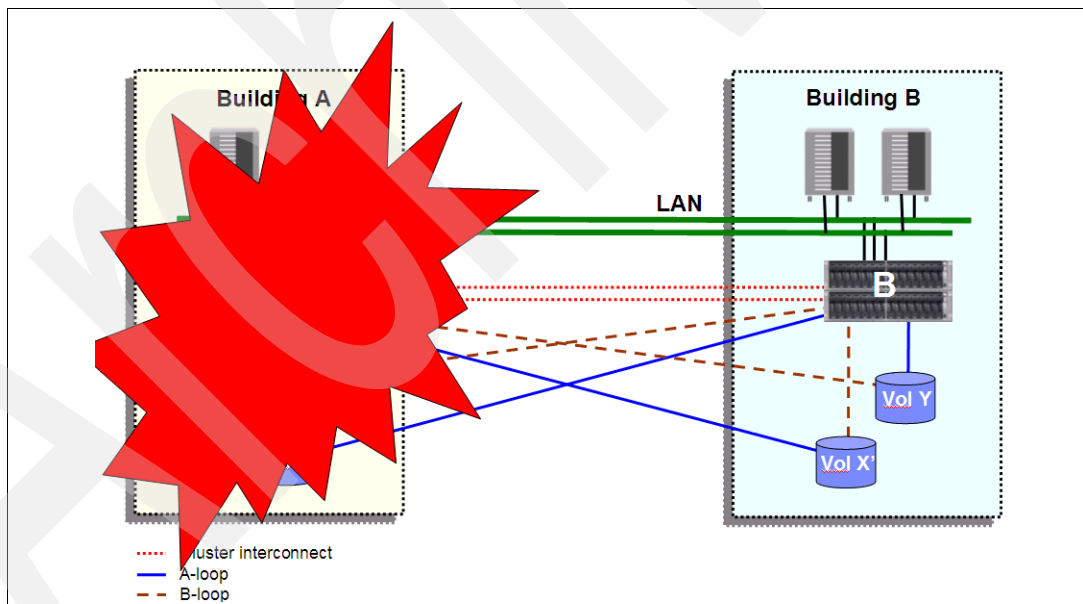


Figure 2-59 Metro Cluster failover to site B

With proper TCP/IP or SAN infrastructure, this might allow applications to continue accessing storage without the necessity for any particular failover procedure. This is possible because the TCP/IP address would remain the same and in a proper switched environment the image of N series 2 would appear to be N series 1 from a server perspective.

Metro Cluster is available on N5000 and N7000 series.

Metro Cluster types

There are two types of Metro Clusters available, depending on the environment and distance to be covered.

- ▶ **Stretch Cluster** is used only for short distances within a campus environment. Stretch Clusters allow you to connect a pair of N series systems as a supported distance of up to 300m. In this case, the connections are direct fibre channel protocol (FCP) connections between the two N series.
- ▶ **Fabric Cluster** is used for longer distances. Fabric Clusters allow the N series to be separated by up to 100km, but that is across a switched FCP infrastructure as opposed to direct connections permitted in a stretch cluster.

In either case, the Metro Cluster handles the switch only at the disk system level. Any failover that is required at the server level requires separate clustering technology.

Archived

Rapid Data Recovery

Rapid Data Recovery is based on maintaining a second disk-resident copy of data that is consistent at a point-in-time as close to the time of a failure as possible. This consistent set of data allows for the restart of systems and applications without having to restore data and re-applying updates that have occurred since the time of the data backup. It is possible that there might be a loss of a minimal number of in-flight transactions.

Rapid Data Recovery spans Tier 4 through Tier 6. Rapid Data Recovery is distinguished from Continuous Availability by the fact that the automation required to be a Tier 7 solution is not present.

In this chapter we describe Rapid Data Recovery in the following environments:

- ▶ System z and mixed z + distributed (GDPS)
- ▶ System z, mixed z + distributed, and distributed (TPC for Replication)
- ▶ UNIX and Windows (SAN Volume Controller)
- ▶ System i
- ▶ FlashCopy Manager and PPRC Migration Manager

3.1 System Storage Rapid Data Recovery: System z and mixed z+Open platforms (GDPS/PPRC HyperSwap Manager)

In this section we provide an overview of the GDPS/PPRC HyperSwap Manager, to provide Rapid Data Recovery for System z and mixed z+Open platforms. We cover these topics:

- ▶ Resiliency Family positioning
- ▶ Description
- ▶ Value
- ▶ Components

More information about the complete family of Geographically Dispersed Parallel Sysplex (GDPS) Business Continuity offerings can be found in 2.1, “Geographically Dispersed Parallel Sysplex (GDPS)” on page 10.

3.1.1 Resiliency Portfolio Positioning

Used in a two site implementation, the GDPS/PPRC HyperSwap Manager provides a Rapid Data Recovery Tier 6 Business Continuity solution for System z and z+Open data; see Figure 3-1. Note that it falls short of being a Tier 7 solution due to the lack of System z processor, System z workload, and Coupling Facility recovery automation provided by a full GDPS/PPRC implementation.

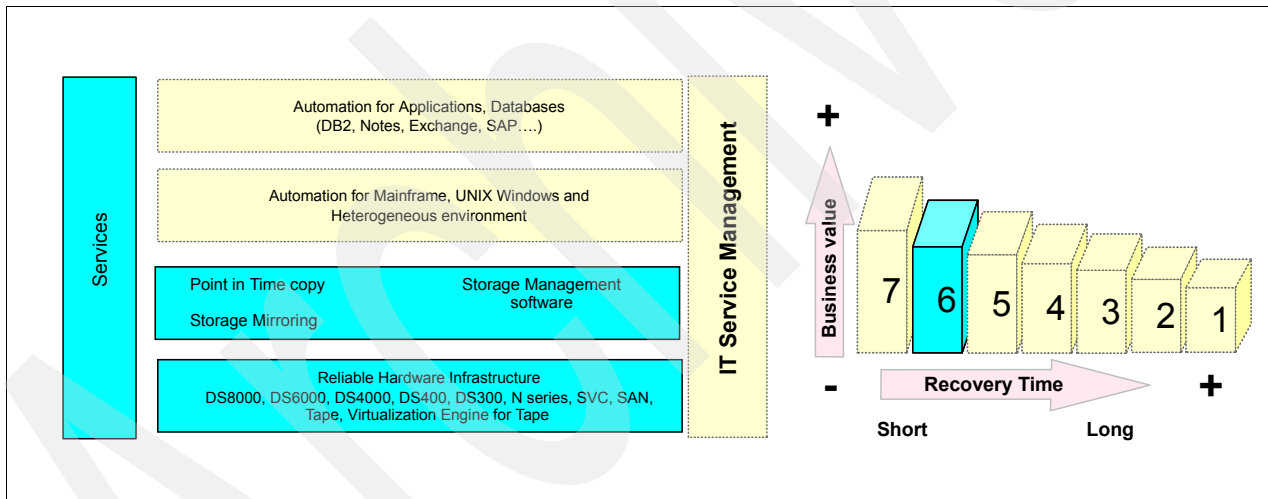


Figure 3-1 System Storage Rapid Data Recovery for System z

3.1.2 Description

Rapid Data Recovery for System z is provided by an IBM Global Services service offering, GDPS/PPRC HyperSwap Manager (GDPS/PPRC HM), in the GDPS suite of offerings. It uses IBM System Storage Metro Mirror (previously known as Synchronous PPRC) to mirror the data between disk subsystems. Metro Mirror is a hardware-based mirroring and remote copying solution for the IBM System Storage DS8000, DS6000, and ESS (see Figure 3-2).

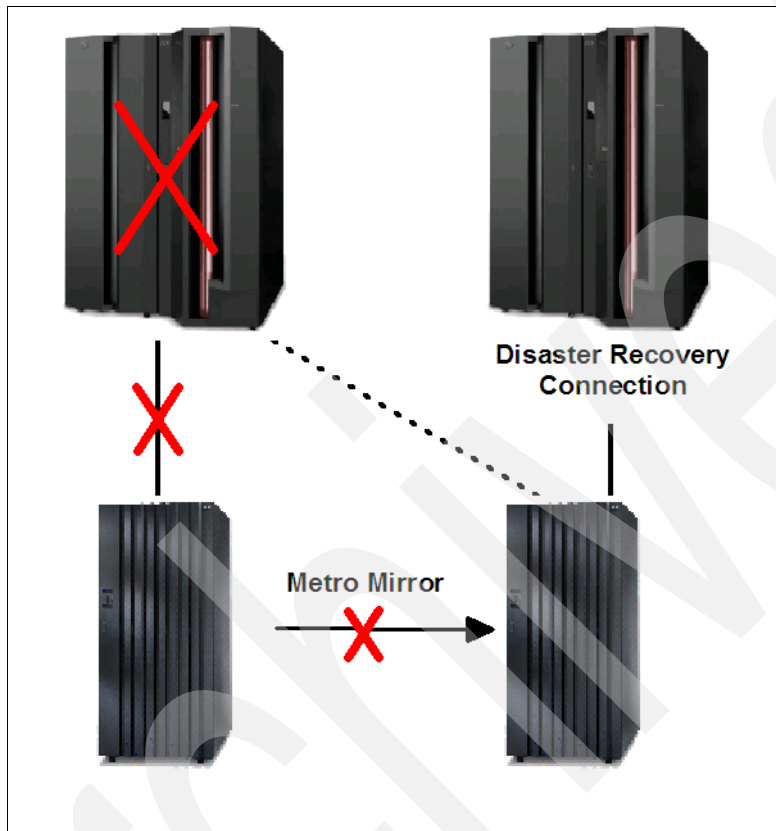


Figure 3-2 IBM System Storage Metro Mirror

Continuous Availability for System z data

GDPS/PPRC HyperSwap Manager is primarily designed for single site or multiple site System z environments, to provide Continuous Availability of disk-resident System z data by masking disk outages due to failures. Planned outage support is also included, for example, for planned disk maintenance. GDPS/PPRC HM would be considered a Tier 6 solution as the automated recovery is extended to the disk mirroring, but is not present for the System z processor, System z workloads, or Coupling Facilities.

When a disk failure occurs, GDPS/PPRC HM invokes HyperSwap to automatically switch disk access of System z data to the secondary disk system; see Figure 3-3. When a primary disk outage for maintenance is required, GDPS/PPRC HM provided user interface panels can be used to invoke a HyperSwap switch of System z data access to the secondary disks. After the disk repair or maintenance has been completed, HyperSwap can be invoked to return to the original configuration.

Disk Reconfiguration with HyperSwap

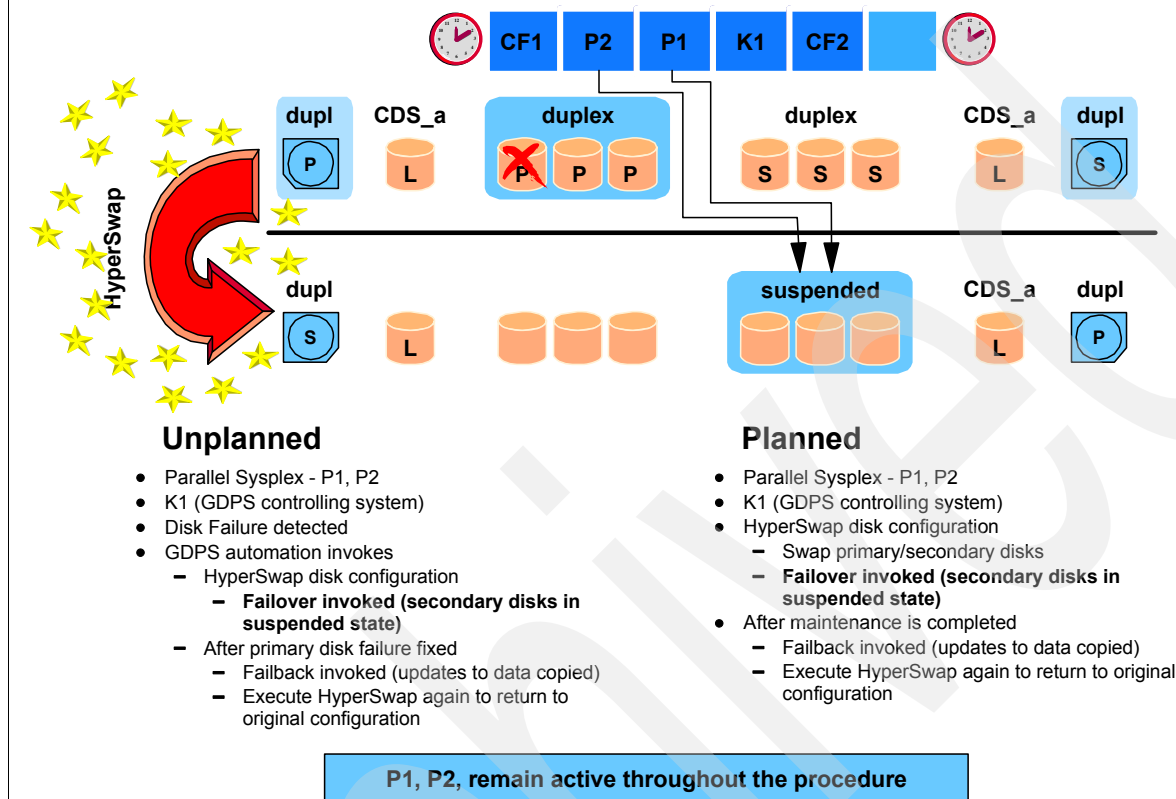


Figure 3-3 HyperSwap disk reconfiguration

When a disaster occurs at the primary site, GDPS/PPRC HyperSwap Manager can swap to the secondary disk system(s), at the same time assuring data consistency at the remote site via GDPS/PPRC HyperSwap Manager's control of Metro Mirror Consistency Groups.

After GDPS/PPRC HyperSwap Manager has swapped to the alternate disk systems, the operations staff must take the necessary steps to restore a production environment using the secondary disk system; see Figure 3-4. This includes such activities as:

- ▶ Powering up processors
- ▶ IPLs with alternate startup parameters and configurations
- ▶ Making network changes
- ▶ Restarting subsystems and applications

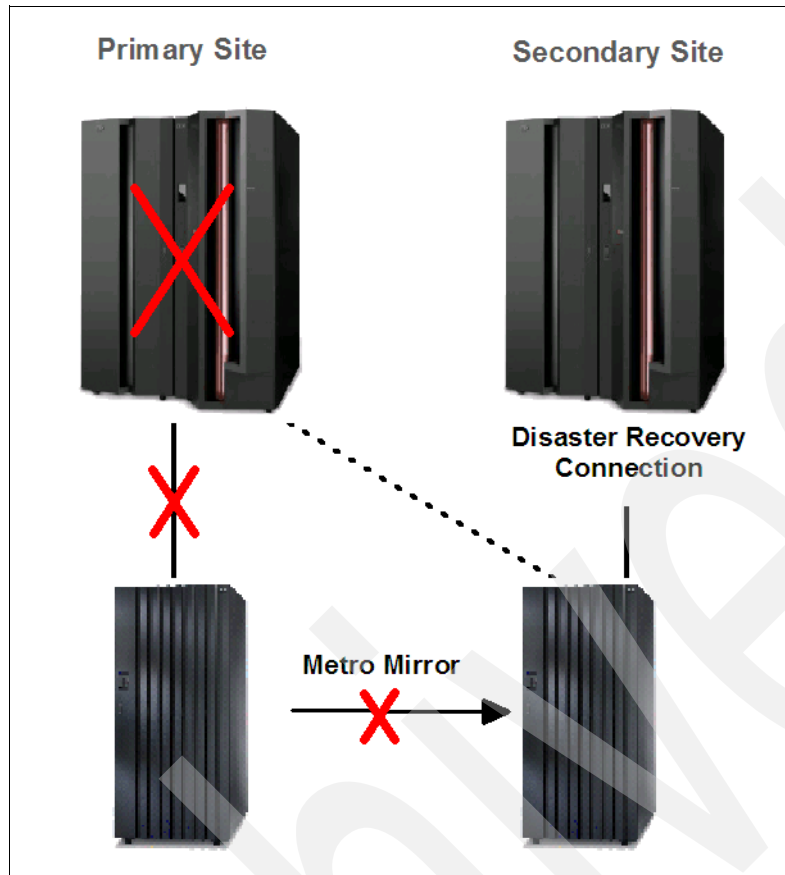


Figure 3-4 Disaster recovery configuration

GDPS/PPRC Open LUN management

When Metro Mirror is implemented for DS8000, DS6000 and ESS disk with System z data, the GDPS Open LUN Management function also allows GDPS to manage the Metro Mirroring of open systems data within the same primary disk system.

After HyperSwap is done, open systems servers must be restarted. The open systems data is assured to be data consistent and time consistent with all other volumes or LUNs in the Metro Mirror Consistency Group.

When HyperSwap is invoked due to a failure or by a command, full support of the Consistency Group FREEZE of the open data takes place to maintain data consistency.

GDPS/PPRC HyperSwap Manager increases business resiliency by extending Parallel Sysplex availability to disk systems. System z application outages due to a disk subsystem failure or planned maintenance are avoided by invoking HyperSwap to direct disk access to the secondary disk system.

GDPS/PPRC HyperSwap Manager reduces the time required to implement disk mirroring, and simplifies management of a Metro Mirror configuration by providing a single point of control, thus reducing storage management costs.

In conjunction with specially priced Tivoli NetView and System Automation products, GDPS/PPRC HyperSwap Manager provides an affordable entry level offering where high levels of availability are required, but not all of the functions provided by the full GDPS/PPRC offering. GDPS/PPRC HyperSwap Manager provides a lower cost option for providing continuous or near-continuous access to data within a Sysplex. GDPS/PPRC HyperSwap Manager can be used to help eliminate single points of failure within the disk systems as well as provide availability across planned disk maintenance.

Product components

Rapid Data Recovery for System z requires the following components:

- ▶ IBM System z servers
- ▶ IBM System Storage DS8000, DS6000, ESS and other Metro Mirror level 3 disk systems
- ▶ IBM System Storage Metro Mirror and FlashCopy
- ▶ IBM Tivoli NetView and System Automation
- ▶ IBM Global Services (GDPS/PPRC HyperSwap Manager and implementation)

3.1.3 Additional information

For additional information, see the GDPS/PPRC HyperSwap Manager IBM announcement letter at:

<http://www.ibm.com/servers/eserver/zseries/zos/news.html>

Or, visit the GDPS Web site:

<http://www.ibm.com/systems/z/gdps/>

3.2 System Storage Rapid Data Recovery for UNIX and Windows

Next, we turn our attention to Rapid Data Recovery in a UNIX/Windows mixed environment, when disk mirroring is used to replicate the data.

In a Business Continuity solution based on disk remote mirroring, we want to restart a database application following an outage without having to restore the database. This process has to be consistent, repeatable and fast, measurable in minutes. A database recovery where the last set of image copy tapes is restored, then the log changes are applied to bring the database up to the point of failure, is a process that can be measured in hours or even days.

You want to avoid this situation in a Tier 4 or higher solution. However, in a real disaster (fire, explosion, earthquake) you can never expect your complex to fail all at the same moment. Failures are intermittent and gradual, and the disaster can occur over many seconds or even minutes. This is known as a *Rolling Disaster*. A viable disk mirroring disaster recovery solution must be architected to avoid data corruption caused during a Rolling Disaster.

In any operating system, the sequence in which updates are being made is what maintains the integrity of the data. If that sequence is changed, data corruption occurs. The correct sequence must be maintained within a volume, across volumes, and across multiple storage devices.

For instance, in Figure 3-5 we show the relationship between a database and its log, which demonstrates the requirement for maintaining I/O integrity. Data consistency across the storage enterprise must be maintained to insure data integrity.

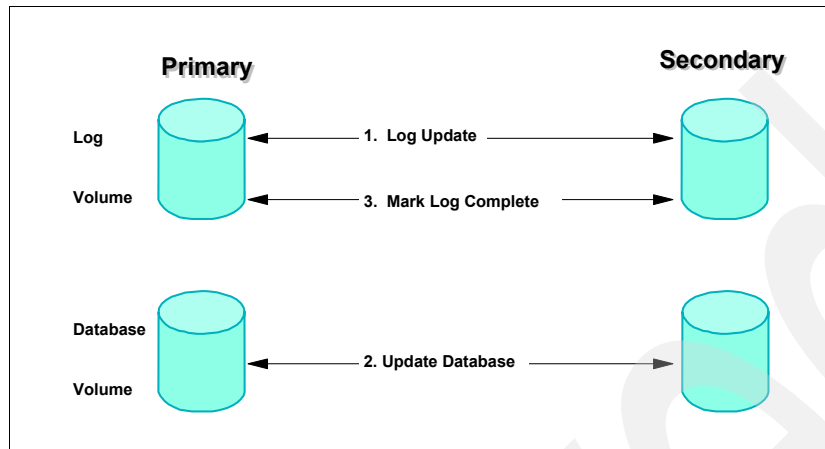


Figure 3-5 Sample update sequence in a database

The order of *Dependent Writes* across volumes must be maintained at remote locations. Failure to do so results in data corruption because the data consistency is lost.

In Figure 3-6 we illustrate this concept with an example. The intention to update the database is logged in the database log files at both the primary and secondary volumes (step 1). The database datafile is updated at the primary volume, but the update does not reach the remote volume which contains the mirrored datafile. The primary location is not aware of the write failure to the secondary volume (step 2). The database update is marked complete in the log files at both the primary and remote locations (step 3). The result is that the secondary site log files say the update was done, but the updated data is not in the database at the secondary location. There is no way to know that the data was corrupted.



For this reason it is strongly recommended that, at a minimum, *Consistency Groups* be put in place for any mirroring solution. Consistency Groups are an implementation of technology that assists with the consistency of application data capable of dependent writes. To guarantee a fully consistent remote copy, multiple volumes require a Consistency Group functionality. DS8000/DS6000/ESS Metro Mirror and Global Mirror microcode already has the Consistency Group function for both System z and open systems. The SAN Volume Controller Metro Mirror and DS4000 Global Mirror microcode has a Consistency Group function for open systems.

Figure 3-7 illustrates this mechanism. If a write cannot complete, the storage system does not back out incomplete transactions on its own. Instead, the application has to recognize that the transaction was incomplete and take the appropriate actions. Once the storage system holds application I/O to the affected primary volumes, the write dependent mechanism of the application prevents the Metro Mirror secondaries from becoming inconsistent.

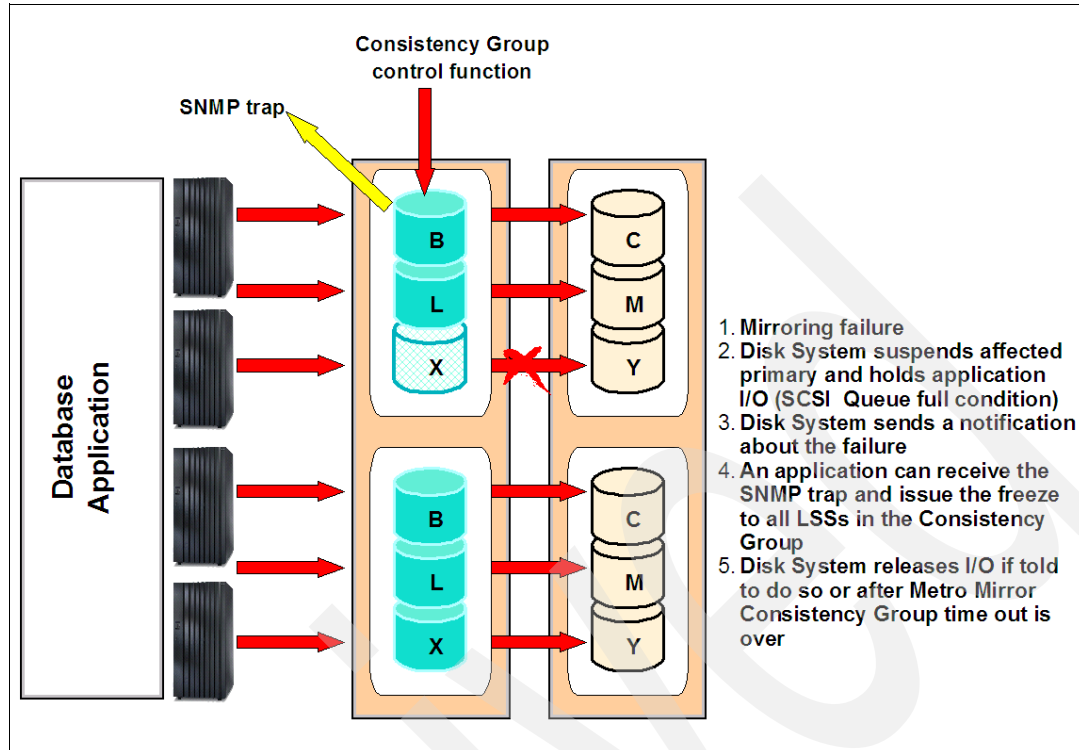


Figure 3-7 Dependent write protection of database integrity

The **freeze** command, which is available through TSO, the ESS API, ICKDS, or the DS8000 Copy Services Web User Interface, stops the write activity on all of the active Metro Mirror primary and secondary volumes of a given source and target Logical Subsystem (LSS) pair. This function ensures consistency between the primary and secondary volumes and can affect any LSS volumes that are in an active Metro Mirror state between the frozen LSS pair: duplex, duplex pending SYNC, or duplex pending XD states. It does not affect the suspended and simplex volumes that might be in the LSS. The **freeze** operation has three effects:

1. The paths that connect the pair of LSSs being frozen are terminated.
2. The active volumes under the frozen LSS pair are suspended. This state transition, to suspended, is then communicated to the host with alert messages. These alert messages can be used by automation routines to trigger other recovery operations.
3. If the consistency group option was enabled at path definition time, then the ELB or SCSI Queue Full condition is instigated, so that the write activity to the primary LSS is temporarily queued. During this interval, other automated operations can be triggered; such as, freezing other application-related LSS pairs.

When using **freeze** through ICKDSF or the Web interface, it takes effect on each LSS individually. This is useful for creating a point-in-time copy of the data, but because of slight delays between the issuing of each iteration of the **freeze** command, it is unsuitable for preventing rolling disasters and should be done at periods of low utilization to ensure the restartability of the secondary data.

When Metro Mirror is used in conjunction with automation, however, such as Geographically Dispersed Parallel Sysplex (GDPS) or TPC for Replication a **freeze** command can be simultaneously issued to all LSSs within the configuration. This ensures globally consistent data across all LSSs in the secondary copy of data during a disaster.

The main target of the Rapid Data Recovery for UNIX and Windows solution, based on the combination of DS8000, DS6000, and ESS Copy Services with TPC for Replication, is to protect data to a consistent point of recovery, avoiding a recover action. Remember, a Restart allows for a quick recovery, while a Recover might cause an extended recovery action potentially resulting in downtime.

Using DS8000 Metro Mirror Consistency Group, this solution freezes your environment at a known point instead of mirroring literally hundreds of time-offset failures in a short amount of time.

More information about the **Freeze** and **run** commands can be found in the IBM Redbooks: *IBM System Storage DS8000 Series: Copy Services with IBM System z*, SG24-6787 and *IBM System Storage DS8000 Series: Copy Services in Open Environments*, SG24-6788.

3.3 System Storage Rapid Data Recovery for System z and mixed z+Distributed platforms using TPC for Replication

Rapid Data Recovery for System z and Distributed systems is provided by TPC for Replication, which is an effective, user friendly GUI offering for defining and managing the setup of data replication and recovery. The TPC for Replication solution has no dependencies on the operating system because no client installation is required; it is operating system agnostic.

In environments with mixed workloads, System z, UNIX, Windows, etc. TPC for Replication provides a single common GUI to control the entire environment from the same set of panels, common across different operating systems. This simplifies management and control of the entire environment.

The IBM TotalStorage Productivity Center for Replication V3.1 is designed to manage the advanced copy services provided by IBM Enterprise Storage Server (ESS) Model 800, IBM System Storage DS8000, IBM System Storage DS6000 and IBM SAN Volume Controller (SVC). It is available in two complementary packages:

- ▶ **IBM TotalStorage Productivity Center for Replication V3.1** offers the following basic functions:
 - Manages advanced copy services capabilities for the IBM ESS Model 800, IBM DS8000, IBM DS6000, and IBM SVC. FlashCopy, Metro Mirror, and Global mirror support (global mirror is supported for ESS, DS8000, DS6000 only) with a focus on automating administration and configuration of these services, operational control (starting, suspending, resuming) of copy services tasks, and monitoring and managing the copy sessions.
 - Monitors the progress of the copy services so you can verify the amount of replication that has been done as well as the amount of time required to complete the replication operation.
 - Manages one way (single direction) data replication to protect you against primary site failure.
- ▶ **IBM TotalStorage Productivity Center for Replication Two Site BC V3.1** helps to aid Business Continuity by adding the following capabilities when used with IBM TotalStorage Productivity Center for Replication, its prerequisite:
 - Designed to provide disaster recovery management through planned and unplanned failover and fallback automation for the IBM ESS Model 800, IBM DS8000, and IBM DS6000.

- Designed to provide high availability capability for the IBM TotalStorage Productivity Center for Replication V3.1 server, to help manage your replication even if the main IBM TotalStorage Productivity Center for Replication server experiences a failure.

In the following sections we provide an overview of TPC for Replication to provide System Storage Rapid Data Recovery for System z and mixed z+Open platforms, covering these topics:

- ▶ System Storage Resiliency Portfolio positioning
- ▶ Solution description
- ▶ Functionality of TPC for Replication
- ▶ Components
- ▶ Terminology
- ▶ Additional information

3.3.1 System Storage Resiliency Portfolio positioning

The basic functions of TPC for Replication provide management of FlashCopy, Metro Mirror, and Global Mirror capabilities for the IBM DS8000, IBM DS8000, IBM ESS, and IBM SAN Volume Controller.

TPC for Replication can simplify management of advanced copy services by:

- ▶ Automating administration and configuration of these services with a wizard-based session and copy set definitions
- ▶ Providing simple operational control of copy services tasks, including starting, suspending and resuming
- ▶ Offering tools for monitoring and managing copy sessions.

TPC for Replication provides a Rapid Data Recovery Tier 6 Business Continuity solution for System z and z+Open data. See Figure 3-8.

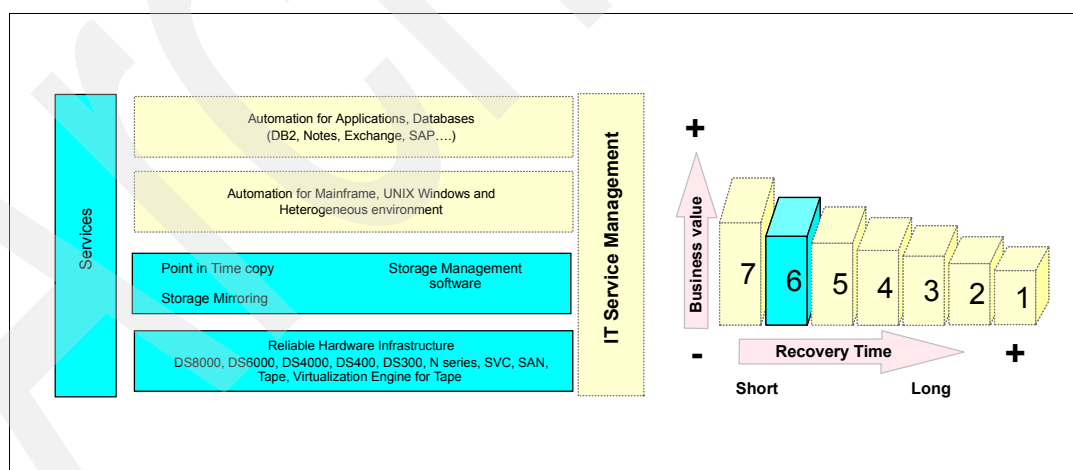


Figure 3-8 System Storage Rapid Data Recovery for System z and distributed environments

3.3.2 Solution description

The System Storage Rapid Data Recovery for z/OS, UNIX, and Windows solution, based on TotalStorage Productivity Center for Replication, is a multisite disaster recovery solution that manages data on volumes for Distributed Systems as well as data for System z.

TPC for Replication manages the advanced copy services provided by IBM DS8000, DS6000, ESS800, and SAN Volume Controller (SVC). It is available in two complementary packages, shown in Figure 3-9:

► **TPC for Replication: Basic (one direction)**

Provides a one way disaster recovery management of disk advanced copy services. It includes FlashCopy, Metro Mirror, and Global Mirror support (Global Mirror is supported for ESS, DS8000, DS6000 only). It focuses on automating administration and configuration of these services, operational control, (starting, suspending, and resuming) of copy services tasks, and monitoring and managing the copy sessions.

► **TPC for Replication Two Site Business Continuity**

Provides failover and failback disaster recovery management through planned and unplanned outage for the DS8000, DS6000, and ESS Model 800. It monitors the progress of the copy services to verify the amount of replication that has been done as well as the amount of time required to complete the replication operation. TPC for Replication Two Site BC requires TPC for Replication base as a prerequisite.

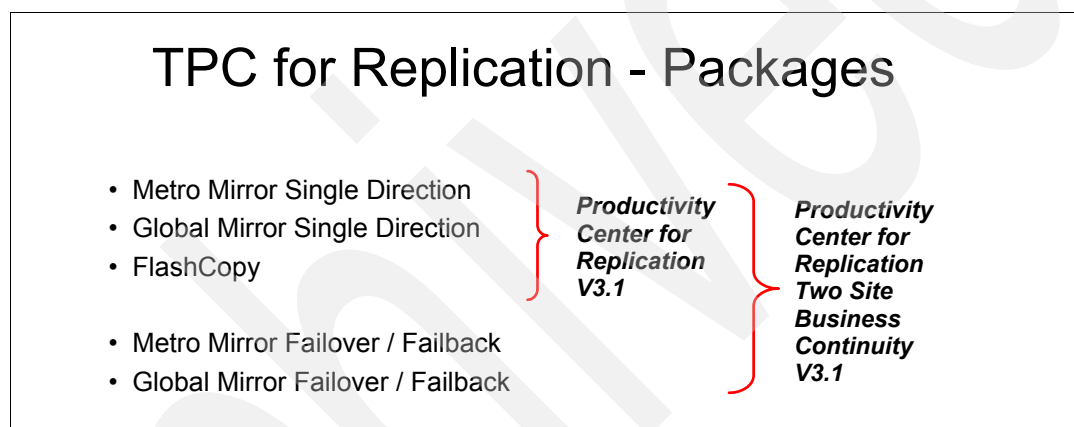


Figure 3-9 TPC for Replication packages

TPC for Replication, Two Site BC, can provide high availability capability through a standby TPC for Replication server. The standby server takes over management of the environment, through manual intervention if the primary TPC for Replication server experiences a failure

TPC for Replication manages advanced copy services from multiple IBM disk systems using a single interface, common functions, and terminology. The GUI simplifies all functions of copy services configurations through easy-to-use, guided wizards and optional, automated source and target volume matching. Copy services configurations are stored persistently in a database, so that manual or automated commands can be executed against a configuration at any time.

TPC for Replication greatly simplifies advanced copy services operations by combining multiple hardware-level commands into a single GUI selection or CLI command. For example, for a Global Mirror environment, a single GUI “Start” action:

1. Uses existing Global Copy paths or establish required paths
2. Establishes Global Copy pairs
3. Automatically establishes Flash Copy pairs at the secondary disk system/s
4. Defines the Global Mirror session to all primary Logical Disk Systems
5. Adds volumes to the Global Mirror session
6. Starts Global Mirror consistency formation

TPC for Replication includes built-in automation to maintain consistency in failure situations. Also, TPC for Replication allows for additional customized automation of copy services configuration and operations.

TPC for Replication provides FlashCopy, Metro Mirror and Global Mirror support for the IBM DS8000, DS6000 and ESS800 with Copy Services, and FlashCopy and Metro Mirror support for the IBM SAN Volume Controller. TPC for Replication manages uni-directional data replication, from the primary site to the secondary site.

TPC for Replication Two Site Business Continuity provides additional functions of a high availability TPC for Replication server configuration and incremental resynchronization of remote mirroring volumes from the secondary site to the primary site. Here, TPC for Replication manages the data replication as well as supports failover and failback from site A to site B and vice versa.

For the high availability configuration for both cases above, management of uni-directional replication and management of the Two Site Business Continuity, we strongly recommend that you consider providing two TPC for Replication Servers that back up each other in a standby mode. The two TPC for Replication servers continually update their respective copies of the database of the environment and keep it in synch through regular heartbeat to each other. This means that in an event that results in failure of the primary TPC for Replication server, the standby server takes over the management of the replication environment once a decision is made to switch to that server.

3.3.3 Functionality of TPC for Replication

TPC for Replication includes the following functionality:

- ▶ Managing and configuring copy services environment:
 - Add/delete/modify storage devices
 - Add/delete/modify copy sets (a copy set is a set of volumes containing copies of the same data)
 - Add/delete/modify sessions (a container is a container of multiple copy sets managed by a replication manager)
 - Add/delete/modify logical paths (between storage devices)
- ▶ Monitoring the copy services environment:
 - View session details and progress
 - Monitor sessions - with status indicators and SNMP alerts
 - Diagnostics - error messages

3.3.4 Functionality of TPC for Replication Two Site BC

TPC for Replication Two Site BC includes the following functionality:

- ▶ All functions of TPC for Replication
- ▶ Failover and failback from primary to a disaster recovery site
- ▶ Support for IBM TotalStorage Productivity Center for Replication high-availability server configuration

3.3.5 Environment and supported hardware

TPC for Replication and TPC for Replication Two Site BC require a separate server for the application (or two if using a standby server). The server can run in Windows, Linux, or AIX. For specific supported operating systems, for the supported and detailed server hardware requirements, see:

<http://www.ibm.com/support/docview.wss?rs=1115&uid=ssg1S1002892>:

The server runs and maintains an internal DB2 UDB database. For specific supported versions, see:

<http://www.ibm.com/support/docview.wss?rs=1115&uid=ssg1S1002893>

For detailed requirements on the supported disk storage systems, see:

<http://www.ibm.com/servers/storage/software/center/replication/interop.html>

3.3.6 Terminology

TPC for Replication uses a set of terms which we define in the following sections.

Copy set

A set of volumes that represent copies of the same data. All volumes in a copy set must be of the same type and size. The number of volumes in a copy set and the roles that each volume in a copy set plays in the replication session is determined by the session policy. For an illustration, see Figure 3-10.

Session

A session is a collection of copy sets making up one consistency group, as shown in Figure 3-10.

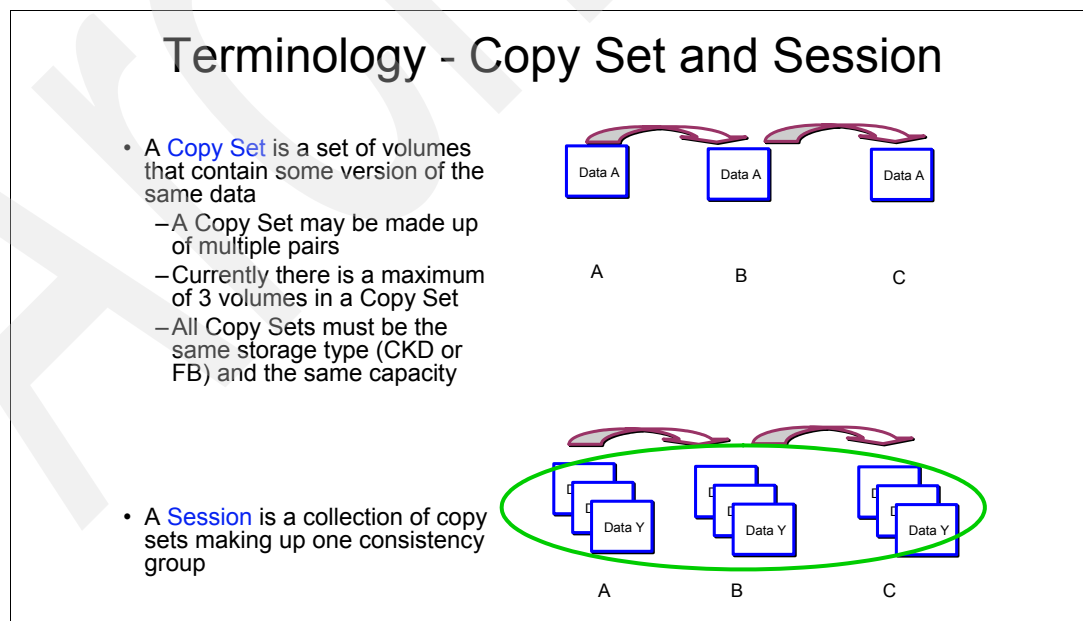


Figure 3-10 Copy Set and Session

Session states

A session can be in one of several states, as shown in Figure 3-11.

- ▶ **Defined:**
 - After creation or termination of a session, the session exists, but is inactive.
- ▶ **Preparing:**
 - The volumes are initializing or resynchronizing. Resynchronization occurs in Metro Mirror and Global Mirror when a volume was prepared once and then suspended: the hardware records the changed tracks so that on the next startup, only the changed tracks are copied. In FlashCopy, this means that the volumes are initializing. The Preparing state for FlashCopy sessions applies only to SVC.
- ▶ **Prepared:**
 - Source to target data transfer is active. In Metro Mirror and Global Mirror, this means that the data written to the source is transferred to the target, and all volumes are consistent and recoverable. In FlashCopy, this means that the volumes are not yet consistent, but the flash is ready to begin.
- ▶ **Suspending:**
 - This state marks the transition into a Suspended state.
 - It is caused by a Suspend command or suspending event.
 - It indicates that the session is in the process of suspending copy operations.
 - This state applies only to Global Mirror sessions and does not apply to SVC.
- ▶ **Suspended:**
 - Data copying has temporarily stopped - no session level copying is occurring.
 - Application writes can continue (freeze and run). Changes are tracked.
 - The Suspended state applies only to Global Mirror and Metro Mirror sessions.
- ▶ **Recovering:**
 - The session is in the process of recovering.
 - The target volumes are write enabled.
 - It automatically becomes Target Available when recovery is complete.
- ▶ **Target available:**
 - Recover command processing has completed. The target volumes are write enabled and available for application updates.

It is important to understand these transitions, because these can and do at times determine which TPC for Replication commands are required to move to the next state.

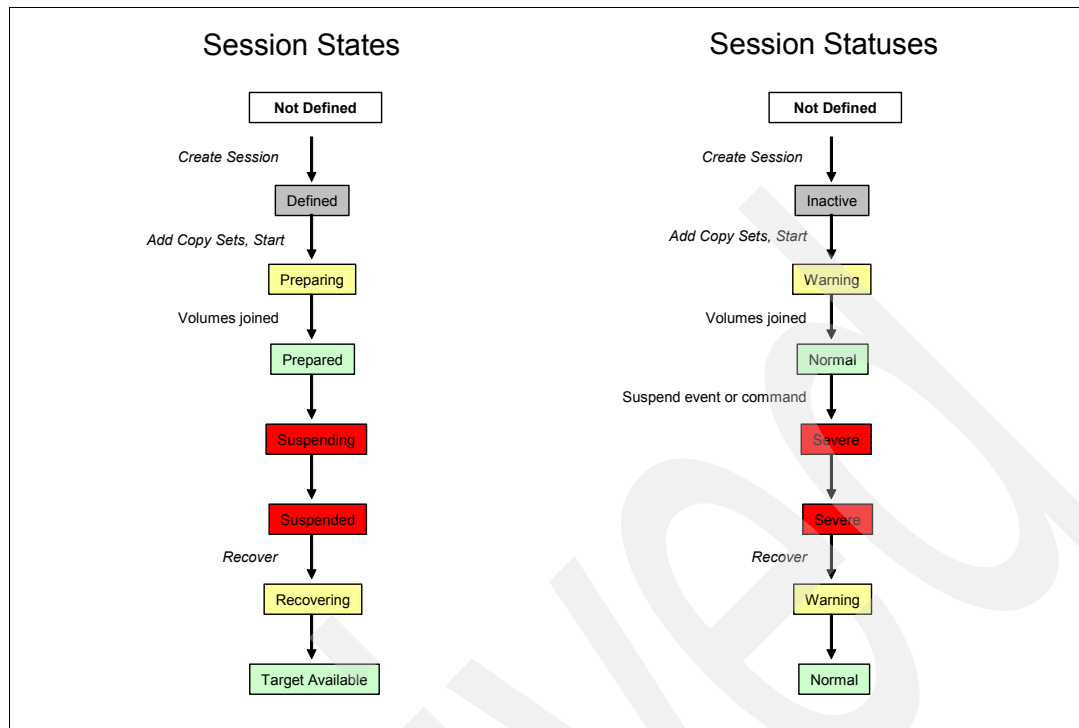


Figure 3-11 Session states and session status

Session status

A session can be in one of several states, as shown in Figure 3-11.

- ▶ **Normal** (green with checkmark)
A consistent copy of the data either exists, or is being maintained.
- ▶ **Warning** (yellow with !)
For Metro Mirror and Global Mirror, the session might have volumes that are being synchronized or about to be synchronized, with no suspended volumes. For FlashCopy, the Warning status is valid only after the start command is issued and before the flash, and means that the session is either preparing or is ready for a flash command but targets do not yet have a consistent copy.
- ▶ **Severe** (red with X)
One or more errors must be dealt with immediately. Possible causes include the following:
 - One or more volumes are suspended.
 - A session is suspended.
 - A volume is not copying correctly.
- ▶ **Inactive** (grey)
The session is in a defined state, with no activity on the hardware.

Roles

The role indicates the function of a volume in a copy set. It indicates the intended use of a volume (host volume, target volume, or journal volume), and the location of the volume; for example, the primary (1) or secondary (2) site.

A *host volume* (H) is a volume that an application, such as a database, reads data to and from. A host volume can be the source volume when the application connected to the host volume is actively issuing read and write input/output (I/O). A host volume can also be the target of the copy function, in which case you cannot write to the volume.

A *target volume* (T) is for FlashCopy only, and is an intermediate volume that receives data from a source. Depending on the session type, that data might or might not be consistent.

A *journal volume* (J) holds a consistent copy of the data until a new consistent copy is formed. The journal volume restores the last consistent point during a recovery.

Role pairs

A role pair is the association of two volumes/roles in a session. For example, in a Metro Mirror session, the role pair can be the association between the volume roles Host1 and Host2.

Figure 3-12 shows the terms we have defined, and their relationship with each other.

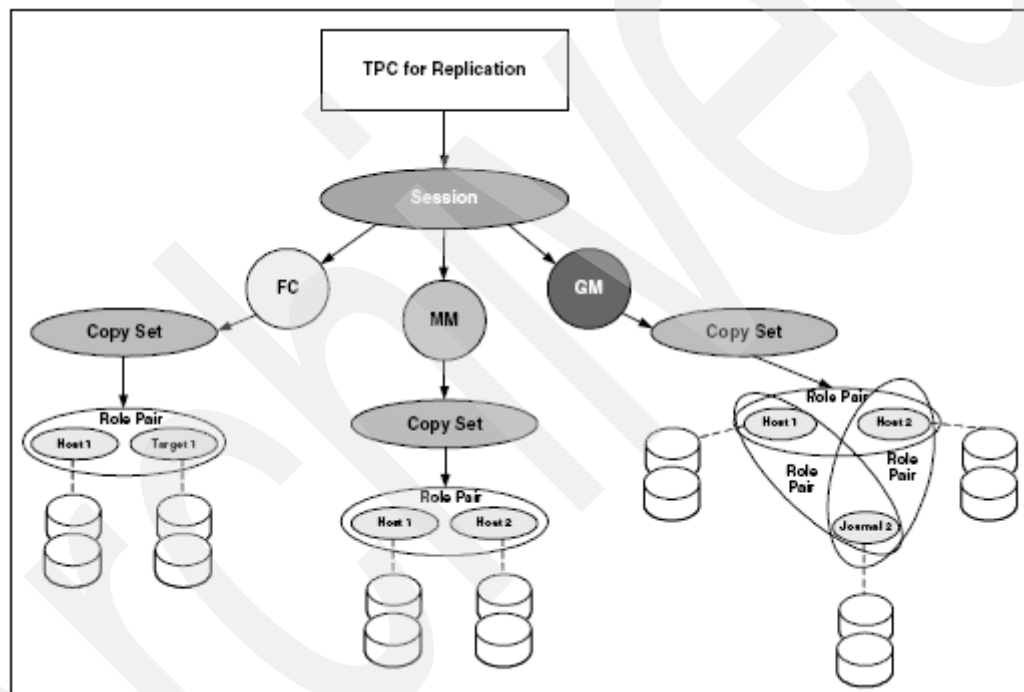


Figure 3-12 TPC for Replication Terminology

3.3.7 TPC for Replication session types and commands

TotalStorage Productivity Center for Replication enables you to configure a number of different sessions with different copy types as follows:

- ▶ FlashCopy
- ▶ Metro Mirror Single Direction
- ▶ Global Mirror Single Direction
- ▶ Metro Mirror Failover / Failback
- ▶ Global Mirror Failover / Failback

Session Commands

Table 3-1 shows the commands that can be issued against any defined session.

Table 3-1 TPC for Replication commands

Command	Meaning
Flash	Perform the FlashCopy operation using the specified options.
Initiate Background Copy	Copy all tracks from the source to the target immediately, instead of waiting until the source track is written to. This command is valid only when the background copy is not already running.
Start	Perform any steps necessary to define the relationship before performing a FlashCopy operation. For ESS/DS, you do not have to issue this command. For SVC, use this command to put the session in the prepared state.
Recover	Issued to suspended sessions. Performs the steps necessary to make the target available as the new primary site. On completion of this command, the session becomes Target Available. This command does not apply for SVC.
Start	Sets up all relationships in a single-direction session and begins the process to start forming consistency groups on the hardware.
Start H → H2	Indicates direction between two hosts in a Global Mirror failover/failback session. Suspend stops all consistency group formation when the data is actively being copied. This command can be issued at any point during a session when the data is actively being copied.
Start H2 → H1	Indicates direction of a failover/failback session. If a recover has been performed on a session such that the production site is now H2, you can issue this command to start moving data back to Site 1. However, this start does not provide consistent protection, because it copies back only asynchronously because of the long distance. When you are ready to move production back to Site 1, issue a suspend to the session — this puts the relationships into a synchronized state and suspends them consistently. This command is not supported for SVC.
Terminate	Removes all physical copies from the hardware. Can be issued at any point in an active session. If you want the targets to be data consistent before removing their relationship, you must issue the Suspend command, the Recover command, and then the Terminate command.

TPC for Replication GUI

The TPC for Replication GUI presents a single point of control to configure, manage, and monitor copy services. The GUI reports on the status and availability of the administration components as well as management information for the established copy operations in real-time.

Monitoring includes:

- ▶ Overall session status: Indicates the session status, which can be normal, warning, or severe.
- ▶ Overall storage subsystem status: Indicates the connection status of the storage system.
- ▶ Management server status (applicable only if you are using the BC license).

Status indicators are used to simply describe the various states for defined TPC for Replication components. In addition, various icons are used to represent the status. These are:

- ▶ **Green:** TPC Copy Services is in “normal” mode. Session is in Prepared state for all defined volumes and maintaining a current consistent copy of the data. Or, the session has successfully processed a Recover command and is in TargetAvailable state with all volume consistent and no exceptions.
- ▶ **Yellow:** TPC Copy Services is not maintaining a current consistent copy at this time but is working toward that goal. In other words, sessions might have volumes that are actively being copied or pending to be copied, there are no suspended volumes and copy services is “temporarily” inconsistent but actions are in place to come into duplex state. No action is required to make this become Green, because states do automatically change the session to Green, without client interaction.
- ▶ **Red:** TPC Copy Services has one or more exceptions that have to be dealt with immediately. This could be one or more suspended volumes, a down session (both planned and unplanned), or a volume that should be copying and for some reason is not.

Note: With two TPC for Replication servers running, and if you are logged on to the active server, this indicates the status of the standby server. Conversely, if you are logged on to the standby server, this indicates the status of the active server.

3.3.8 Additional information

For additional information, you can refer to the following publications:

- ▶ *IBM TotalStorage Productivity Center for Replication: Installation and Configuration Guide*, SC32-0102
- ▶ *IBM TotalStorage Productivity Center for Replication: User's Guide*, SC32-0103
- ▶ *IBM TotalStorage Productivity Center for Replication: Command-Line Interface User's Guide*, SC32-0104
- ▶ *IBM TotalStorage Productivity Center for Replication: Quick Start Guide*, GC32-0105
- ▶ *IBM TotalStorage Productivity Center for Replication: Two Site BC Quick Start Guide*, GC32-0106
- ▶ *IBM TotalStorage Productivity Center for Replication on Windows 2003*, SG24-7250

3.3.9 TPC for Replication architecture

With TPC for Replication, you can have a single consistency group across disparate environments, provided that they are in the same TPC for Replication session. You can also choose to have separate consistency groups for different environments. In this case you would have different sessions for each environment.

TPC for Replication is a scalable and flexible solution that protects business data and maintains consistent restartable data. It can be used to manage copy relationships for both of the following types of outages:

- ▶ Planned outages such as hardware and software upgrades
- ▶ Unplanned outages such as an actual disaster

Solution highlights

Here we summarize the main solution highlights and benefits. This solution:

- ▶ Provides automation capability
- ▶ Simplifies disaster recovery operations
- ▶ Improves protection of critical business data guaranteeing data consistency and ensuring restart capabilities
- ▶ Reduces risk of downtime
- ▶ Enables regular, low-risk testing, and avoids human errors introduced into the process
- ▶ Simplifies the configuration:
 - Simple configuration, with the same look and feel for different disk systems
 - Configuration can be verified
- ▶ Simplifies the utilization:
 - Simple GUI to fix problems
 - No predefined Copy Services tasks necessary

Figure 3-13 provides an overview of a TPC for Replication architecture for a two site configuration.

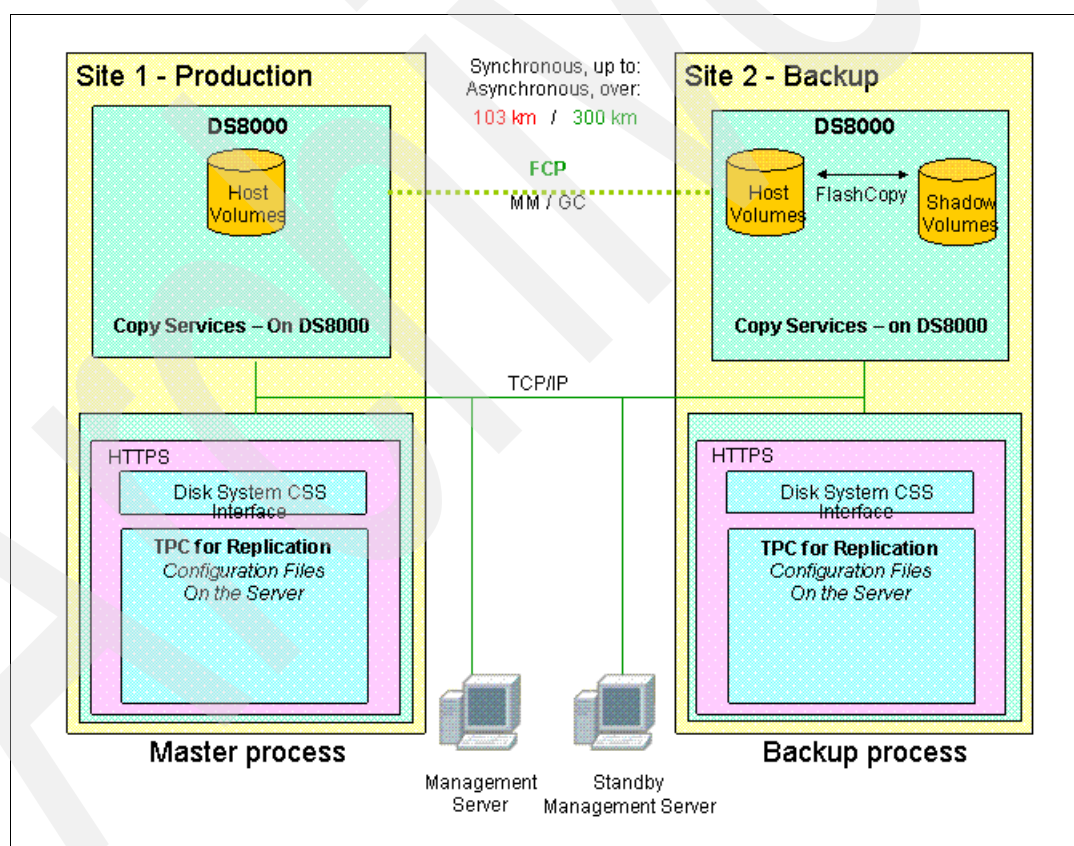


Figure 3-13 TPC for Replication V3 Architecture: 2 site configuration

Note: Although the IBM TPC for Replication architecture has the capability to manage a three site DR solution; at the time of writing, TPC for Replication does not support a three site DR solution (that is, a Metro/Global Mirror solution).

3.4 IBM System Storage SAN Volume Controller (SVC)

Rising storage requirements are posing an unprecedented challenge of managing disparate storage systems. IBM System Storage SAN Volume Controller (SVC) brings diverse storage devices together in a virtual pool to make all the storage appear as one logical device to centrally manage and to allocate capacity as required. It also provides a single solution to help achieve the most effective on demand use of key storage resources.

The SVC addresses the increasing costs and complexity in data storage management. It addresses this increased complexity by shifting storage management intelligence from individual SAN controllers into the network and by using virtualization.

There is a scalable hardware and software solution that facilitates aggregation of storage from different disk subsystems. It provides storage virtualization and thus a consistent view of storage across a (SAN).

The IBM SVC provides a resiliency level of Tier 6, when coupled with either Metro Mirror or Global Mirror.

SVC's storage virtualization:

- ▶ Consolidates disparate storage controllers into a single view
- ▶ Improves application availability by enabling data migration between disparate disk storage devices nondisruptively
- ▶ Improves disaster recovery and business continuity
- ▶ Reduces both the complexity and costs of managing SAN-based storage
- ▶ Increases business application responsiveness
- ▶ Maximize storage utilization
- ▶ Does dynamic resource allocation
- ▶ Simplifies management and improves administrator productivity
- ▶ Reduces storage outages
- ▶ Supports a wide range of servers and storage systems

For the most up-to-date list of supported operating systems, hosts, HBAs, SAN switches, and storage systems, refer to the following Web site:

<http://www.ibm.com/servers/storage/software/virtualization/svc/interop.html>

The SAN Volume Controller falls under Tier of 4 for FlashCopy and Tier 6 for Metro Mirror and Global Mirror as shown in Figure 3-14.

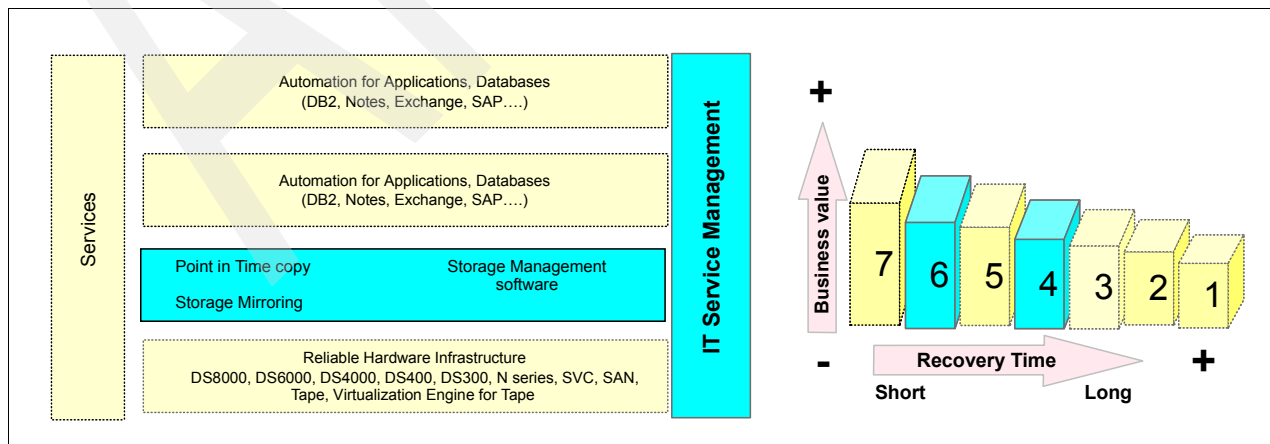


Figure 3-14 Tier level graph

Figure 3-15 summarizes how the SVC works — it acts as a layer between the application hosts and the physical storage, so that the hosts are insulated from change in the underlying storage layer.

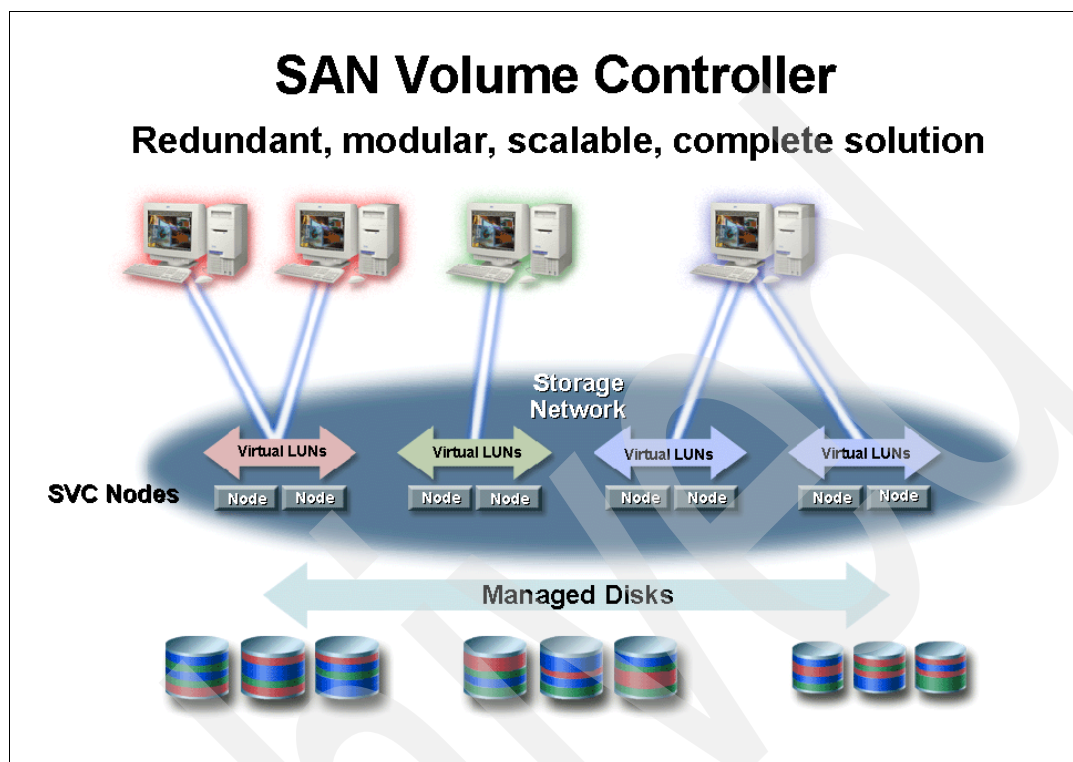


Figure 3-15 SAN Volume Controller

As shown in Figure 3-15, all the disk systems can be managed within pairs of nodes under SAN Volume Controller. Each host has virtual LUNs or vDisks that the SAN has zoned in such a way that the host or application servers cannot see the back-end storage. When a host or an application server performs I/O to a vDisk assigned to it by the SAN Volume Controller, it can access that vDisk via either of the two nodes in the I/O group.

Note: For further information about the IBM SVC, see 11.2, “IBM System Storage SAN Volume Controller” on page 365.

3.4.1 SAN Volume Controller: FlashCopy services

This section describes the fundamental principles of SAN Volume Controller FlashCopy which falls under BC Tier 4. FlashCopy makes a copy of source virtual disks to a set of target virtual disks. After the copy operations complete, the target virtual disks have the contents of the source virtual disks as they existed at a single point-in-time, known as a *T0 copy*. This method of copy, as compared to conventional disk to disk copy, might help to reduce the recovery time, backup time, and application impact.

FlashCopy must be part of a Consistency Group that addresses the issue of using applications that have related data which spans multiple virtual disks. Consistency Groups contain a number of FlashCopy mappings, which are necessary for FlashCopy to perform. All the FlashCopy mappings in the Consistency Groups are started at the same time, resulting in a point-in-time copy which is consistent across all FlashCopy mappings that are contained in the Consistency Group.

Important: For FlashCopy, source virtual disks and target virtual disks must be the same size. The minimum volume granularity for SVC FlashCopy is a complete virtual disk. Source and target virtual disks must both be managed by the same SV C cluster, but can be in different I/O groups within that cluster. It is not possible for a virtual disk to simultaneously be the source for one FlashCopy mapping and the target for another.

Consistency Groups address the issue where the objective is to preserve data consistency across multiple vDisks, because the applications have related data which spans multiple vDisks. A requirement for preserving the integrity of data being written is to ensure that dependent writes are executed in the application's intended sequence (Figure 3-16).

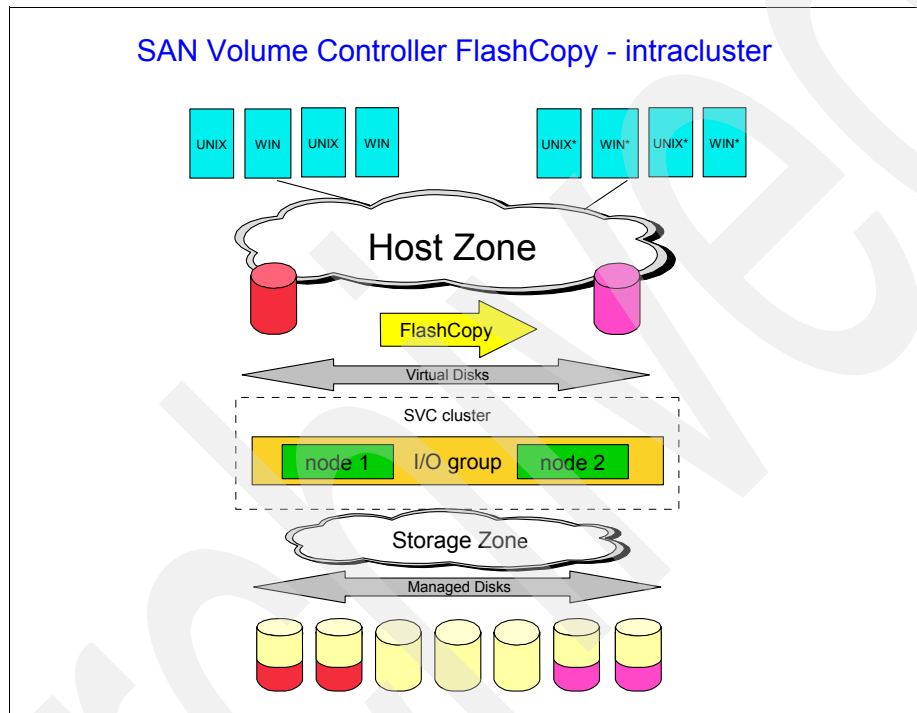


Figure 3-16 SAN Volume Controller FlashCopy - intracluster

FlashCopy requests a copy from the source virtual disk to the target virtual disk, using either the Web GUI, CLI or scripts. Then it creates a FlashCopy relationship between the source and target virtual disks. It establishes an algorithm in which it:

- ▶ Flushes write data in cache for a source virtual disk or disks
- ▶ Places cache into write-through mode on the source
- ▶ Discards file system cache on the target
- ▶ Establishes a sync point on all virtual disks
- ▶ Enables cache on both source and target

The practical uses for FlashCopy are:

- ▶ Backup
- ▶ Application testing
- ▶ Data backup with minimal impact on production
- ▶ Moving and migrating data
- ▶ Moving workload

UNIX with SVC: FlashCopy

In this section we describe the fundamentals for individual UNIX systems when using SVC FlashCopy. We explain how to bring FlashCopy target volumes online to the same host as well as to a second host.

The SVC FlashCopy functionality copies the entire contents of a source volume to a target volume. FlashCopy is a point-in-time copy and, after successful completion of the FlashCopy, the volumes under the backup server can then access the copied vDisks.

As shown in Figure 3-17, to the host, the storage device is a vDisk containing data. FlashCopy copies the production vDisk to another vDisk within the same SVC.

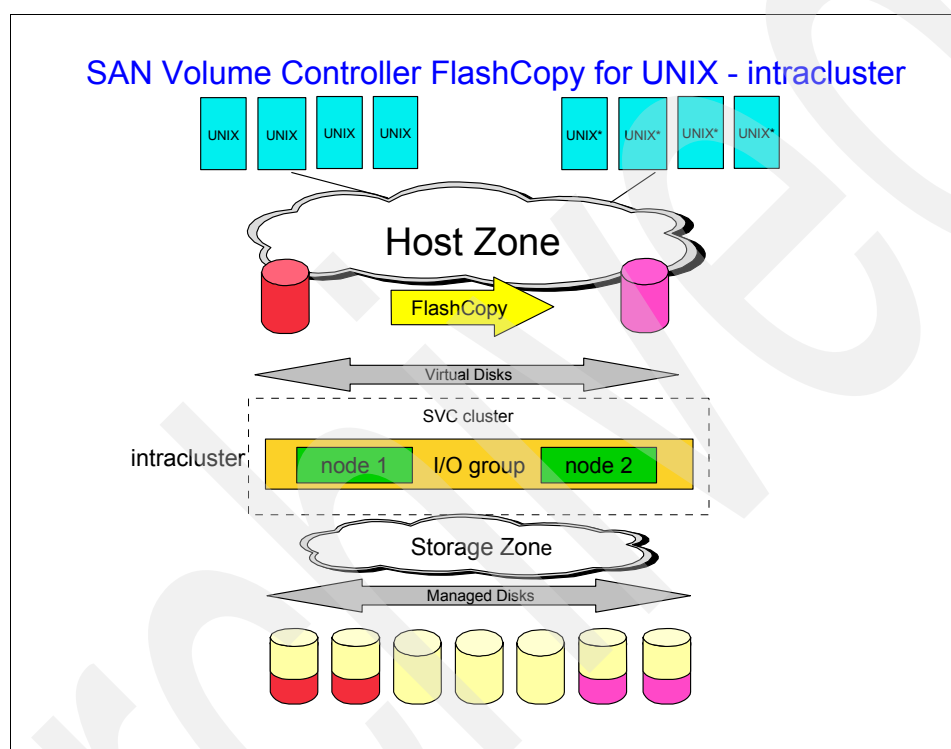


Figure 3-17 SAN Volume Controller FlashCopy for UNIX environment - Intracluster

Figure 3-18 shows that if the production UNIX volume fails, the target volume can be switched over to resume production responsibilities using FlashCopy. FlashCopy is a point-in-time copy which means that prior to the outage on the production UNIX system, FlashCopy is initiated and the secondary volume contains the latest point-in-time data copy of the production data. Assuming that there is no subsequent FlashCopy performed prior to the outage on the production UNIX system, the backup system has access the latest FlashCopied data that is available for recovery.

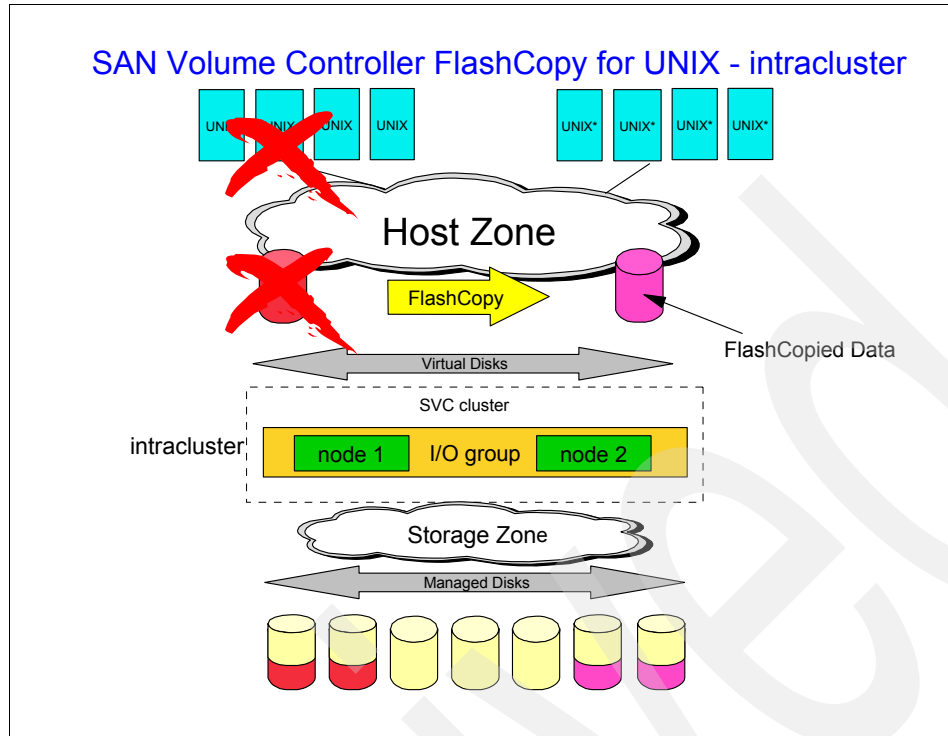


Figure 3-18 SVC FlashCopy for UNIX environment switch over - Intracuster

Recovery point objective and data currency are very dependent on how frequently FlashCopy is performed. This gives you a certain degree of protection of data but it does not give you a geographical recovery. In the recovery example in Figure 3-18 setup, it does not protect the entire SVC cluster from an outage. You can consider Metro Mirror to provide a higher resiliency for your data.

Windows with SVC FlashCopy

Microsoft Windows 2003 includes Volume Shadow Copy Service (VSS), which provides a new storage infrastructure for users of Microsoft Windows. These services allow the storage administrator to manage complex storage configurations more effectively, thus helping to realize the business goal of highly available data.

SVC allows the storage administrator to manage SVC's FlashCopy using an Application Programming Interface (API). To enable support for VSS, you must install IBM TotalStorage support for Microsoft Volume Shadow Copy Service (IBM VSS Provider). Refer to the *IBM System Storage SAN Volume Controller Configuration Guide*, SC26-7902.

The IBM TotalStorage Hardware Provider interfaces to the Microsoft VSS and to the CIM Agent on the master console, thereby seamlessly integrating VSS FlashCopy on the SVC. When a backup application on the Windows host initiates a snapshot command, VSS notifies the IBM TotalStorage Hardware Provider that a copy is required. VSS prepares volumes for a snapshot and quiesces applications and flushes file system buffers to prepare for the copy. The SVC makes the shadow copy with the FlashCopy Copy Service and the VSS service notifies the writing applications that I/O operations can resume after the backup was successful.

For more information about Microsoft Volume Shadow Copy Service refer to:

<http://www.microsoft.com>

Advantages of SVC FlashCopy

SVC FlashCopy gives you a certain degree of protection against outages for disaster recovery. FlashCopy can assist in offloading the backup window to the secondary or backup site or for application testing purposes. Data mining can be performed on the secondary FlashCopy, from which you can now extract data without affecting the production application.

Limitations of SVC FlashCopy

There are limits to how SVC FlashCopy can protect a production site from outages. It does not protect the production site from any form of wide scale disasters or outages within the data center, assuming that the entire SVC cluster is within the same data center.

3.4.2 SVC remote mirroring

The SVC supports two forms of remote mirroring, synchronous remote copy (implemented as Metro Mirror), and asynchronous remote copy (implemented as Global Mirror).

Synchronous remote copy

SVC Metro Mirror is a fully synchronous remote copy technique which ensures that updates are committed at both primary and secondary VDisks before the application is given completion to an update.

Figure 3-19 illustrates how a write operation to the master VDisk is mirrored to the cache for the auxiliary VDisk before an acknowledge of the write is sent back to the host issuing the write. This ensures that the secondary is real-time synchronized, that is, fully up-to-date, in case it is required in a failover situation.

However, this also means that the application is fully exposed to the latency and bandwidth limitations of the communication link to the secondary site. This might lead to unacceptable application performance, particularly when placed under peak load. This is the reason for the distance limitations when applying Metro Mirror.

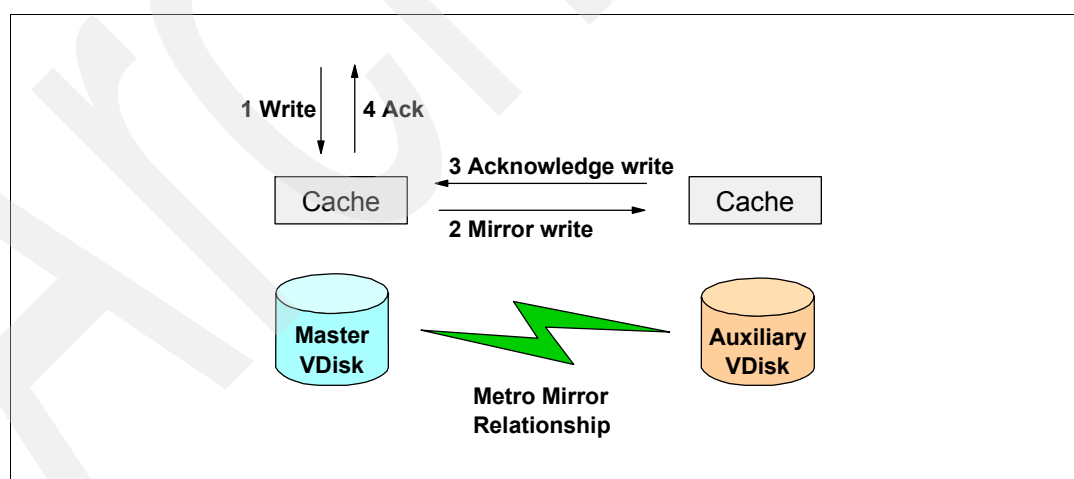


Figure 3-19 Write on VDisk in Metro Mirror relationship

Asynchronous remote copy

In an asynchronous remote copy, the application is given completion to an update when it is sent to the secondary site, but the update is not necessarily committed at the secondary site at that time. This provides the capability of performing remote copy over distances exceeding the limitations of synchronous remote copy.

Figure 3-20 illustrates that a write operation to the master VDisk is acknowledged back to the host issuing the write before it is mirrored to the cache for the auxiliary VDisk. In a failover situation, where the secondary site requirements to become the primary source of your data, some updates might be missing at the secondary site. The application must have some external mechanism for recovering the missing the updates and reapplying them, for example, transaction log replay.

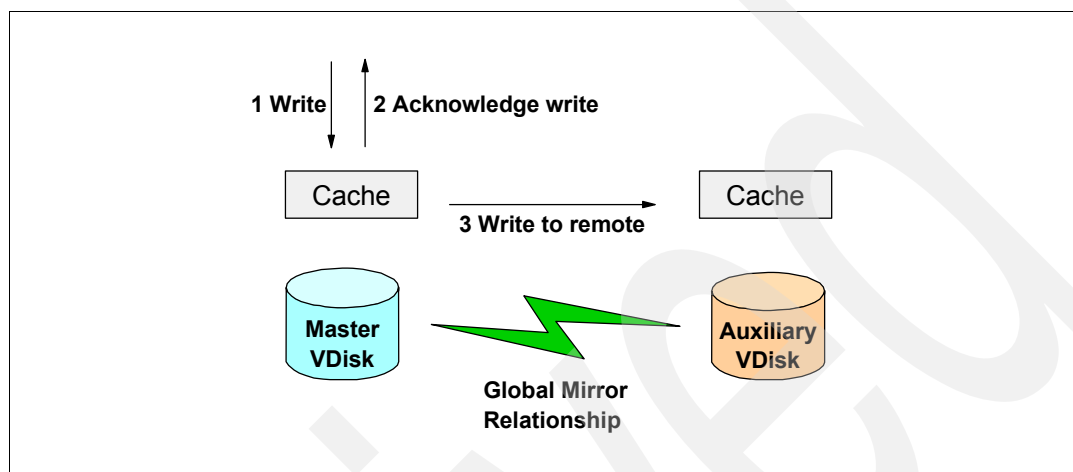


Figure 3-20 Write on VDisk in Global Mirror relationship

Next we describe each of these capabilities in further detail.

3.4.3 SVC Metro Mirror

In this section we describe the fundamental principles of SVC Metro Mirror (formerly PPRC), which falls under Tier 6. The general application of Metro Mirror seeks to maintain two copies of a data set often separated by some distance. SVC assumes that the Fibre Channel (FC) fabric to which it is attached contains hardware, which is capable of achieving the long distance requirement for the application. For more detail on extended distance replication, see *IBM System Storage SAN Volume Controller*, SG24-6423.

Important: When considering long distance communication connecting other SAN fabric components, you can consider a channel extender. These can involve protocol conversion to asynchronous transfer mode (ATM) or Internet Protocol (IP). Maximum distance between two nodes within an I/O group is 300 m (shortwave 2 Gbps) and 10 km (longwave). There are other more sophisticated components such as channel extenders or special SFP modules that can be used to extend the distance to thousands of kilometers. Performance does *degrade* as distance increases.

Metro Mirror enables two SVCs to connect to each other and establishes communications as in the local SVC cluster fabric.

One copy of the vDisk data is referred as the source copy (or *primary copy*), and this copy provides the reference for normal run time operation. Updates to this source copy is shadowed to the target copy. The target copy can be read but cannot be updated. If the source copy fails, the target copy can be enabled for I/O resumption. Typical usage of Metro Mirror involves two sites where the source provides service during normal operations and the target is only activated when a failure on source is detected.

Figure 3-21 and Figure 3-22 show two different SVC Metro Mirror setups.

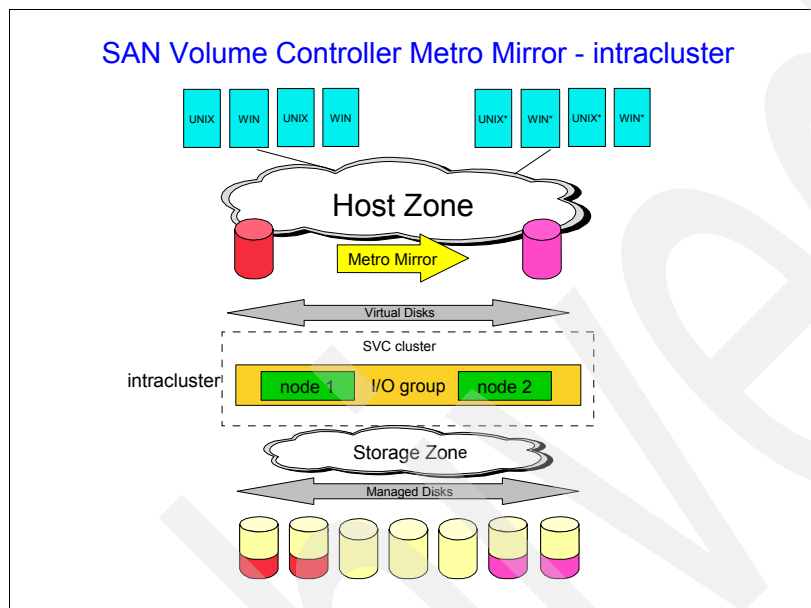


Figure 3-21 SVC Metro Mirror - Intracluster

SVC Metro Mirror in an intracluster environment can be a replacement for SVC FlashCopy. FlashCopy gives you a point-in-time data for recovery, while Metro Mirror gives you mirrored data during a failover process. Figure 3-21 shows the SVC Metro Mirror in an intracluster environment. Figure 3-22 shows the SVC Metro Mirror in an intercluster environment.

Note: If you plan to use intracluster Metro Mirror, make sure that both the master and auxiliary vDisk are in the same I/O group.

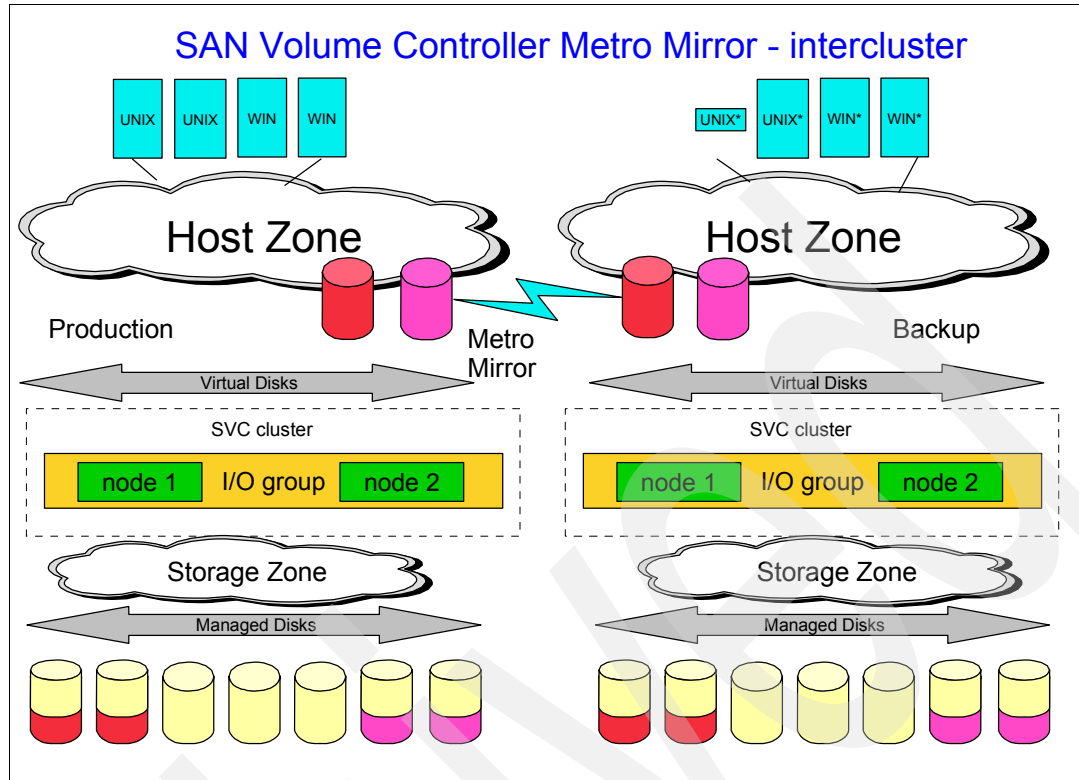


Figure 3-22 SVC Metro Mirror - Intercluster

SVC allows only read-only access to the target copy when it contains the consistent image. This is only intended to allow boot time and operating system discovery to complete without error, so that any hosts at the secondary site can be ready to start up the applications with minimum delay if required.

Enabling the secondary copy for active operations requires some SVC, operating system, and probably application specific work. This has to be performed as part of the entire failover process. The SVC software at the secondary site must be instructed to stop the relationship, which has the effect of making the secondary logical unit accessible for normal I/O access. The operating system might have to mount the file systems, or do similar work, which can typically only happen when the logical unit is accessible for writes. The application logs might have to be recovered.

The goal of Metro Mirror is to have a rapid (much faster, compared to recovering from a FlashCopy or tape backup copy) and seamless recovery. Most clients aim to automate this through failover management software. SVC provides SNMP traps that can be used to notify failover automation products.

Intracluster Metro Mirror can only be performed between vDisks in the same I/O group. We recommend that you use a dual-site solution using two SVCs, where the first site is considered the primary production site, and the second site is considered the failover site, which is activated when a failure of the first site is detected.

Intercluster Metro Mirror requires a pair of clusters, connected by a number of high bandwidth communication links to handle the writes from the Metro Mirror process. For Metro Mirror intercluster, the two SVCs must not share the same disk system, otherwise data loss might result. If the same mDisk becomes visible on two different SVCs, then this is an error that can cause data corruption.

A channel extender is a device for long distance communication connecting other SAN fabric components. Generally, these might involve protocol conversion to asynchronous transfer mode (ATM) or Internet Protocol (IP) or some other long distance communication protocol.

The SVC supports operation with Fibre Channel DWDM extenders and dark fiber. The supported distance depends on the SAN fabric vendor. You can also contact your IBM representative to discuss extended distance solutions.

UNIX with SVC Metro Mirror

SVC with Metro Mirror on UNIX gives you a Tier 6 level of resiliency. A variety of outages can happen on your production site, for example, an entire data center infrastructure failure. Figure 3-23 shows the logical setup for SVC using intercluster Metro Mirror.

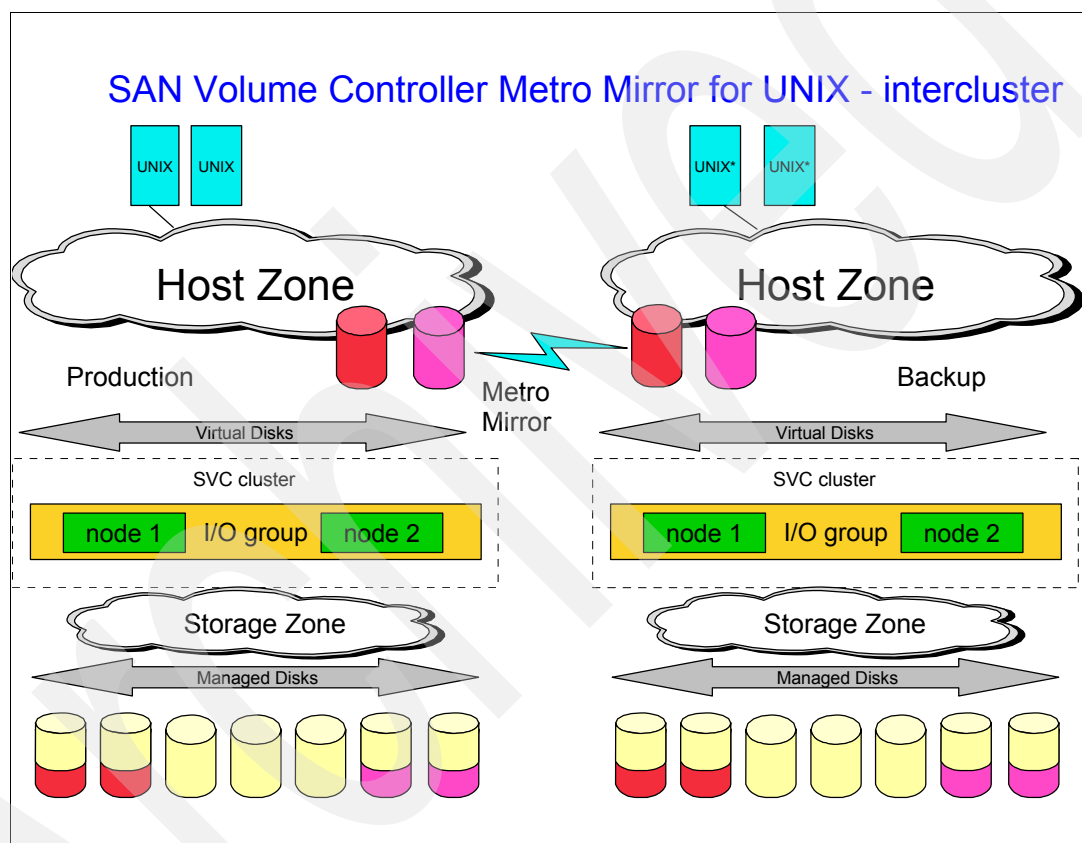


Figure 3-23 Intercluster Metro Mirror for UNIX environment

The concept of vDisk and zoning is the same as for FlashCopy. The only difference between FlashCopy and Metro Mirror is the distance between the SAN Volume Controller clusters. Refer to the IBM Redbook, *IBM System Storage SAN Volume Controller*, SG24-6423 for more information about distance limitations for Metro Mirror.

Figure 3-23 here and Figure 3-24, following, show SVC Metro Mirror in an intercluster and intracluster environment.

Enabling the secondary site copy for active operation requires using SVC for an entire failover process. The SVC software at the secondary site must be instructed to stop the relationship which has the affect of making the secondary site's logical units accessible for normal I/O access. The operating system might have to remount file systems, which can only happen when the logical unit is accessible for writes. The application might have some logs of work to recover.

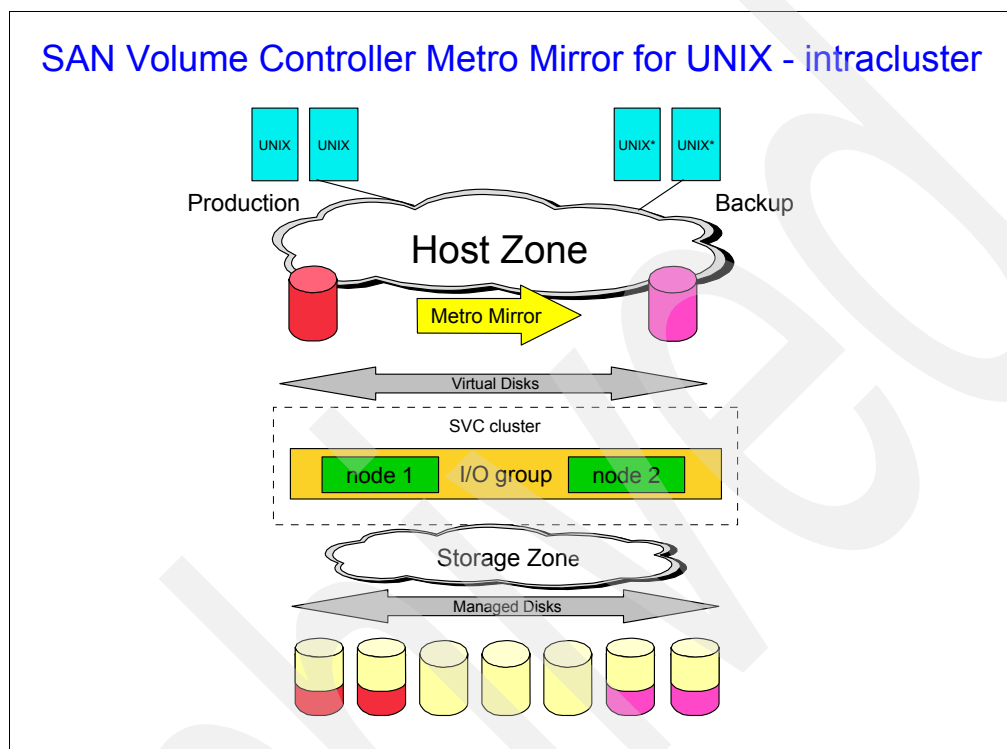


Figure 3-24 SAN Volume Controller Metro Mirror for UNIX environment - Intracluster

The goal for Metro Mirror is to make a more rapid recovery than one from a tape backup recovery, but it is not seamless. You can automate this through failover management software — SVC provides SNMP traps that can be used to notify failover automation products. One possible automation software product is TPC for Fabric; see 12.4, “IBM TotalStorage Productivity Center” on page 394.

Figure 3-25 shows the relationship between the production and the secondary sites for Metro Mirror. This assumes there is an outage occurring at the production site which causes the entire SVC setup to fail. Metro Mirror at this instance loses all communication for any possible update to the secondary SVC. You can immediately stop the relationship so that the secondary or secondary logical unit can take over the production responsibilities. You can use automation software to automate the switch over process or manual intervention from a system administrator.

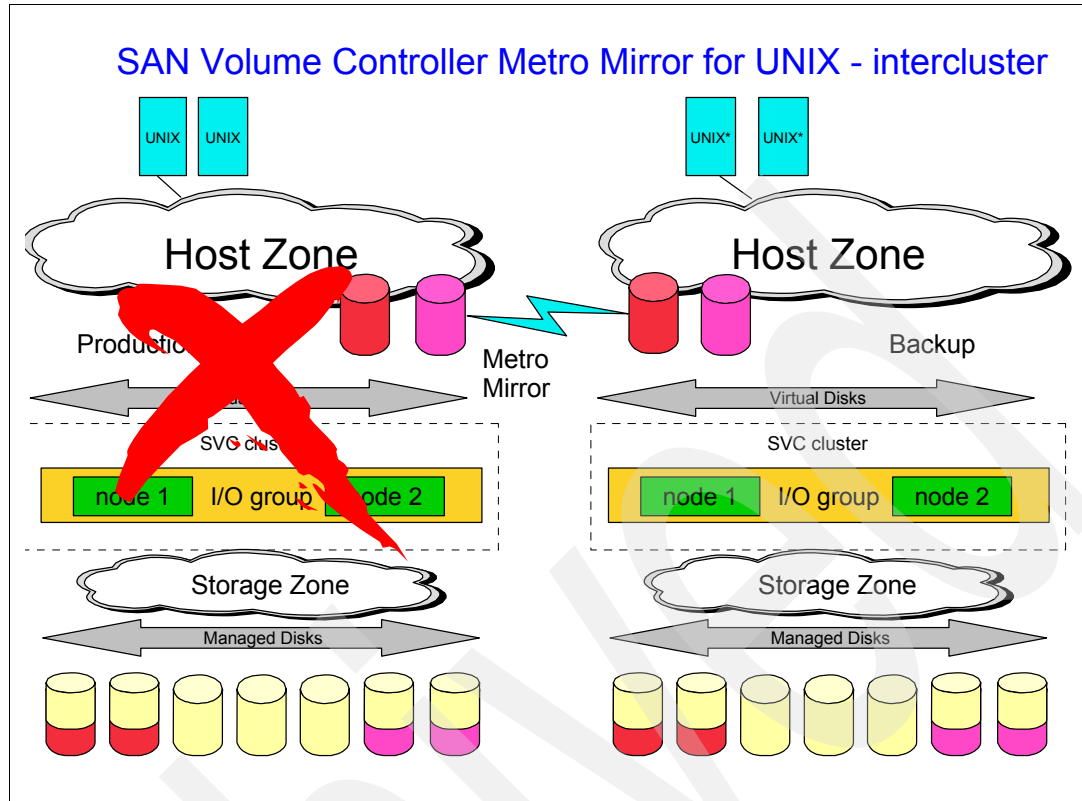


Figure 3-25 SVC Metro Mirror intercluster disaster scenario

Attention: There might be instances where communication is lost during the Metro Mirror process. In a normal situation, a copy can be consistent because it contains data which was frozen at some point-in-time. Write I/O might have continued to the production disks and not have been copied to the backup.

This state arises when it becomes impossible to keep data up-to-date and maintain transaction consistency. Metro Mirror tracks the changes that happen at the production site, but not the order of such changes, nor the details of such changes (write data). When communication is restored, it is impossible to make the backup synchronized without sending write data to the backup out-of-order, and therefore losing consistency. There are two approaches that can be used to circumvent this problem:

- ▶ Take a point-in-time copy of the consistent secondary before allowing the secondary to become inconsistent. In the event of a disaster before consistency is achieved, the point-in-time copy target provides a consistent, though out-of-date, image.
- ▶ Accept the loss of consistency, and loss of useful secondary, while making it synchronized.

Windows with SVC: Metro Mirror

As for UNIX, Metro Mirror can be used either intracluster or intercluster. During a disaster at the production site for intracluster, the vDisk fails to replicate across to the backup site. You can stop the relationship so that the secondary logical unit can take over the production responsibilities. For an intracluster Metro Mirror setup, there is a single point of failure that is the primary and secondary vDisk are located within the same I/O group. Any failure to the I/O groups poses a potential disaster.

For intercluster Metro Mirror setup, should the SVC I/O group with the production fail, you can stop the relationship, so that the secondary logical unit can take over production responsibilities. Intercluster Metro Mirror is recommended if a true hot backup site is retained for disaster recovery purposes. See Figure 3-21 on page 134 and Figure 3-22 on page 135, which show how a Windows platform can be connected to SAN Volume Controller with a UNIX system.

When a production Windows system fails, you can fail over to the secondary Windows system which has the mirrored copy of data. After you stop the relationship on the SVC Metro Mirror the secondary logical unit can take over the production responsibilities.

SVC Metro Mirror in an intercluster environment provides you a higher resiliency level of protection as compared to an intracluster environment. SVC Metro Mirror can protect you from an entire production SVC cluster outage or failure. You can stop the SVC Metro Mirror relationship and the secondary site can take over the production responsibilities.

Advantages of SVC Metro Mirror

SVC Metro Mirror provides a geographical recovery in the event of a wide scale outage or disaster at the production site. Metro Mirror takes over if it should encounter any failure in the production site.

Limitations of SAN Volume Controller Metro Mirror

With synchronous remote copy, the updates are committed at both the production and backup sites, thus introducing latency and bandwidth limitations on the communication link to the secondary site. This might pose a significant adverse effect on application performance if latency and bandwidth limitations are not addressed in the planning and designing stage.

3.4.4 Global Mirror

SVC Global Mirror (GM) copy service provides and maintains a consistent mirrored copy of a source VDisk to a target VDisk. Data is written from the source to the target asynchronously. This method was previously known as Asynchronous Peer-to-Peer Remote Copy.

Global Mirror works by defining a GM relationship between two VDisks of equal size and maintains the data consistency in an asynchronous manner. Therefore, when a host writes to a source VDisk, the data is copied from the source VDisk cache to the target VDisk cache. At the initiation of that data copy, confirmation of I/O completion is transmitted back to the host.

Note: The minimum firmware requirement for GM functionality is v4.1.1. Any cluster or partner cluster not running this minimum level does *not* have GM functionality available. Even if you have a GM relationship running on a downlevel partner cluster and you only wish to use intracluster GM, the functionality is not available to you.

The SVC provides both intracluster and intercluster Global Mirror:

Intracluster Global Mirror

Although GM is available for intracluster, it has no functional value for production use. Intracluster GM provides the same capability for less overhead. However, leaving this functionality in place simplifies testing and does allow client experimentation and testing (for example, to validate server failover on a single test cluster).

Intercluster Global Mirror

Intercluster GM operations require a pair of SVC clusters that are commonly separated by a number of moderately high bandwidth links. The two SVC clusters must each be defined in an SVC cluster partnership to establish a fully functional GM relationship.

Note: When a local and a remote fabric are connected together for GM purposes, the ISL hop count between a local node and a remote node should not exceed seven hops.

Supported methods for synchronizing

This section describes three methods that can be used to establish a relationship.

Full synchronization after Create

This is the *default method*. It is the simplest, in that it requires no administrative activity apart from issuing the necessary commands. However, in some environments, the bandwidth available makes this method unsuitable.

The sequence for a single relationship is:

- ▶ A **mkrcrelationship** is issued without specifying **-sync** flag.
- ▶ A **starttrrelationship** is issued without **-clean**.

Synchronized before Create

In this method, the administrator must ensure that the master and auxiliary virtual disks contain identical data before creating the relationship. There are two ways in which this might be done:

- ▶ Both disks are created with the **-fmt disk** feature so as to make all data zero.
- ▶ A complete tape image (or other method of moving data) is copied from one disk to the other.

In either technique, no write I/O must take place to either Master or Auxiliary before the relationship is established.

Then, the administrator must ensure that:

- ▶ A **mkrcrelationship** is issued with **-sync** flag.
- ▶ A **starttrrelationship** is issued without **-clean** flag.

If these steps are not performed correctly, the relationship is reported as being consistent, when it is not. This is likely to make any secondary disk useless. This method has an advantage over the full synchronization, in that it does not require all the data to be copied over a constrained link. However, if the data has to be copied, the master and auxiliary disks cannot be used until the copy is complete, which might be unacceptable.

Quick synchronization after Create

In this method, the administrator must still copy data from master to auxiliary. But it can be used without stopping the application at the master. The administrator must ensure that:

- ▶ A **mkrcrelationship** is issued with **-sync** flag.
- ▶ A **stoprcrelationship** is issued with **-access** flag.
- ▶ A tape image (or other method of transferring data) is used to copy the entire master disk to the auxiliary disk.

Once the copy is complete, the administrator must ensure that:

- ▶ A **startrcrelationship** is issued with **-clean** flag.

With this technique, only the data that has changed since the relationship was created, including all regions that were incorrect in the tape image, are copied from master and auxiliary. As with “Synchronized before Create” on page 140, the copy step must be performed correctly, or else the auxiliary is be useless, although the copy reports it as being synchronized.

The importance of write ordering

Many applications that use block storage have a requirement to survive failures such as loss of power, or a software crash, and not lose data that existed prior to the failure. Since many applications have to perform large numbers of update operations in parallel to that storage, maintaining write ordering is key to ensuring the correct operation of applications following a disruption.

An application that is performing a large set of updates was probably designed with the concept of dependent writes. These are writes where it is important to ensure that an earlier write has completed before a later write is started. Reversing the order of dependent writes can undermine the applications algorithms and can lead to problems such as detected, or undetected, data corruption.

Dependent writes that span multiple VDisks

The following scenario illustrates a simple example of a sequence of dependent writes, and in particular what can happen if they span multiple VDisks. Consider the following typical sequence of writes for a database update transaction:

1. A write is executed to update the database log, indicating that a database update is to be performed.
2. A second write is executed to update the database.
3. A third write is executed to update the database log, indicating that the database update has completed successfully.

Figure 3-26 illustrates the write sequence.

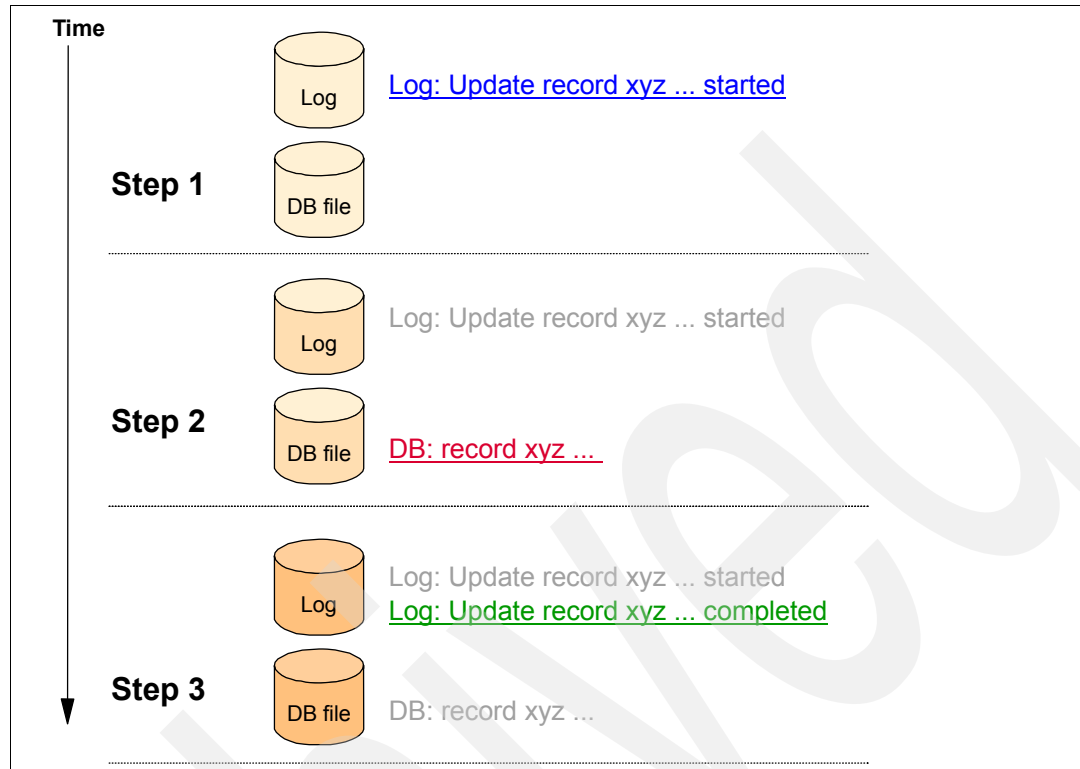


Figure 3-26 Dependent writes for a database

The database ensures the correct ordering of these writes by waiting for each step to complete before starting the next.

Note: All databases have logs associated with them. These logs keep records of database changes. If a database has to be restored to a point beyond the last full, offline backup, logs are required to roll the data forward to the point of failure.

But imagine if the database log and the database itself are on different VDisks and a Global Mirror relationship is stopped during this update. In this case you have to exclude the possibility that the GM relationship for the VDisk with the database file is stopped slightly before the VDisk containing the database log. If this were the case, then it could be possible that the secondary VDisks see writes (1) and (3) but not (2).

Then, if the database was restarted using the backup made from the secondary disks, the database log would indicate that the transaction had completed successfully, when it is not the case. In this scenario, the integrity of the database is in question.

To overcome the issue of dependent writes across VDisks, and to ensure a consistent data set, the SVC supports the concept of consistency groups for GM relationships. A GM consistency group can contain an arbitrary number of relationships up to the maximum number of GM relationships supported by the SVC cluster.

GM commands are then issued to the GM consistency group, and thereby simultaneously for all GM relationships defined in the consistency group. For example, when issuing a GM **start** command to the consistency group, all of the GM relationships in the consistency group are started at the same time.

Using Global Mirror

To use Global Mirror, a relationship must be defined between two VDisks.

When creating the GM relationship, one VDisk is defined as the master, and the other as the auxiliary. The relationship between the two copies is asymmetric. When the GM relationship is created, the master VDisk is initially considered the primary copy (which is often referred to as the source), and the auxiliary VDisk is considered the secondary copy (often referred to as the target).

The master VDisk is the production VDisk, and updates to this copy are real time mirrored to the auxiliary VDisk. The contents of the auxiliary VDisk that existed when the relationship was created are destroyed.

Note: The copy direction for a GM relationship can be switched so the auxiliary VDisk becomes the production and the master VDisk becomes the secondary.

While the GM relationship is active, the auxiliary copy (VDisk) is not accessible for host application write I/O at any time. The SVC allows read-only access to the auxiliary VDisk when it contains a “consistent” image. This is only intended to allow boot time operating system discovery to complete without error, so that any hosts at the secondary site can be ready to start up the applications with minimum delay if required.

For instance, many operating systems have to read Logical Block Address (LBA) 0 to configure a logical unit. Although read access is allowed at the secondary in practice, the data on the secondary volumes cannot be read by a host. The reason for this is that most operating systems write a “dirty bit” to the file system when it is mounted. Because this write operation is not allowed on the secondary volume, the volume cannot be mounted.

This access is only provided where consistency can be guaranteed. However, there is no way in which coherency can be maintained between reads performed at the secondary and later write I/Os performed at the primary.

To enable access to the secondary VDisk for host operations, the GM relationship must be stopped, specifying the **-access** parameter.

While access to the auxiliary VDisk for host operations is enabled, the host must be instructed to mount the VDisk and related tasks before the application can be started, or instructed to perform a recovery process.

The GM requirement to enable the secondary copy for access differentiates it from, for example, third party mirroring software on the host, which aims to emulate a single, reliable disk regardless of which system is accessing it. GM retains the property that there are two volumes in existence, but suppresses one while the copy is being maintained.

Using a secondary copy demands a conscious policy decision by the administrator that a failover is required, and the tasks to be performed on the host involved in establishing operation on the secondary copy are substantial. The goal is to make this rapid (much faster when compared to recovering from a backup copy) but not seamless.

The failover process can be automated through failover management software. The SVC provides SNMP traps and programming (or scripting) using the CLI to enable this automation.

SVC Global Mirror features

SVC Global Mirror supports the following features:

- ▶ Asynchronous remote copy of VDisks dispersed over metropolitan scale distances is supported.
- ▶ SVC implements the GM relationship between VDisk pairs, with each VDisk in the pair being managed by an SVC cluster.
- ▶ SVC supports intracluster GM, where both VDisks belong to the same cluster (and IO group). However, as stated earlier, this functionality is better suited to Metro Mirror.
- ▶ SVC supports intercluster GM, where each VDisk belongs to their separate SVC cluster. A given SVC cluster can be configured for partnership with another cluster. A given SVC cluster can only communicate with one other cluster. All intercluster GM takes place between the two SVC clusters in the configured partnership.
- ▶ Intercluster and intracluster GM can be used concurrently within a cluster for different relationships.
- ▶ SVC does not require a control network or fabric to be installed to manage Global Mirror. For intercluster GM the SVC maintains a control link between the two clusters. This control link is used to control state and co-ordinate updates at either end. The control link is implemented on top of the same FC fabric connection as the SVC uses for GM I/O.
- ▶ SVC implements a configuration model which maintains the GM configuration and state through major events such as failover, recovery, and resynchronization to minimize user configuration action through these events.
- ▶ SVC maintains and polices a strong concept of consistency and makes this available to guide configuration activity.
- ▶ SVC implements flexible resynchronization support enabling it to re-synchronize VDisk pairs which have suffered write I/O to both disks and to resynchronize only those regions which are known to have changed.

How Global Mirror works

There are several steps in the Global Mirror process:

1. An SVC cluster partnership is created between two SVC clusters (for intercluster GM).
2. A GM relationship is created between two VDisks of the same size.
3. To manage multiple GM relationships as one entity, the relationships can be made part of a GM consistency group. This is to ensure data consistency across multiple GM relationships, or simply for ease of management.
4. The GM relationship is started, and when the background copy has completed the relationship is consistent and synchronized.
5. Once synchronized, the auxiliary VDisk holds a copy of the production data at the primary which can be used for disaster recovery.
6. To access the auxiliary VDisk, the GM relationship must be stopped with the access option enabled before write I/O is submitted to the secondary.
7. The remote host server is mapped to the auxiliary VDisk and the disk is available for I/O.

Global Mirror relationships

Global Mirror relationships are similar to FlashCopy mappings. They can be stand-alone or combined in consistency groups. The **start** and **stop** commands can be issued either against the stand-alone relationship, or the consistency group.

Figure 3-27 illustrates the Global Mirror relationship.

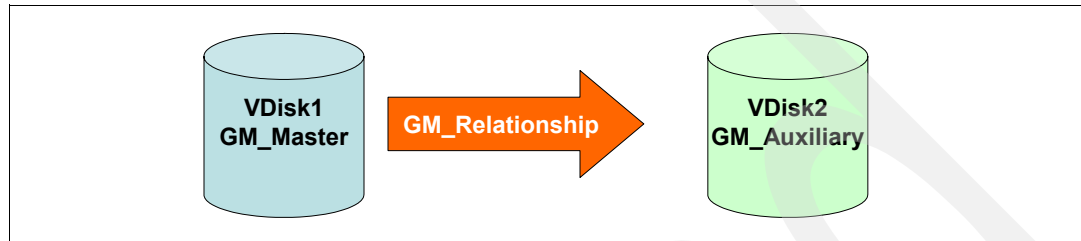


Figure 3-27 GM relationship

A GM relationship is composed of two VDisks (known as master and auxiliary) of the same size. The two VDisks can be in the same I/O group, within the same SVC cluster (intracluster Global Mirror), or can be on separate SVC clusters that are defined as SVC partners (intercluster Global Mirror).

Note: Be aware that:

- ▶ A VDisk can only be part of one Global Mirror relationship at a time.
- ▶ A VDisk that is a FlashCopy target cannot be part of a Global Mirror relationship.

Global Mirror consistency groups

Certain uses of GM require the manipulation of more than one relationship. GM consistency groups provides the ability to group relationships, so that they are manipulated in unison.

Consistency groups address the issue where the objective is to preserve data consistency across multiple Global Mirrored VDisks because the applications have related data which spans multiple VDisks. A requirement for preserving the integrity of data being written is to ensure that “dependent writes” are executed in the application's intended sequence.

GM commands can be issued to a GM consistency group, which affects all GM relationships in the consistency group, or to a single GM relationship if not part of a GM consistency group.

Figure 3-28 shows the concept of GM consistency groups. Since the GM_Relationship 1 and GM_Relationship 2 are part of the consistency group, they can be handled as one entity, while the stand-alone GM_Relationship 3 is handled separately.

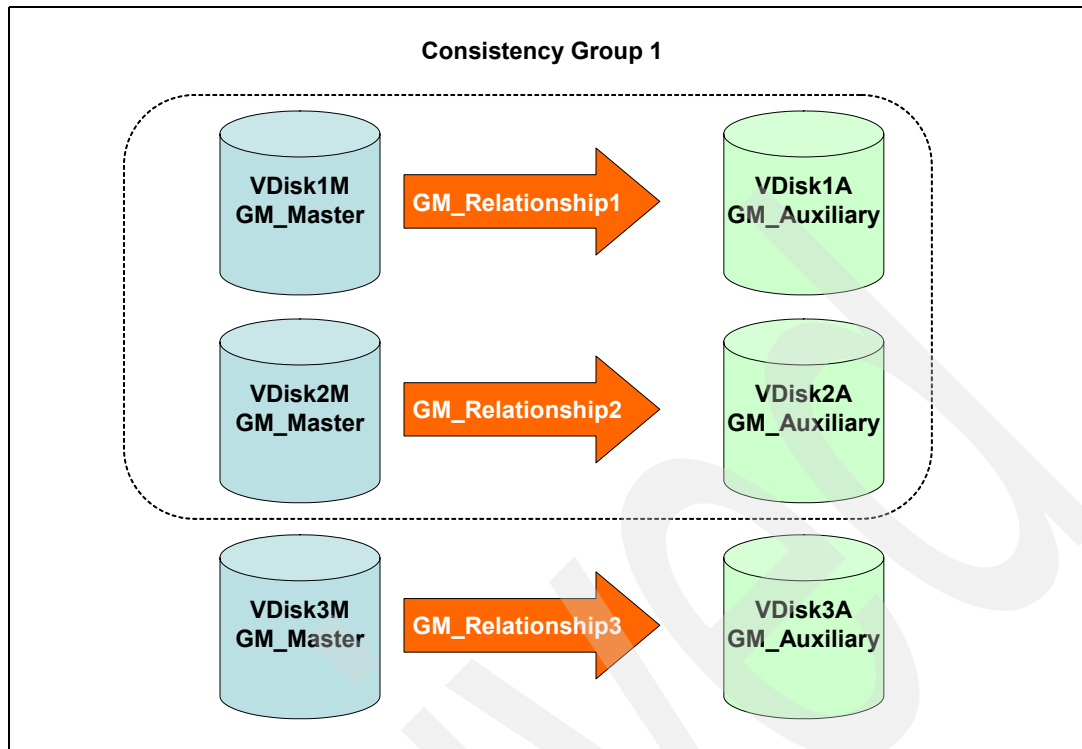


Figure 3-28 Global Mirror consistency group

The following rules for consistency groups apply:

- ▶ Global Mirror relationships can be part of a consistency group, or be stand-alone and therefore handled as single instances.
- ▶ A consistency group can contain zero or more relationships. An empty consistency group, with zero relationships in it, has little purpose until it is assigned its first relationship, except that it has a name.
- ▶ All the relationships in a consistency group must have matching master and auxiliary SVC clusters.

Although it is possible that consistency groups can be used to manipulate sets of relationships that do not have to satisfy these strict rules, that manipulation can lead to some undesired side effects. The rules behind consistency mean that certain configuration commands are prohibited where this would not be the case if the relationship was not part of a consistency group.

For example, consider the case of two applications that are completely independent, yet they are placed into a single consistency group. In the event of an error there is a loss of synchronization, and a background copy process is required to recover synchronization. While this process is in progress, GM rejects attempts to enable access to the secondary VDisks of either application.

If one application finishes its background copy much more quickly than the other, GM still refuses to grant access to its secondary, even though this is safe in this case, because the GM policy is to refuse access to the entire consistency group if any part of it is inconsistent.

Stand-alone relationships and consistency groups share a common configuration and state model. All the relationships in a non-empty consistency group have the same state as the consistency group.

Global Mirror states and events

In this section we explain the different states of a GM relationship, and the series of events that modify these states. In Figure 3-29, the GM relationship state diagram shows an overview of the states that apply to a GM relationship in the connected state.

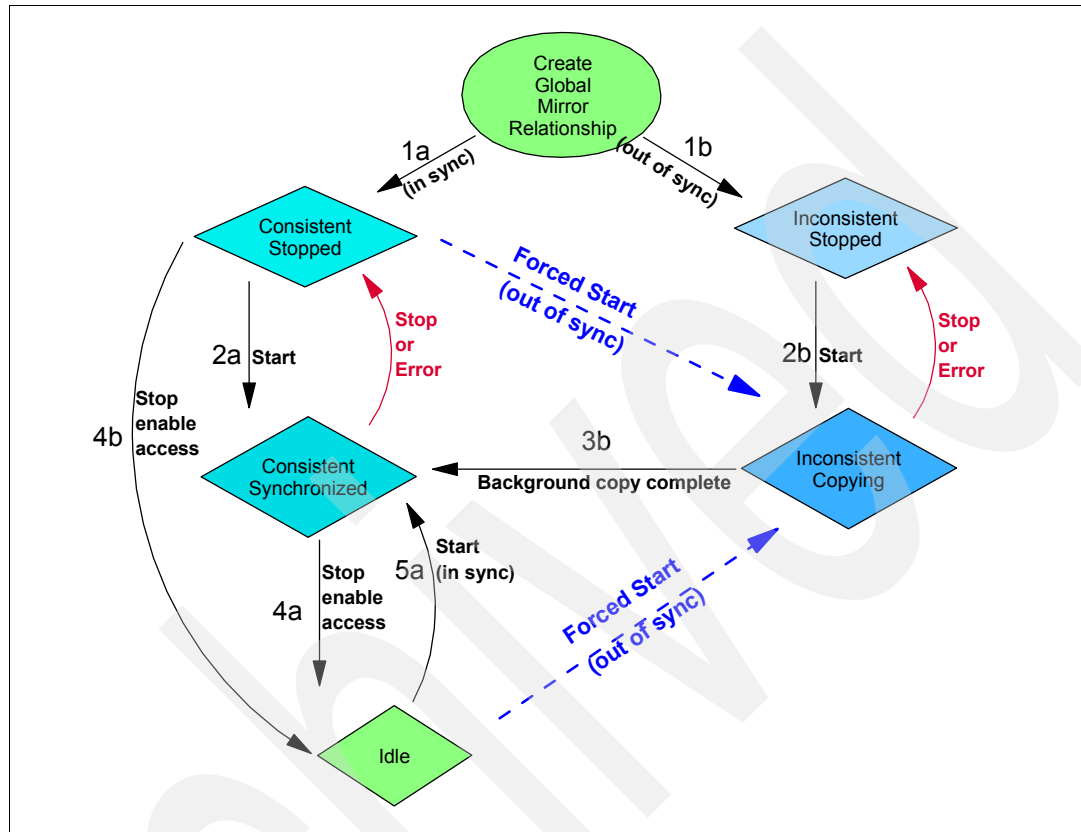


Figure 3-29 Global Mirror mapping state diagram

When creating the GM relationship, you can specify if the auxiliary VDisk is already in sync with the master VDisk, and the background copy process is then skipped. This is especially useful when creating GM relationships for VDisks that have been created with the format option:

1. Step 1 is done as follows:
 - a. The GM relationship is created with the **-sync** option and the GM relationship enters the *Consistent stopped* state.
 - b. The GM relationship is created without specifying that the master and auxiliary VDisks are in sync, and the GM relationship enters the *Inconsistent stopped* state.
2. Step 2 is done as follows:
 - a. When starting a GM relationship in the *Consistent stopped* state, it enters the *Consistent synchronized* state. This implies that no updates (write I/O) have been performed on the primary VDisk while in the *Consistent stopped* state, otherwise the **-force** option must be specified, and the GM relationship then enters the *Inconsistent copying* state, while background copy is started.
 - b. When starting a GM relationship in the *Inconsistent stopped* state, it enters the *Inconsistent copying* state, while background copy is started.

3. Step 3 is done as follows:
 - a. When the background copy completes, the GM relationship transitions from the *Inconsistent copying* state to the *Consistent synchronized* state.
4. Step 4 is done as follows:
 - a. When stopping a GM relationship in the *Consistent synchronized* state, specifying the `-access` option which enables write I/O on the secondary VDisk, the GM relationship enters the *Idling* state.
 - b. To enable write I/O on the secondary VDisk, when the GM relationship is in the *Consistent stopped* state, issue the command `svctask stoprcrelationship` specifying the `-access` option, and the GM relationship enters the *Idling* state.
5. Step 5 is done as follows:
 - a. When starting a GM relationship which is in the *Idling* state, it is required to specify the `-primary` argument to set the copy direction. Given that no write I/O has been performed (to either master or auxiliary VDisk) while in the *Idling* state, the GM relationship enters the *Consistent synchronized* state.
 - b. In case write I/O has been performed to either the master or the auxiliary VDisk, then the `-force` option must be specified, and the GM relationship then enters the *Inconsistent copying* state, while background copy is started.

Stop or Error: When a GM relationship is stopped (either intentionally or due to an error), a state transition is applied:

- ▶ For example, this means that GM relationships in the *Consistent synchronized* state enter the *Consistent stopped* state and GM relationships in the *Inconsistent copying* state enter the *Inconsistent stopped* state.
- ▶ In case the connection is broken between the SVC clusters in a partnership, then all (intercluster) GM relationships enter a disconnected state. For further information, refer to the following topic, “Connected versus disconnected”.

Note: Stand-alone relationships and consistency groups share a common configuration and state model. This means that all the GM relationships in a non-empty consistency group have the same state as the consistency group.

State overview

The SVC defined concepts of state are key to understanding the configuration concepts and are therefore explained in more detail below.

Connected versus disconnected

This distinction can arise when a Global Mirror relationship is created with the two virtual disks in different clusters.

Under certain error scenarios, communications between the two clusters might be lost. For instance, power might fail causing one complete cluster to disappear. Alternatively, the fabric connection between the two clusters might fail, leaving the two clusters running but unable to communicate with each other.

When the two clusters can communicate, the clusters and the relationships spanning them are described as connected. When they cannot communicate, the clusters and the relationships spanning them are described as disconnected.

In this scenario, each cluster is left with half the relationship and has only a portion of the information that was available to it before. Some limited configuration activity is possible, and is a subset of what was possible before.

The disconnected relationships are portrayed as having a changed state. The new states describe what is known about the relationship, and what configuration commands are permitted.

When the clusters can communicate again, the relationships become connected once again. Global Mirror automatically reconciles the two state fragments, taking into account any configuration or other event that took place while the relationship was disconnected. As a result, the relationship can either return to the state it was in when it became disconnected or it can enter a different connected state.

Relationships that are configured between virtual disks in the same SVC cluster (intracluster) are never described as being in a disconnected state.

Consistent versus inconsistent

Relationships that contain VDisks operating as secondaries can be described as being consistent or inconsistent. Consistency groups that contain relationships can also be described as being consistent or inconsistent. The consistent or inconsistent property describes the relationship of the data on the secondary to that on the primary virtual disk. It can be considered a property of the secondary VDisk itself.

A secondary is described as consistent if it contains data that could have been read by a host system from the primary if power had failed at some imaginary point in time while I/O was in progress and power was later restored. This imaginary point in time is defined as the recovery point. The requirements for consistency are expressed with respect to activity at the primary up to the recovery point:

- ▶ The secondary virtual disk contains the data from all writes to the primary for which the host had received good completion and that data had not been overwritten by a subsequent write (before the recovery point)
- ▶ For writes for which the host did not receive good completion (that is, it received bad completion or no completion at all) and the host subsequently performed a read from the primary of that data and that read returned good completion and no later write was sent (before the recovery point), the secondary contains the same data as that returned by the read from the primary.

From the point of view of an application, consistency means that a secondary virtual disk contains the same data as the primary virtual disk at the recovery point (the time at which the imaginary power failure occurred).

If an application is designed to cope with unexpected power failure this guarantee of consistency means that the application is able to use the secondary and begin operation just as though it had been restarted after the hypothetical power failure.

Again, the application is dependent on the key properties of consistency:

- ▶ Write ordering
- ▶ Read stability for correct operation at the secondary

If a relationship, or set of relationships, is inconsistent and an attempt is made to start an application using the data in the secondaries, a number of outcomes are possible:

- ▶ The application might decide that the data is corrupt and crash or exit with an error code.
- ▶ The application might fail to detect that the data is corrupt and return erroneous data.
- ▶ The application might work without a problem.

Because of the risk of data corruption, and in particular undetected data corruption, Global Mirror strongly enforces the concept of consistency and prohibits access to inconsistent data.

Consistency as a concept can be applied to a single relationship or a set of relationships in a consistency group. Write ordering is a concept that an application can maintain across a number of disks accessed through multiple systems and therefore consistency must operate across all those disks.

When deciding how to use consistency groups, the administrator must consider the scope of an application's data, taking into account all the interdependent systems which communicate and exchange information.

If two programs or systems communicate and store details as a result of the information exchanged, then either of the following actions might occur:

- ▶ All the data accessed by the group of systems must be placed into a single consistency group.
- ▶ The systems must be recovered independently (each within its own consistency group). Then, each system must perform recovery with the other applications to become consistent with them.

Consistent versus synchronized

A copy which is consistent and up-to-date is described as synchronized. In a synchronized relationship, the primary and secondary virtual disks are only different in regions where writes are outstanding from the host.

Consistency does not mean that the data is up-to-date. A copy can be consistent and yet contain data that was frozen at some point in time in the past. Write I/O might have continued to a primary and not have been copied to the secondary. This state arises when it becomes impossible to keep up-to-date and maintain consistency. An example is a loss of communication between clusters when writing to the secondary.

When communication is lost for an extended period of time, Global Mirror tracks the changes that happen at the primary, but not the order of such changes, nor the details of such changes (write data). When communication is restored, it is impossible to make the secondary synchronized without sending write data to the secondary out-of-order, and therefore losing consistency.

Two policies can be used to cope with this:

- ▶ Take a point-in-time copy of the consistent secondary before allowing the secondary to become inconsistent. In the event of a disaster before consistency is achieved again, the point-in-time copy target provides a consistent, though out-of-date, image.
- ▶ Accept the loss of consistency, and loss of useful secondary, while making it synchronized.

Global Mirror configuration limits

Table 3-2 lists the SVC Global Mirror configuration limits.

Table 3-2 SVC Global Mirror configuration limits

Parameter	Value
Number of Global Mirror consistency groups	256 per SVC cluster
Number of Global Mirror relationships	1024 per SVC cluster
Total VDisk size per I/O group	16TB is the per I/O group limit on the quantity of primary and secondary VDisk address space that can participate in Global Mirror relationships

3.4.5 Summary

SAN Volume Controller is designed to increase the flexibility of your storage infrastructure by enabling changes to the physical storage with minimal or no disruption to applications. SAN Volume Controller combines the capacity from multiple disk storage systems into a single storage pool, which can be managed from a central point. This is simpler to manage and helps increase utilization. It also allows you to apply advanced copy services across storage systems from many different vendors to help further simplify operations.

3.5 System i storage introduction

The System i uses the concept of single-level storage — where all disk storage is regarded as being in a single big bucket, and programmers do not have to consider where data was stored. Figure 3-30 shows the overall System i single level storage.

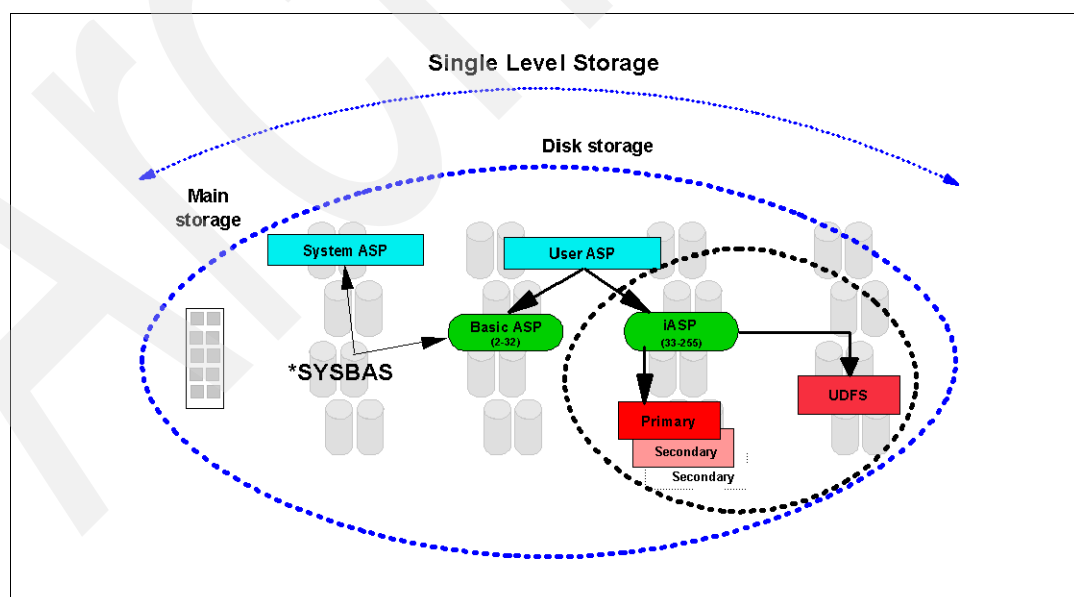


Figure 3-30 System i single level storage

Programs, files, and other structures are regarded simply as *objects* and are grouped into libraries.

In the following section we discuss external storage systems that are supported on System i:

- ▶ DS8000 Storage Server
- ▶ DS6000 Storage Server
- ▶ Enterprise Storage Server

3.5.1 System i storage architecture

Before we go into details on available tools and solutions for System i on rapid data recovery, we further explain the System i storage architecture. In System i architecture we distinguish between three types of disk pools or auxiliary storage pools (ASP):

- ▶ System ASP
- ▶ User ASP
- ▶ Independent ASP

The system creates the system ASP (ASP1) and is *always* configured. It contains the Licensed Internal Code, licensed programs, system libraries, and temporary system work space. The system ASP also contains all other configured disk units that are not assigned to a user ASP.

Grouping together a physical set of disk units and assigning them a number (2 through 32) creates a basic user ASP. User ASPs can be used to isolate objects on specific disk units.

An independent disk pool, or independent auxiliary storage pool (IASP), is a collection of disk units that can be brought online or taken offline independent of the rest of the storage on a system, including the system ASP, user ASPs, and other independent ASPs.

When an IASP is associated with a switchable hardware, it becomes a switchable IASP and can be switched between one System i system server and other System i system in a *clustered* environment or between partitions on the same system. DB2 objects can be contained in an IASP. This disk pool can contains objects, directories, or libraries that contain database files, application code, or different object attributes.

3.5.2 Independent auxiliary storage pool (IASP)

There are a number of reasons listed here why Independent Auxiliary Storage Pools should be used. One benefit of IASPs is that any disk failure in an IASP does not affect any other jobs or users on the system not using IASP (unlike basic ASP, which does eventually cause the system to halt). From a maintenance point of view, Reclaim Storage can now be run for each IASP without affecting the rest of the system, such as running in a Restricted State.

The most obvious benefit allows switched disk, where one tower of disks can be switched between two System i servers, or a disk Input/Output Processor (IOP) or Input/Output Adapter (IOA) between partitions. Multiple IASPs can have their addresses reclaimed concurrently.

The main advantage of IASP is server consolidation. Prior to IASPs, there are two choices:

- ▶ Restructuring the libraries used in an application
- ▶ Logical partitioning (LPAR)

The first choice requires a lot of application redesign, while LPAR does not remove the maintenance task of running multiple instance of OS/400. Using IASPs can allow a single instance of OS/400 to have multiple databases where the same library name can be used in different IASPs. Taking these advantages of IASPs and *combining* with external storage can bring even greater benefits in availability, backup, and disaster recovery.

Important: Before attempting to use IASPs and external storage, make sure that you have fully understood the limitations and which particular resiliency level you are achieving. The method, tools, and solutions discussed cannot be taken as a complete overall solution for your System i. You require careful consideration and System i specific requirements to perform the particular method, tools, or solutions. Contact IBM Global Services or your IBM Business Partner before attempting.

3.5.3 System i and FlashCopy

FlashCopy for System i is provided by DS8000, DS6000, and ESS systems. It provides a point-in-time copy of a disk. This gives a fast checkpoint so that full system backups can be performed with minimum application downtime.

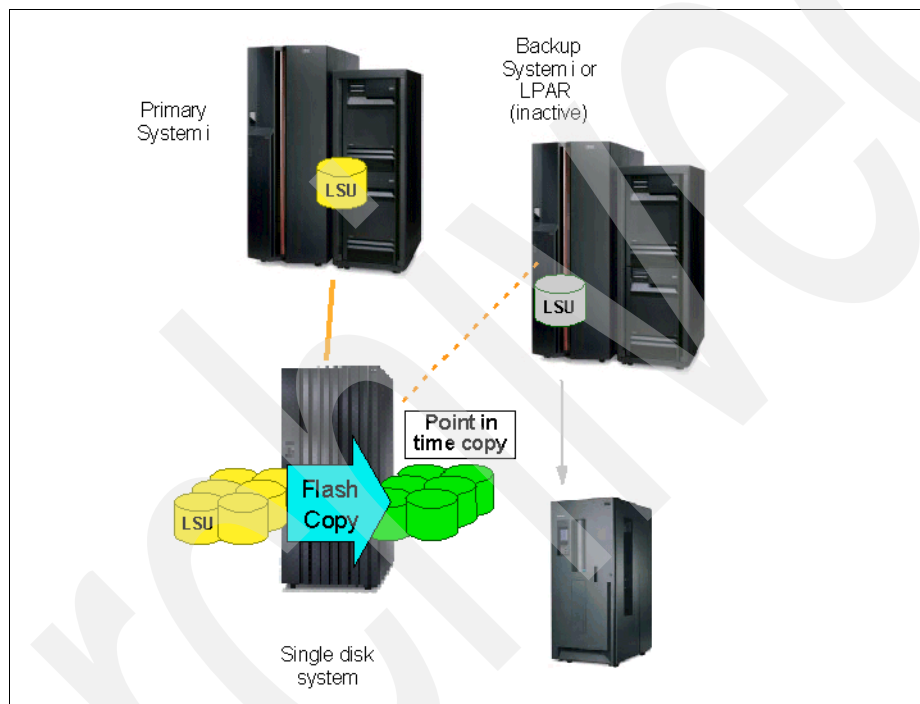


Figure 3-31 Basic FlashCopy process for System i

When FlashCopy was first available for System i, it was necessary to copy the entire storage space, including the Load Source Unit (LSU). However, the LSU must reside on an internal disk and this must first be mirrored to a LUN in the external storage. It is *not possible* to IPL from external storage and it is necessary to D-Mode IPL the target system or partition from CD to recover the Remote LSU. This might not be a practical solution and FlashCopy is not widely used for System i as opposed to other platforms where LSU considerations do not apply.

In order to ensure that the entire single-level storage is copied, memory has to be flushed. This can be accomplished by the following method:

1. Power down system (PWRDWN SYS).
2. Go into Restricted Mode (ENDSBS *ALL).

This method causes the system to be unavailable to all users. It might not be acceptable to you if you want your System i to be available at all times or have no tolerance level of system down time.

IASP FlashCopy

IASPs remove the tight coupling of the ASP from the LSU. This allows the IASP to FlashCopy independently of the LSU and other disks which make up the basic system disk pool. This has two major benefits:

- ▶ Less data is required to be copied.
- ▶ Remote LSU recovery is not necessary.

You should decide how to structure the application. If you are only considering offline backups as a primary concern, you might want to keep the application in system/basic user disk pool (*SYSBAS), but if it is used in conjunction with Metro Mirror (see 3.5.4, “Metro Mirror and Global Mirror” on page 155), you might want to include the application in an IASP.

The target system can be a live system (or partition) used for other functions such as test or development, and when backups are to be done, the FlashCopy target can be attached to the partition without affecting the other users. Alternatively, the target can be a partition on the production system, which has no CPU or memory initially allocated to it. In this case, when backups are taken, CPU and memory resources are then taken away or borrowed from the production environment (or others) and moved to the backup partition for the duration of the backup.

Like the basic FlashCopy, it is necessary to *ensure* to flush memory for those objects to be saved so that all objects reside in the external storage subsystem where the FlashCopy is performed. However, unlike basic FlashCopy, this is achieved by varying off the IASP. The rest of the system is unaffected.

Figure 3-32 shows FlashCopy for IASP.

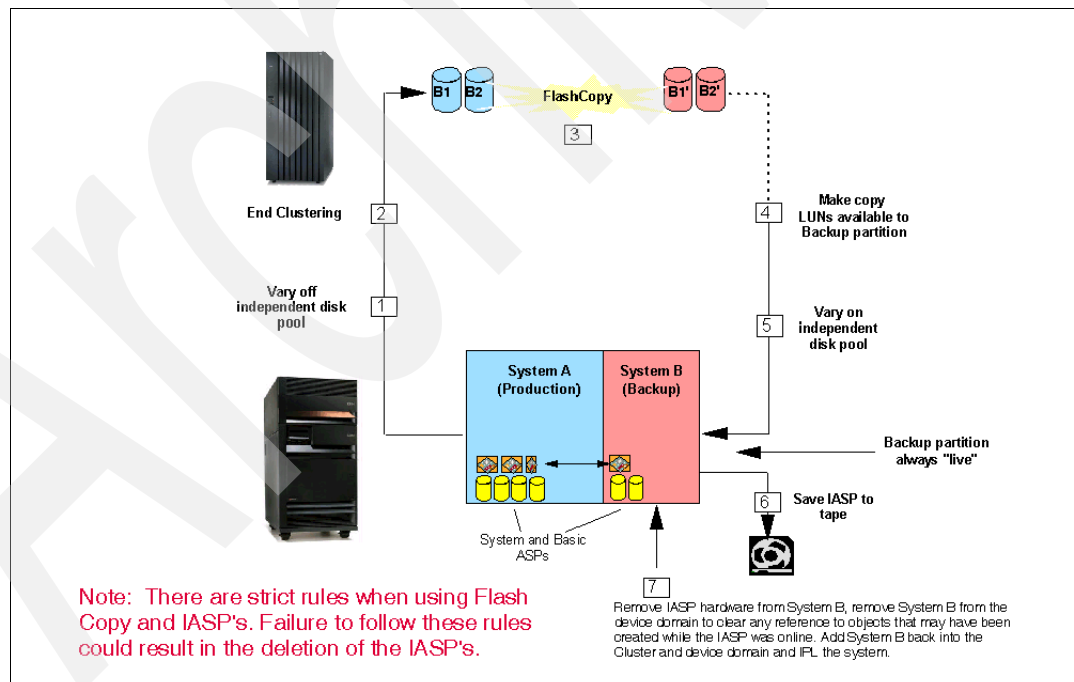


Figure 3-32 IASP FlashCopy for System i

It is critical that the steps shown are correctly performed, to preserve data integrity. To help manage this, IBM has the *iSeries™ Copy Services toolkit*. Using FlashCopy with IASPs without using this toolkit is *not supported* by IBM.

Currently, there is *no* CLI for running FlashCopy from OS/400, so this must be either be done using the Storage Management Console for DS8000 or DS6000. As each IASP is logically attached or detached on the target, the toolkit requires a separate IOP/IOA for each IASP. This approach is much more appealing to System i users than the basic FlashCopy discussed earlier.

3.5.4 Metro Mirror and Global Mirror

Conceptually, *Metro Mirror* is similar to FlashCopy, except that rather than a point-in-time copy being made, the copy is *synchronous*. Normally, Metro Mirror is done between two external storage systems, separated by distance to provide disaster recovery, although it is possible to run Metro Mirror within a single storage system.

When using a remote external storage system, the synchronous connection can be up to 300 km if using Fibre Channel (FC) for DS8000 and DS6000. Greater distances are supported on special request. However, with synchronous operation, the greater the distance is, the greater the impact on disk response times due to network latency. You have to consider the network bandwidth requirement and sizing when considering Metro Mirror. Figure 3-33 shows Metro Mirror for the System i environment.

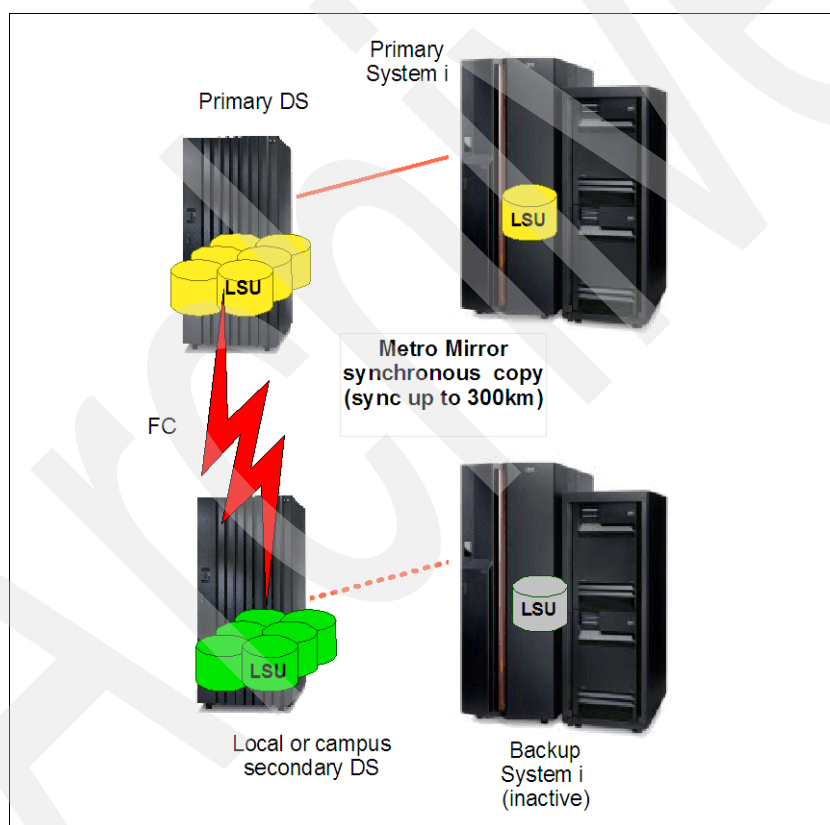


Figure 3-33 Metro Mirror for System i

Global Mirror, provides an *asynchronous* connectivity similar to Metro Mirror. For Global Mirror, distance can be much longer without having the impact on performance. Refer to the following IBM Redbooks on DS8000 and DS6000 on Global Mirror for more information:

- ▶ *IBM System Storage DS8000 Series: Concepts and Architecture*, SG24-6786
- ▶ *IBM System Storage DS6000 Series: Concepts and Architecture*, SG24-6781

The same considerations apply to the LSU as for FlashCopy. Unlike FlashCopy, which would likely be done on a daily or nightly basis, Metro Mirror is generally used for disaster recovery, and the additional steps required to make the Metro Mirror target usable are more likely to be accepted due to the infrequent nature of invoking the disaster recovery copy. Recovery time might be affected, but the recovery point is to the point of failure. For Metro Mirror and Global Mirror, there is no native Command-Line Interface (CLI) available in OS/400.

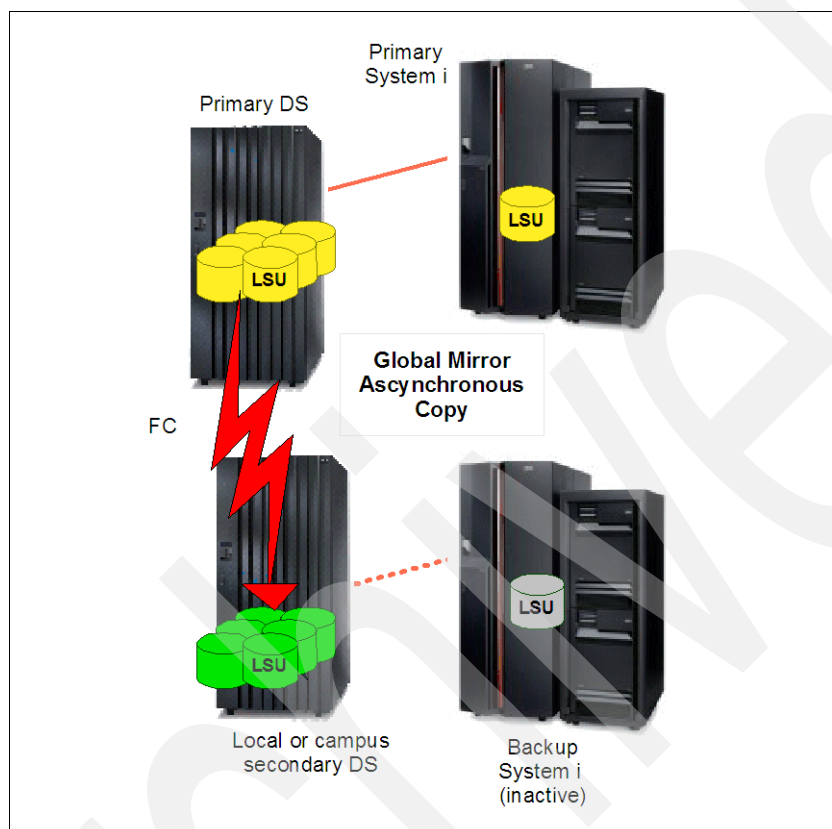


Figure 3-34 Global Mirror for System i

Important: With Metro Mirror and Global Mirror, some steps must be taken while invoking the disaster recovery target disk pool. You must perform a D-Mode IPL and recover the remote Load Source Unit, which attributes to an abnormal IPL. Your total recovery time must include this special step to obtain Recovery Time Objective (RTO).

Metro Mirror with IASPs

As discussed in 3.5.2, “Independent auxiliary storage pool (IASP)” on page 152, when using IASPs with FlashCopy, we can use this technology with Metro Mirror as well. Instead of the entire system (single level storage pool) being copied, only the application resides in an ASP and in the event of a disaster, the Metro Mirror target copy can be used and attached to the disaster recovery server.

Additional considerations include maintaining user profiles and other additional requirements for System i on both system. This is no different from using switched disk between two local System i servers on a High Speed Link (HSL). However, with Metro Mirror, the distance can be much greater than the limitation of HSL which is 250m. With Global Mirror, there is little performance impact on the production server.

With FlashCopy, you would likely have only data in IASP for disaster recovery, and with Metro Mirror, you would require the application on the disaster recovery system, where this can reside in system/basic user disk pool (*SYSBAS) in an IASP. If application is in an IASP, the entire application would have to be copied with Metro Mirror function. If the application is in *SYSBAS, good Change Management facilities are required to ensure that both systems have the same level of application. You *must ensure* that system objects in *SYSBAS such as User Profiles are synchronized. The disaster recovery instance can be a dedicated system or more likely a shared system with development, testing, or other function. Figure 3-35 shows the Metro Mirror with IASP on System i.

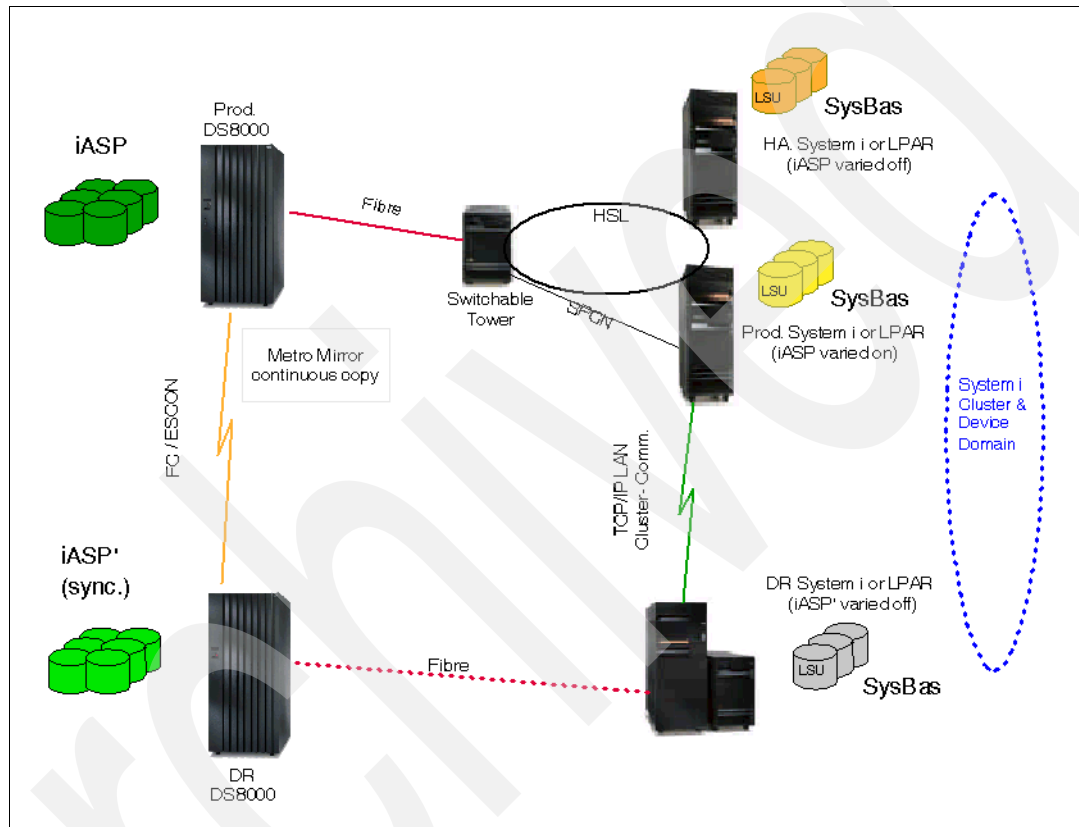


Figure 3-35 Metro Mirror for System i IASP

3.5.5 Copy Services Toolkit

The iSeries Copy Services Toolkit FlashCopy disk functions by doing offline saves using a second system. The production System i has the production copy of the application and database in an IASP. The production System i can be a secondary partition of an System i system.

To facilitate the use of FlashCopy of an IASP to a second system, both the production and backup System i *must exist* in a System i Cluster environment. Within the cluster, both systems must also reside in the same device domain. The use of the device domain requires a license for HA Switchable Resources (OS/400 option 41) on both nodes.

Important: Planning for Copy Services for DS8000 is part of the services engagement associated with purchasing Copy Service for System i. This includes planning for the IASPs, locations of the Disk IOPs/IOAs, and ensuring that all requirements and restrictions are followed. Contact IBM Global Services for more information about this Toolkit.

3.5.6 System i high availability concepts: FlashCopy

In the following sections we further explore examples of external storage used in conjunction with System i in a Business Continuity environment. However, this *does not include* any advanced functions of the disk systems themselves — rather, it is meant to introduce some of the OS/400 High Availability concepts.

Note: The following concept examples are based on DS8000. Disk systems DS6000 and ESS are also supported on System i and they have very similar Copy Services as DS8000. You can refer to the IBM Redbooks listed in 3.5.10, “Additional information” on page 166.

First we discuss basic connectivity for IBM System Storage attachment to System i. Availability is provided by the in-built redundancy in IBM System Storage subsystem as well as OS/400 switched disk. The I/O tower containing the IOPs/IOAs to attach to the storage subsystem is switchable between the production server and the HA server (or partition). System/basic user disk pool (*SYSBAS) would normally be held on internal disk drives in the System i although this is not absolutely necessary. Only Load Source Unit (LSU) has to be internal for System i. The switchable disks in this case would be the external storage subsystem. Figure 3-36 shows the basic connectivity for System i with DS8000.

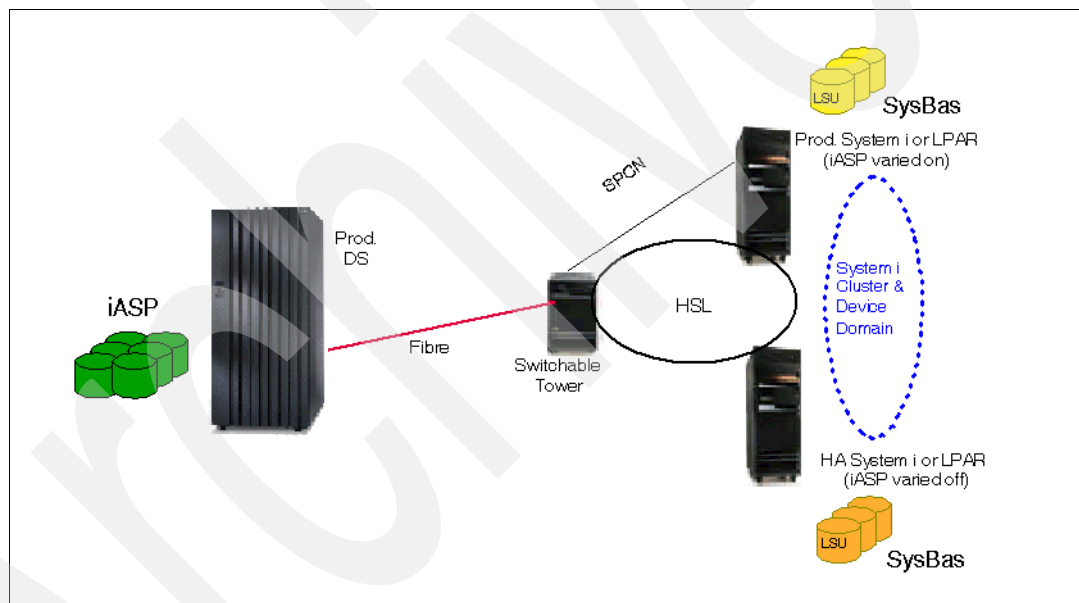


Figure 3-36 System i basic connectivity with DS8000

System i high availability: Concept 1

OS/400 supports multi-pathing — up to eight IOAs can address the same external storage LUNs. For greater resiliency, the additional IOA should be in a separate HSL loop, and each path to the LUNs should go via separate (non-connected) switches or directors. For simplicity, the SAN components are not shown in the following diagram and subsequent figures also do not show multi-path connections. However, this should always be considered for SAN reliability and redundancy. Figure 3-37 shows the multi-path redundancy for DS8000 connectivity to System i.

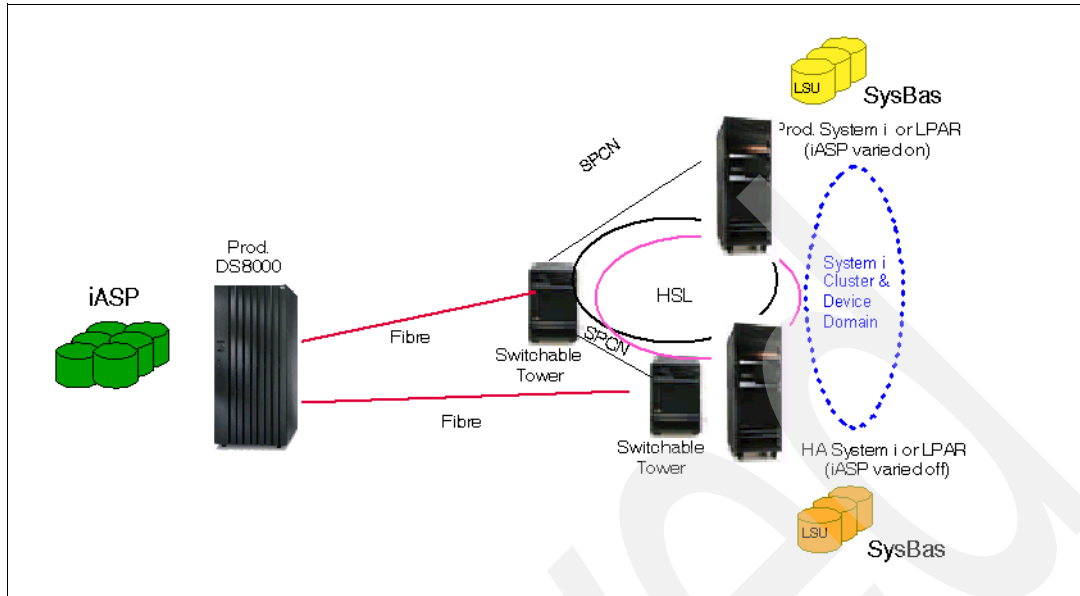


Figure 3-37 System i multi-path connectivity to DS8000

System i high availability: Concept 2

In this section, we further extend the concept of a switchable DS8000. In this example, two System i systems provide high availability for the other, with the production workloads (different applications) split between the two System i servers as shown in Figure 3-38.

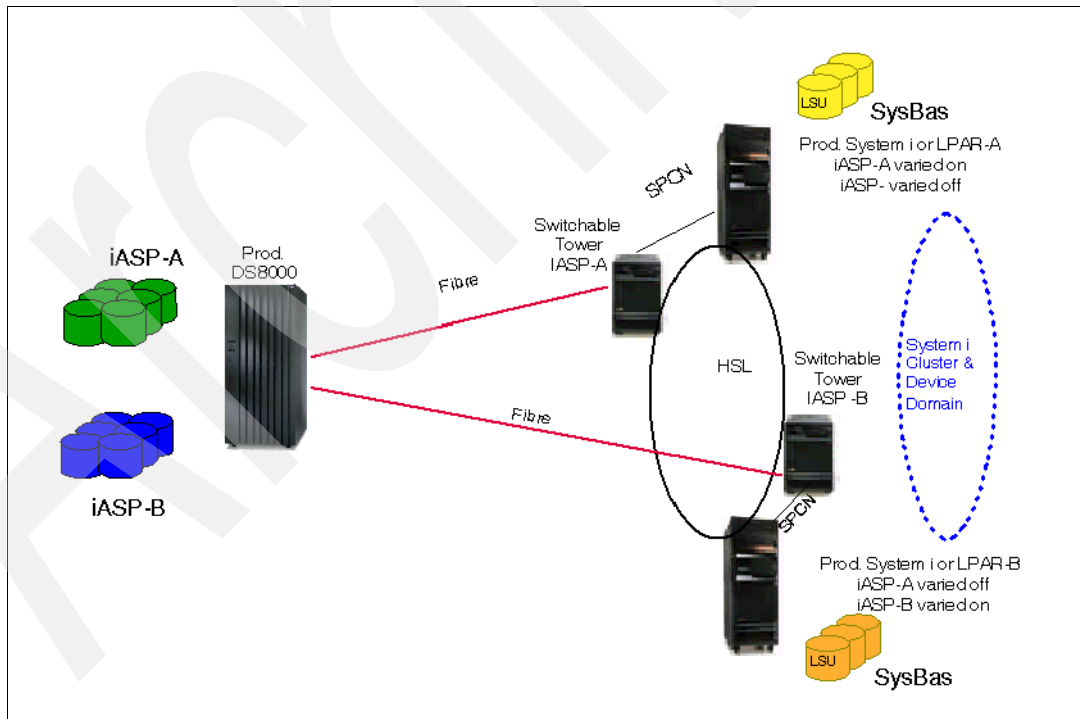


Figure 3-38 Split System i workload connectivity with DS8000

System i offline backup solutions with IASPs

In this section, we show examples of FlashCopy to provide a point-in-time copy of an IASP. This copy of the IASP is then made available to another system or partition (LPAR) for offline tape backups.

All the solution examples in this section requires the System i Copy Services for DS8000, DS6000, and ESS service offering described in 3.5.5, “Copy Services Toolkit” on page 157. When a backup is required, the application must be quiesced and the IASP *must be* varied off to flush all the data from memory to disk. When this is completed (usually a matter of seconds), the IASP can be varied back on to the production server or LPAR, after which the application can be restarted and the clone of the IASP can be varied on to the backup server or LPAR where backups can be taken.

Attention: This process *must be* performed through the Copy Services Toolkit.

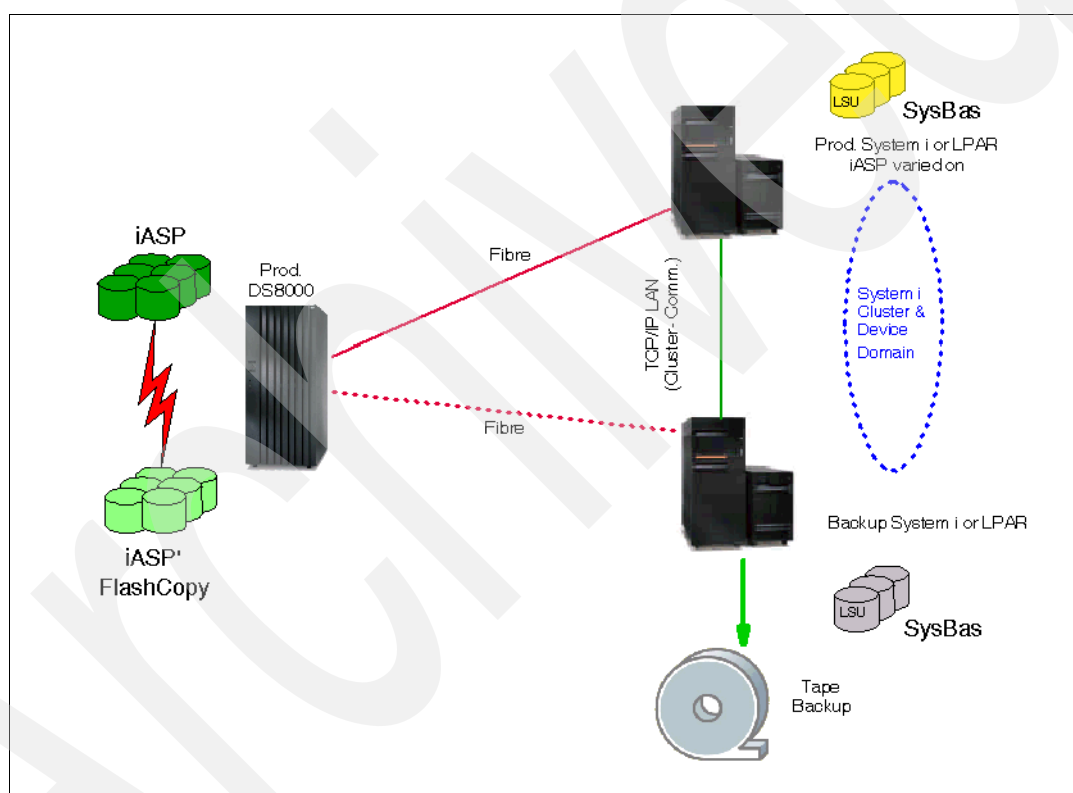


Figure 3-39 Offline backup for System i using Copy Services for System i

Figure 3-39 shows the offline backup schematic for System i. In any circumstances, backups would be taken in a secondary partition and this could utilize memory and processor resources from the production partition. However, this is not necessary because any System i server supporting Fibre Channel and OS/400 V5R2 could be used.

Multiple System i offline backups solutions

A single System i system or LPAR can act as the target for multiple production systems/LPARs. Backups for each server or LPAR can be run concurrently or one after another. However, it is necessary for each IASP clone to be managed individually by Copy Services for System i toolkit and each IASP *must have* dedicated IOPs/IOAs. Figure 3-40 and Figure 3-41 show the offline backups for multiple System i servers.

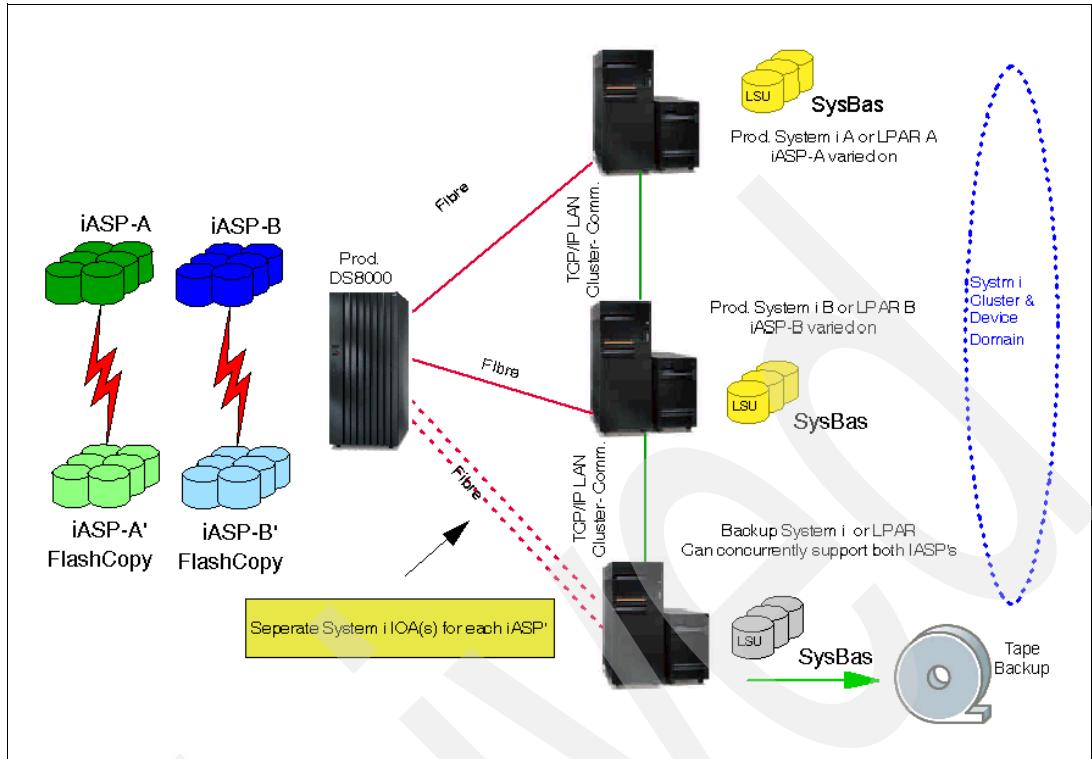


Figure 3-40 Multiple offline backups for System i

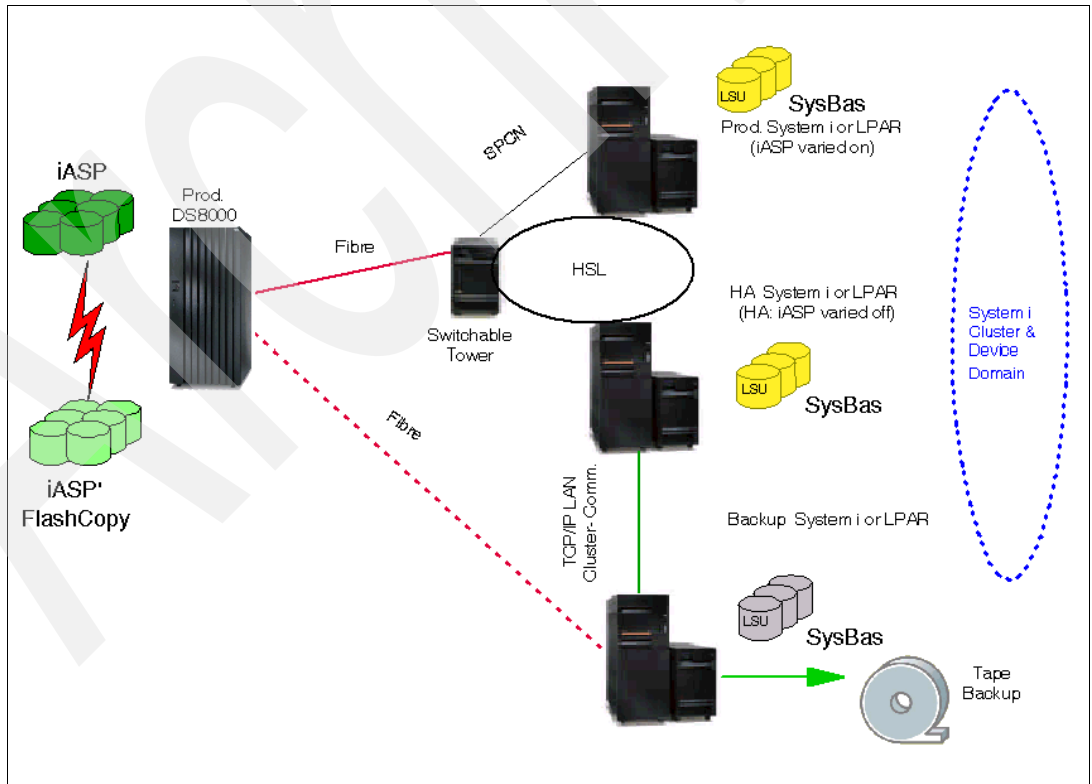


Figure 3-41 Concurrent multiple offline backups for System i

3.5.7 System i high availability concepts: Metro Mirror

Metro Mirror provides a continuous remote copy of an IASP, which is then made available to another system or LPAR for disaster recovery. In principle, there is no difference when running either Metro Mirror or Global Mirror although the hardware requirements are different.

In addition to Metro Mirror, we show some examples of Metro Mirror being used in combination with FlashCopy to provide both disaster recovery and off-site online backups. All the solution examples in this section require System i Copy Services Toolkit as discussed in 3.5.5, “Copy Services Toolkit” on page 157. Figure 3-42 shows the basic Metro Mirror function for System i connectivity.

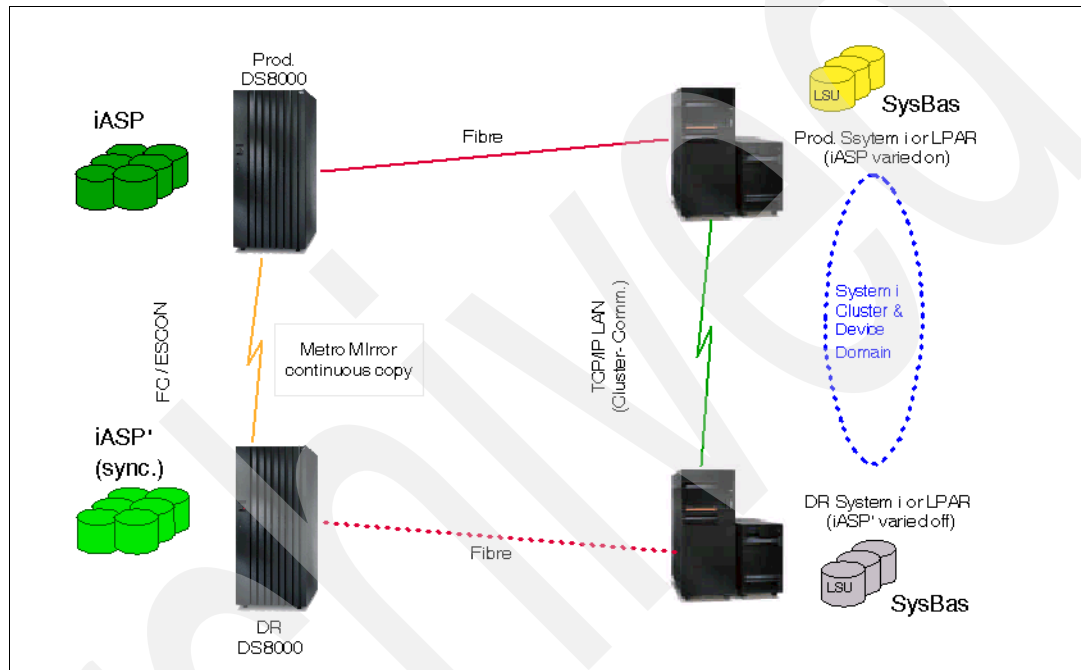


Figure 3-42 DS8000 Metro Mirror for System i

Metro Mirror function with IASP

As Metro Mirror provides a continuous remote copy of an IASP, the production server has its application(s) residing in one or more IASPs. In the event of a failure of either the production server or local storage subsystem, the service can be switched to the remote server and storage subsystem.

Important: This is not switching of the IASP in the way that you can switch a tower of internal disks. You require the Metro Mirror Toolkit to perform this Metro Mirror function.

For planned outages, the application would be quiesced and the IASP would be varied off from the production server. This ensures a clean termination of the production workload and a speedy vary on to the remote server. In the event of an unplanned outage, the vary on of the IASP on the remote site would go through database recovery steps in the same way in a switched tower of internal disks. To minimize the database recovery time, the application *should use* journaling to the IASP (or a secondary IASP) so that the journal receivers are available to assist with the recovery. Or, commitment control could be used to ensure transaction integrity.

Figure 3-43 shows Metro Mirror for System i IASPs.

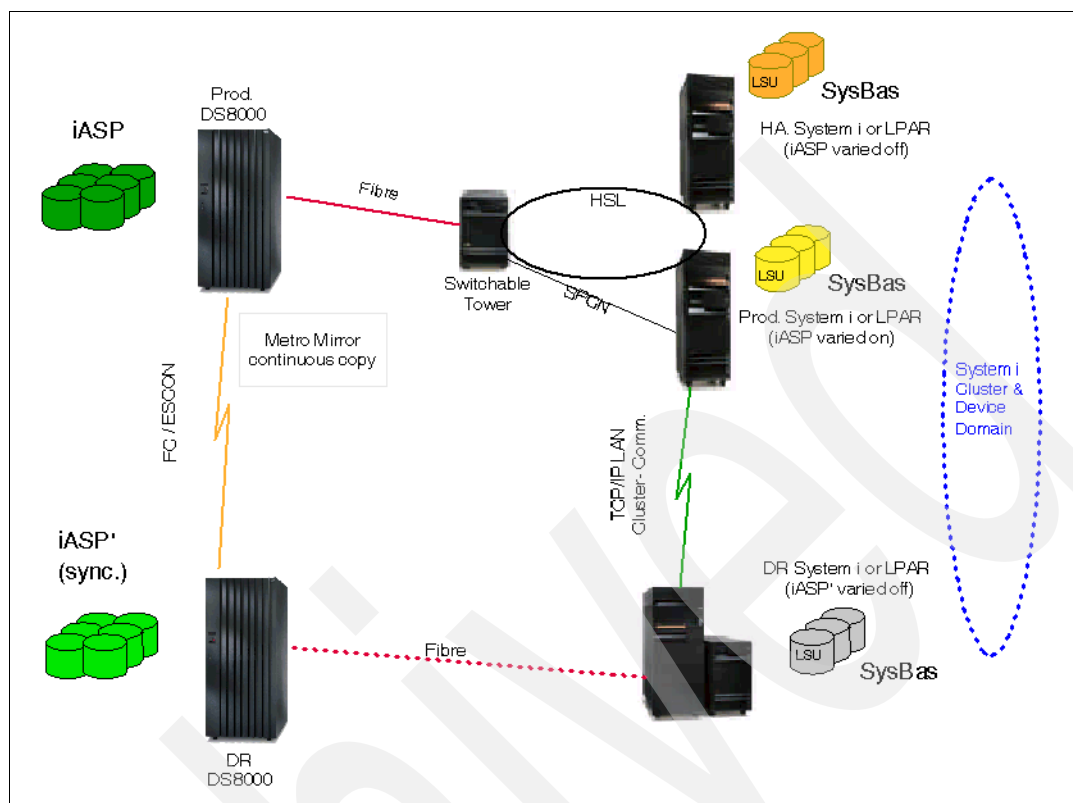


Figure 3-43 Metro Mirror for System i IASPs

System i remote offline backups with Metro Mirror

Metro Mirror is extended to disaster recovery capabilities by providing a third copy of data using FlashCopy in the remote site. The second copy (IASP) is the disaster recovery copy, and the third copy is used for taking offline backups. Normally, we would expect two partitions (or separate servers) in the remote site to allow more timely failover. If a failure of the production server should occur while backups were being taken, you would have to abort the backups to bring up the service on the disaster recovery system, or wait until backups are completed. In either of these cases, you would not have the ability to do offline backups, because *only one* OS/400 instance is available at the remote site.

Although this solution requires three copies of data, it might be possible to use 2x size Disk Drive Modules (DDMs) at the remote site. Performance of the Metro Mirror writes might suffer slightly during the time that the backups were being taken off the third copy of data (due to these sharing of the same DDMs) but this might be an acceptable compromise.

In the following diagrams, notice that Figure 3-45 is similar to Figure 3-44, but the FlashCopy is taken locally rather than at the remote site. In the event of only a production server being unavailable, production would be switched to the remote site and the direction of Metro Mirror could be reverse to allow an up-to-date copy at the production site and FlashCopy to be done from there. However, there might be no server to do the backups if LPAR was used. If the entire local site is down, and the DS8000 disk system is unavailable, the backups cannot be done off a copy of IASP. Alternatives would have to be provided at the disaster recovery site for backing up the data.

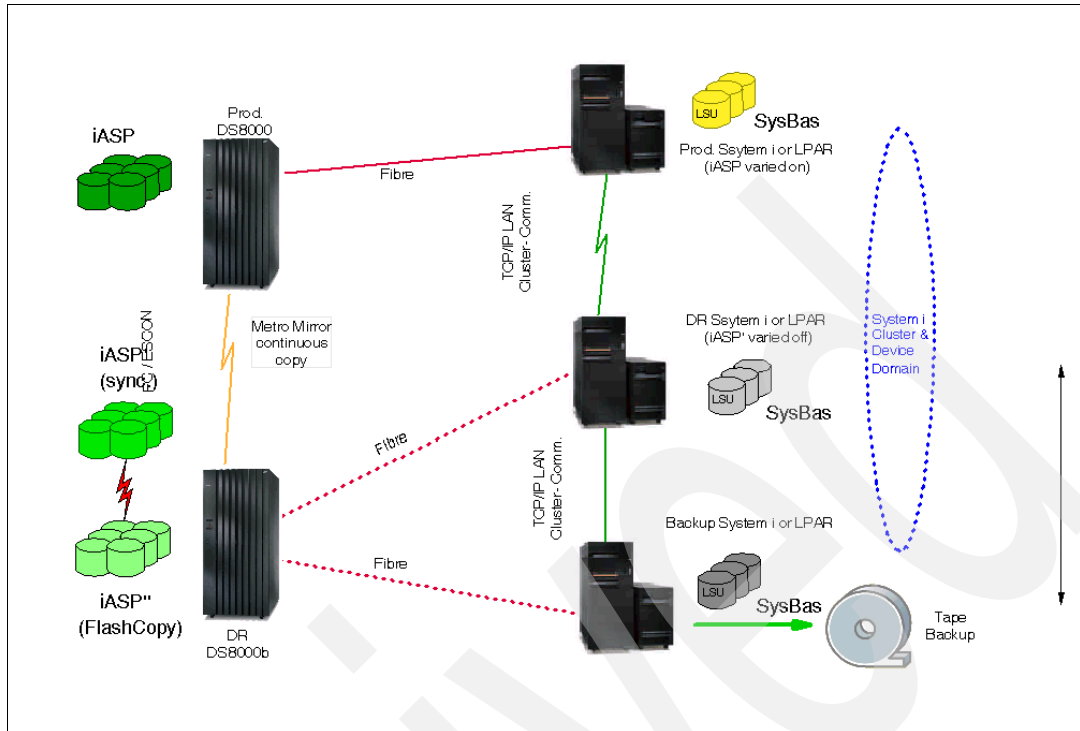


Figure 3-44 Remote offline backups for System i example 1

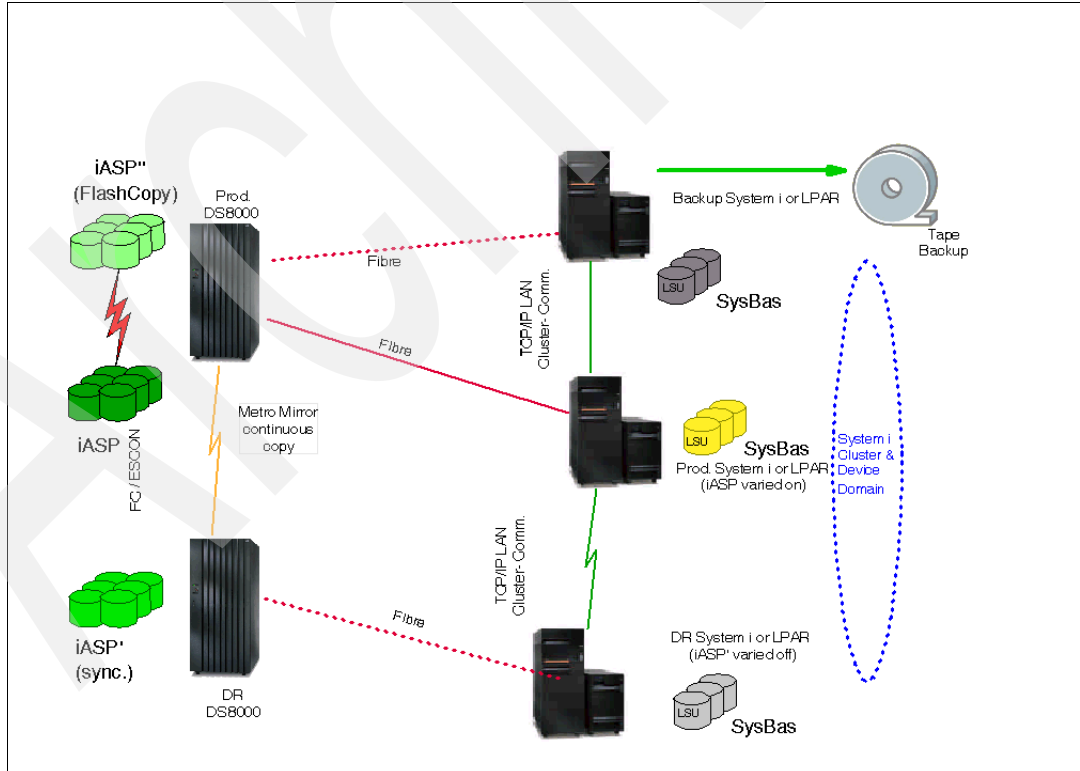


Figure 3-45 Remote offline backups for System i example 2

3.5.8 Cross Site Mirroring solutions

Cross Site Mirroring (XSM) was introduced in V5R3. It allows you to maintain two identical copies of an IASP at two sites that are geographically separated. XSM ensures greater protection and availability than available on a single system. It increases the options for defining backup nodes to allow for failover or switchover methods.

In the following examples shown in this section, we use both local High-Availability (HA) (by switching the IASP between two local nodes on the same HSL) and disaster recovery (by having a mirrored copy in a remote site). It is not necessary for the IASP to be switchable, and a disaster recovery solution could be provided by only having a single node at the production site. However, *full synchronization is required* for any persistent transmission interruptions or communication loss between the source and the target systems for an extended period of time.

To minimize the potential for this situation, we recommend that you use redundant communication links and use XSM in at least a three system configuration where the production copy of the IASP can be switched to another system or partition at the same site that can maintain geographic mirroring. Because this is a switch of an IASP, it is quite feasible that this local HA system could be utilized for test, development, or other critical workloads during normal operations.

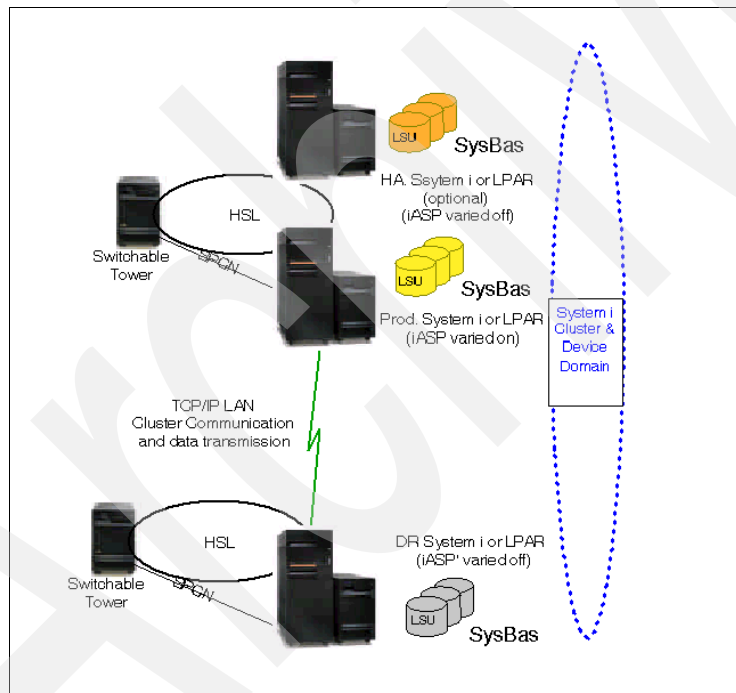


Figure 3-46 System i Cross Site Mirroring

Here are some important considerations before utilizing XSM:

- ▶ Facilities exist to suspend or detach the mirrored copy should be avoided if possible due to having to perform a full re-sync from source to target.
- ▶ Target copy can only be accessed when detached, and consequently a full re-sync is required when re-attached.

Important: Using XSM with FlashCopy allows you to have a static copy of data to be created while the IASP was varied off, thus avoiding re-synchronization issues involved with splitting the XSM Mirror. However, although this would be technically possible, *it is not recommended* and a much better solution would be to use Metro Mirror to maintain the remote copy and use FlashCopy of the Metro Mirror target. In this way, both the disaster recovery and backup facilities are being done with a common set of functions using DSS Copy Services and the toolkit, rather than a hybrid of two technologies.

3.5.9 Summary

All disaster recovery tools and concepts suggested for System i require careful planning and design phases in order to achieve the desired objectives. Improper implementation can lead to system outages or data corruption, which would inevitably cause you to restore the entire System i system. Review the IBM Redbooks and Web sites listed next for more detailed discussions on all the suggested disaster recovery tools, or contact IBM Global Services or Business Partners.

3.5.10 Additional information

Refer to the following IBM Redbooks for more information and discussion:

- ▶ *IBM System i5, eServer i5, and iSeries Systems Builder IBM i5/OS Version 5 Release 4 - January 2006*, SG24-2155
- ▶ *Clustering and IASPs for Higher Availability on the IBM eServer iSeries Server*, SG24-5194
- ▶ *IBM eServer iSeries Independent ASPs: A Guide to Moving Applications to IASPs*, SG24-6802
- ▶ *Introduction to Storage Area Networks*, SG24-5470
- ▶ *iSeries in Storage Area Networks A Guide to Implementing FC Disk and Tape with iSeries*, SG24-6220
- ▶ *IBM System Storage DS8000 Series: Concepts and Architecture*, SG24-6786
- ▶ *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547
- ▶ *Data Resilience Solutions for IBM i5/OS High Availability Clusters*, REDP-0888

Also refer to the following Web sites for more information and discussion:

- ▶ <http://poupublic.boulder.ibm.com/iseriess>
- ▶ <http://www.ob/cp,serverseserver/iseriess/service/itc/pdf/Copy-Services-ESS.pdf>

3.6 FlashCopy Manager and PPRC Migration Manager

FlashCopy Manager and PPRC Migration Manager are IBM Storage Services solutions for z/OS users of FlashCopy and Metro Mirror. For this environment, these packaged solutions are designed to:

- ▶ Simplify and automate the z/OS jobs that set up and execute a z/OS FlashCopy, Metro Mirror, or Global Copy environment
- ▶ Improve the speed of elapsed execution time of these functions
- ▶ Improve the administrator productivity to operate these functions

These two related tools use a common style of interface, operate in a very similar fashion, and are designed to complement each other. A user familiar with one of the offerings should find the other offering easy to learn and use.

In the following sections we discuss these topics:

- ▶ FlashCopy Manager overview and description
- ▶ PPRC Migration Manager overview and description
- ▶ Positioning PPRC Migration Manager and GDPS

3.6.1 FlashCopy Manager overview

FlashCopy Manager is a z/OS-based FlashCopy integration and automation package, delivered through IBM Storage Services. It is designed to provide significant ease of use, automation, and productivity for z/OS users of FlashCopy.

3.6.2 Tier level and positioning within the System Storage Resiliency Portfolio

Tier level: 4 (for the FlashCopy Manager)

FlashCopy allows businesses to make point-in-time copies of their mission critical data for testing and or backup and recovery. FlashCopy Manager is considered a Tier 4 Business Continuity tool, because the FlashCopy data is as current as the most recent invocation of FlashCopy; typically, this might be several hours old.

3.6.3 FlashCopy Manager solution description

FlashCopy Manager is a series of efficient, low overhead assembler programs and ISPF panels that allow the z/OS ISPF user to define, build, and run FlashCopy jobs for any sized FlashCopy z/OS environment.

In particular, z/OS users who have a large number of FlashCopy pairs, in the hundreds or thousands of FlashCopy pairs, should find FlashCopy Manager of significant value.

3.6.4 FlashCopy Manager highlights

FlashCopy Manager provides a straightforward set of ISPF menu panels, used in a question and answer format, to obtain the necessary parameters for building FlashCopy jobs. FlashCopy Manager makes all standard FlashCopy V 1 and V 2 options available through an ISPF interface. These options are fully described in 7.7.1, “Point-In-Time Copy (FlashCopy)” on page 270).

FlashCopy Manager programs provide the capability to:

- ▶ Identify FlashCopy source volumes via specification of source or target volumes, by:
 - Volser (with wildcards)
 - DFSMS storage groups
 - Device number ranges
- ▶ Automatically discover candidate FlashCopy targets using the same capabilities as described in the previous list item
- ▶ Correctly match the candidate targets with defined FlashCopy source volumes
- ▶ Build the z/OS jobs to execute the FlashCopy environment
- ▶ Query the progress of the FlashCopy as well as the background copy
- ▶ Provide an audit trail of FlashCopy parameter specification

FlashCopy Manager is written in very efficient assembler language, and performs this discovery, matching, and job build process very quickly with very low overhead. Because of the fast speed of execution, FlashCopy Manager is especially useful for configurations extending into many hundreds or thousands of FlashCopy pairs.

FlashCopy Manager has special programs that dynamically and very quickly obtain device information required to develop the configuration and to create the files and jobs required for FlashCopy establish. FlashCopy Manager is fully aware of, and dynamically discovers and correlates DS8000, DS6000, and ESS800 specifics such as SSID, storage controller serial number, z/OS VTOC structures, and volume size matching (source and target volumes are matched based on equal size).

FlashCopy Manager provides a special high performance program that executes a FlashCopy **freeze** and **run** Consistency Group (CG) formation sequence with a faster execution time than the normal FlashCopy CG capability. This is designed to make possible the creation of a point in time *consistent copy* of z/OS data even in a very large FlashCopy environment.

FlashCopy Manager provides an audit trail, so that the definition and running of jobs under its control can be tracked and managed. FlashCopy Manager provides valuable diagnostic information to the user when each command is executed. It is capable of notifying users of any problems involved in issuing FlashCopy commands, as well as reporting back how long it took to execute the FlashCopy across all volumes or data sets. This data is available through the SYSLOG.

3.6.5 FlashCopy Manager components

FlashCopy Manager requires both hardware and software, as described next.

Hardware

The following hardware is required:

- ▶ Both source and target FlashCopy volumes must be contained within a single IBM Enterprise Storage Server, DS8000, or DS6000. The Point in Time or FlashCopy license must be installed.
- ▶ Both the source and target devices of a FlashCopy pair must be accessible from a single z/OS environment.

Software

At the time of writing this book, the most current level of FlashCopy Manager is V3R5. FlashCopy Manager operates on any current version of z/OS with TSO and ISPF:

- ▶ FlashCopy Manager's load library must be APF authorized.
- ▶ FlashCopy Manager has no known z/OS system release dependencies.

Summary

FlashCopy Manager provides z/OS users an improved interface, featuring an easy to use ISPF interface to dynamically discover, build, and execute, small and large FlashCopy environments to further enhance the use of FlashCopy technology.

3.6.6 PPRC Migration Manager overview

The PPRC Migration Manager provides a series of efficient, low overhead assembler programs and ISPF panels that allow the z/OS ISPF user to define, build, and run DS8000, DS6000, and ESS Metro Mirror and Global Copy jobs for any sized environment.

In particular, z/OS users who have a large Metro Mirror environment, in the hundreds or thousands of pairs, should find PPRC Migration Manager of significant value. PPRC Migration Manager supports both planned and unplanned outages.

3.6.7 Tier level and positioning within the System Storage Resiliency Portfolio

Tier level: 6 (for the PPRC Migration Manager with Metro Mirror)

Tier level: 4 (for the PPRC Migration Manager with Global Copy)

PPRC Migration Manager is considered a Tier 6 Business Continuity tool when it controls synchronous Metro Mirror storage mirroring, as the remote site is in synchronous mode with the primary site.

PPRC Migration Manager is considered a Tier 4 Business Continuity tool when it controls non-synchronous Global Copy, because the remote site data is not in data integrity until the Global Copy *go - to - sync* process is done to synchronize the local site and the remote site.

3.6.8 PPRC Migration Manager description

PPRC Migration Manager provides a straightforward set of ISPF menu panels, used in a question and answer format, to obtain the necessary parameters for building Metro Mirror and Global Copy jobs. PPRC Migration Manager makes all standard Metro Mirror options available through the ISPF interface. These options are fully described in (see 7.7.3, “Remote Mirror and Copy (Peer-to-Peer Remote Copy)” on page 277).

PPRC Migration Manager programs provide the capability to:

- ▶ Identify Metro Mirror or Global Copy source volumes via specification of source or target volumes, by:
 - Volser (with wildcards)
 - DFSMS storage groups
 - Device number ranges
- ▶ Automatically discover candidate Metro Mirror secondaries using the same capabilities described in the previous item
- ▶ Correctly match the candidate targets with defined Metro Mirror source volumes
- ▶ Build the z/OS jobs to establish and control the Metro Mirror environment
- ▶ Query the progress of the Metro Mirror as well as the background copy
- ▶ Provide an audit trail of Metro Mirror parameter specification

Note that PPRC Migration Manager (and FlashCopy Manager) support Global Mirror for z/OS environments.

PPRC Migration Manager is written in very efficient assembler language, and performs this discovery, matching, and job build process very quickly with very low overhead. Combined with fast speed of execution, PPRC Migration Manager is especially useful for configurations extending into many thousands of Metro Mirror pairs.

The PPRC Migration Manager allows the user to define the Metro Mirror environment using the device number and disk subsystem HBA physical ports used in the Metro Mirror links between boxes. PPRC Migration Manager has special programs that dynamically and very quickly obtain the device information required to develop the configuration, and to create the files and jobs required for Metro Mirror Establish.

With these programs, PPRC Migration Manager dynamically discovers and correlates DS6000, DS8000, and ESS specifics such as SSIDs, LSSids, serial numbers, and SAIDs. VOLSER information is also included to enhance user understanding of the environment. Even though VOLSERs can change with time, a **REFRESH** job updates PPRC to use the correct VOLSERs. A reverse process is available to simplify the creation of a PPRC configuration in which the flow of data is from the recovery site to the production site.

These features represent a great improvement in the ease of use for PPRC configurations in either Metro Mirror or Global Copy mode, especially for large configurations of volume pairs.

3.6.9 Diagnostic tools

PPRC Migration Manager is also useful for diagnosing problems within the Metro Mirror relationship. This is accomplished in a number of ways:

- ▶ When data sets are allocated, the information is displayed on the screen. As a result, it is possible to see the work progress and identify if problems have occurred.
- ▶ When an operation changes the state of Metro Mirror relationships, it follows three steps:
 - First it obtains and records the current Metro Mirror and FlashCopy state of the environment.
 - Next it executes the change request.
 - Finally, it obtains and records the new Metro Mirror and FlashCopy state of the environment.

Note: All state change data recorded is documented and made easily available through the SYSLOG.

- ▶ When problems occur, it is possible to run a job to force state saves on all subsystems in the configuration. Doing so allows errors to be diagnosed more easily, no matter where they start.

3.6.10 Support for FlashCopy

It is valuable to have a tertiary, point-in-time (PIT) copy of data available for both disaster recovery and testing purposes. PPRC Migration Manager supports this business requirement by enabling automated commands for using FlashCopy to make a tertiary copy of the Metro Mirror secondaries.

3.6.11 Support for Metro Mirror Consistency Group FREEZE

To support consistency in secondary data, PPRC Migration Manager contains specially designed programs that support the Metro Mirror Consistency group **FREEZE** process. This command is passed through the AO manager and simultaneously affects all subsystems in the configuration. As a result, PPRC Migration Manager is able to control a split-mirror configuration, in which the secondary site data consistency is assured when splitting the Metro Mirror pairs.

3.6.12 Modes of operation

PPRC Migration Manager can be used for data migration (Normal Mode), or for disaster recovery purposes (Clone Mode). These two modes of operation result in two different types of Metro Mirror control jobs, depending on the intended use.

Normal Mode (for data migration)

Normal Mode is the option selected during data migration. This enables a business to bring data from one set of Metro Mirror compatible subsystems to another via Metro Mirror or Global Copy.

Also, to solve security considerations when returning old disk subsystems back to the vendor, Normal Mode has jobs that issue the appropriate ICKDSF commands required to securely clear all data at the end of the migration. As a result, upgrading to new z/OS disk becomes easier, and does not put any intellectual property that had been stored on the old subsystems at risk.

Clone Mode (for unplanned outage disaster recovery)

Clone Mode is the option selected for disaster recovery purposes. Under Clone Mode, jobs are made available to support planned and unplanned outages. The PPRC Migration Manager has jobs that monitor the status of each of the Metro Mirror pairs. If an event occurs that suspends Metro Mirror, it is capable of initiating support for an unplanned outage in the form of a **FREEZE and GO**. This ensures the consistency of data in the secondary disk, so that databases can be restarted if a disaster has occurred, but applications continue to run in the production configuration.

However, it is important to note that this does not include any automation for halting applications or restarting systems in a remote data center. That level of automation is available through GDPS (see 2.1, “Geographically Dispersed Parallel Sysplex (GDPS)” on page 10).

To properly function in a disaster recovery environment, the PPRC Migration Manager code should be run in an LPAR that does not access the controlled storage. This ensures that it is not slowed by any processes that are affecting said devices during periods of high activity.

3.6.13 Use in a disaster recovery environment

Because the PPRC Migration Manager helps to achieve data consistency with zero or little data loss, it would typically be considered a Tier 6 tool. This would be a valuable tool for businesses with a requirement for such consistency on z/OS data only, but no firm requirement for an automated restart of servers or applications.

3.6.14 PPRC Migration Manager prerequisites

PPRC Migration Manager requires both hardware and software, as described next.

Hardware

The following hardware is required:

- ▶ When used for disaster recovery, both the Metro Mirror primary and Metro Mirror secondary devices must be in a DS6000, DS8000, or ESS800 with Metro Mirror. To use the full CLONE functions, the FlashCopy feature must also be enabled.
- ▶ The Metro Mirror secondary devices must be accessible from a z/OS, OS/390®, or MVS™ system during the setup process. They do not have to be varied online during this phase of activity, nor do they have to have labels and VTOCs on them.
- ▶ The location of the FlashCopy targets relative to the Metro Mirror secondary devices is Disk System microcode release dependent.

Restriction: PPRC Migration Manager is intended only for the z/OS environment; hence, it does not support DS4000 or SAN Volume Controller disk mirroring.

PPRC Migration Manager supports Global Mirror for z/OS environments.

Software

These are the software requirements:

- ▶ The PPRC Migration Manager load library must be APF authorized.
- ▶ The TSO Metro Mirror and FlashCopy commands must be installed, and the user and jobs submitted must have RACF® authority to use these commands.
- ▶ The PPRC Migration Manager has no known z/OS, OS/390, or MVS system release dependencies.

Note: The name PPRC Migration Manager has not been changed at the current time. However, PPRC Migration Manager is fully able to control both Metro Mirror and Global Copy on the DS8000, DS6000, and ESS.

3.6.15 Positioning of PPRC Migration Manager and GDPS

GDPS: The comprehensive IBM Business Continuity automated solution for the System z IT environment is the Geographically Dispersed Parallel Sysplex (GDPS) offering, as detailed in 2.1, “Geographically Dispersed Parallel Sysplex (GDPS)” on page 10.

GDPS provides an end to end System z infrastructure integration of servers, storage, software and automation, networking, and installation services. GDPS is serviced with 24x7 worldwide support by IBM Global Services; is a standard product-level participant in the IBM worldwide support structure, and is enhanced and updated on an ongoing basis by IBM.

GDPS HyperSwap Manager, as detailed in 3.1, “System Storage Rapid Data Recovery: System z and mixed z+Open platforms (GDPS/PPRC HyperSwap Manager)” on page 108, is a entry level GDPS offering that manages only disk mirroring (including HyperSwap). With a reduced price to go along with the reduced scope, GDPS HyperSwap Manager might be an ideal GDPS entry point for clients desiring to eventually grow their Business Continuity functionality to a fully automated server, storage, software, networking solution. Users of GDPS HyperSwap Manager can upgrade to full GDPS/PPRC, building upon and protecting the full investment already made in Metro Mirror under GDPS HyperSwap Manager control.

PPRC Migration Manager: While GDPS and GDPS HyperSwap Manager are the strategic solutions for System z-based IT infrastructure Business Continuity, specific System z clients might have lower required levels of recoverability, a more limited scope of protection, or budget constraints that affect the selection of tool disk mirroring subsystem management and outage protection. With those reduced requirements, PPRC Migration Manager provides an option for a basic disk mirroring management and automation package.

For these particular requirements, where a future upgrade to full GDPS is not a part of the client plan, IBM Storage Services offers the PPRC Migration Manager.

PPRC Migration Manager provides a basic, quick to install, yet highly scalable disk mirroring management package for users of DS8000, DS6000 and ESS Metro Mirror and Global Copy.

Note: PPRC Migration Manager does not support Global Mirror on the DS8000, DS6000, and ESS800. PPRC Migration Manager does not support disk mirroring for non-System z data or distributed systems disk mirroring subsystems.

3.6.16 Summary

Where the z/OS client requirements do not extend to a comprehensive server, storage, software, and automation Business Continuity solution, and are more focused on a basic, low-cost automation package, the FlashCopy Manager and PPRC Migration Manager can provide the basic IBM Storage Services automation solutions.

FlashCopy Manager supports z/OS users of DS8000 / DS6000 / ESS FlashCopy. PPRC Migration Manager supports z/OS users of DS8000 / DS6000 / ESS Metro Mirror (sync PPRC) Global Copy (PPRC-XD), and Global Mirror for z/OS environments.

For the basic z/OS environment, these packaged solutions:

- ▶ Simplify and automate the setup of either FlashCopy, Metro Mirror, or Global Copy.
- ▶ Improve the administrative productivity by significantly reducing the effort and amount of time required to manage these functions.

These two related tools both use a common style of interface, operate in a very similar fashion, and are designed to complement each other.

For more information, contact your IBM Storage Services representative.

An online Web site page, with a contact form, is available at:

http://www.ibm.com/servers/storage/services/featured/pprc_mm.html

Archived

Backup and restore

Backup and restore is the most basic business continuity solution segment for protecting and recovering data, by making additional data copies. The data copies allow you to restore data back to the point in time when it was backed up.

The backup and restore process usually involved tape media; however, backup to disk is becoming popular for faster backup and restore time.

The Backup and Restore segment and the BC solution tiers relationship has been described in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.

In this chapter we describe how archive data can take part in a Business Continuity strategy. We also provide an overview of IBM Tivoli Storage Manager to illustrate the various BC solution tiers where backup and restore can be used in Business Continuity.

Some solution examples of Business Continuity Tier 1, 2, and 3 are covered in 4.6, “Solution examples of backup, restore, and archive” on page 211.

4.1 An overview of backup and restore, archive and retrieve

Backup is a daily IT operation task where production, application, systems, and user data are copied to a different data storage media, in case they are required for restore. Restoring from a backup copy is the most basic Business Continuity implementation.

As part of the Business Continuity process, archive data is also a critical data element which should be available. Data archive is different from backup in that it is the only available copy of data on a long term storage media, normally tape or optical disk. The archive data copy is deleted at a specific period of time, also known as retention-managed data.

4.1.1 What is backup and restore?

To protect against loss of data, the backup process copies data to another storage media which is managed by a backup server. The server retains versions of a file according to policy, and replaces older versions of the file with newer versions. Policy includes the number of versions and the retention time for versions.

A client can restore the most recent version of a file, or can restore previous retained versions to an earlier point in time. The restored data can replace (over-write) the original, or be restored to an alternative location, for comparison purposes.

Note: More information about IBM Tivoli Storage Manager is provided in 4.3, “IBM Tivoli Storage Manager overview” on page 182 to illustrate the versatility and comprehensiveness of Tivoli Storage Manager within a BC solution.

4.1.2 What is archive and retrieve?

The archive process copies data to another storage media of which is managed by an archive server for long-term storage. The process can optionally delete the archived files from the original storage immediately or at a predefined period of time. The archive server retains the archive copy according to the policy for archive retention time. A client can retrieve an archived copy of a file when necessary.

Note: More information about the IBM archive solution is provided in 4.4, “IBM Data Retention 550” on page 206.

4.1.3 Tape backup methodologies

Various methodologies can be used to create tape backups. Often the tape technology is chosen first, then the methodology. This might force a client into service levels, recovery point objectives (RPO), and recovery time objectives (RTO), for data/applications that are less than desired. We recommend that the backup requirements and recovery objectives be determined first, followed by the methodology, and finally the technology.

The recovery point objective (RPO) and the recovery time objective (RTO) are described in the chapter “Tier levels of Business Continuity Solutions in the *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.

Normally a daily backup can be achieved, but if sub-daily backups are required along with short restore times, tape might not be suitable. Virtual Tape Server (VTS) in the mainframe and IBM Tivoli Storage Manager in the mainframe and open environments are able to use disk caches, known as storage pools, to speed up the backup and restore.

The data from the disk cache can later be migrated to tape. VTS and Tivoli Storage Manager offer strong automated solutions for such sub-daily back up requirements. Other methods combine FlashCopy and the mirroring of disks for the most recent versions of the data and tape for an asynchronous copy.

Full backups

A full backup is simply that: a complete backup of the data on every volume, respective of a set of volumes that belong together. This is the easiest way to back up a system image. However, it is not the most efficient. A different set of tapes is cycled through the system each day. The downside is that it takes more tapes than other methods. This type of backup is highly dependent on the amount of data, that must be backed up together. It is only applicable, if there is enough downtime available for the backup task to complete. A work around is the use of FlashCopy, which reduces the downtime dramatically.

The main advantage for Disaster Recovery is that you only require the last set of backup tapes to restore your entire system onto the backup hardware. For a disaster recovery solution with no connecting infrastructure between sites, this method might be usable, depending on your RPO and RTO.

Incremental backups

Incremental backups only backup the files that were created or changed since the last backup, full or incremental. The main advantage is that incremental backups are much faster and use much less tape.

The major disadvantage of incremental backups is that multiple tapes are required to restore a set of files. The files can be spread over all the tapes used since the last full backup and you might have to search several tapes to find the files you want to restore. During a Disaster Recovery, you often restore several versions of the same file as you apply the full backup followed by each incremental. In addition you might inadvertently restore files that have been deleted from your system, which you no longer require, therefore requiring you to go back and remove these unwanted files again.

Note: More sophisticated backup software such as DFSMSHsm™ on the mainframe or IBM Tivoli Storage Manager, which use a central database to manage all the backup data and tapes, overcome this disadvantage by doing regular recycle or reclamation tasks and support functions like collocation. In this case, a restore normally can take much less time than with the other methods. For DFSMSHsm recycle information, see *z/OS DFSMSHsm Storage Administration Guide*, SC35-0421. For details on IBM Tivoli Storage Manager reclamation and collocation, see the IBM Redbook, *IBM Tivoli Storage Management Concepts*, SG24-4877.

Differential backups

Differential backups are backups that include all files that were created or modified since the last *full* backup.

The advantage over full backups is that differential backups are quicker and use less media. The advantage over (traditional) incremental backups is that the restore process is more efficient. Regardless of the number of backups taken, a restore requires only the latest full backup set and the latest differential backup set.

The disadvantage of differential backups is that more and more time is required each night to perform a backup as the amount of changed data grows. Therefore more media is also required each time a differential backup is performed. Also, a full system restore requires many files to be restored more than once.

Full / differential / incremental pattern

Another way of utilizing tape is using combinations of the above mentioned methods. For example, you can take full backups once a week, differential backups in the middle of the week, and incremental backups in-between. This process reduces the number of tapes you have to manage, because you can discard your incrementals once you perform a differential backup. It also reduces the total backup window for the week. The restore is not as simple and can be very time consuming, because you (or the backup software) have to figure out which tapes must be restored. Again, a file that exists on the incremental also exists on the differential and the full, thus requiring files to be restored more than once.

Progressive backup

IBM Tivoli Storage Manager has a progressive backup feature. When applied to all file system data (non-database), this feature has the potential for significant savings in tape hardware and network bandwidth. These savings can be achieved because significantly less backup data has to be transferred over the network and stored than with traditional backup and recovery applications.

The progressive/incremental backup feature is made possible by the ability of Tivoli Storage Manager to do file-level tracking in the database and recovery log. With the progressive technique, only incremental backups are required after the first full backup is completed. In addition, when a restore is required, it is not necessary to transfer the full backup plus the differential data (a combination that often contains multiple copies of the same files), instead, the restore process transfers only the actual files required for a full restore. Other benefits of the file-level tracking are the ability to consolidate and reduce tape usage through collocation and reclamation.

For more details on the benefits discussed here, see the IBM Redbook, *IBM Tivoli Storage Management Concepts*, SG24-4877.

Hierarchical backup schemes

Sophisticated backup software, such as Tivoli Storage Manager or the hardware in an IBM Virtual Tape Server, uses a hierarchy of storage. This means, that a disk pool can be defined for initial backup with data later migrating to tape. The advantage is that the initial backup is made to faster, but more expensive media (disk), and the data is then migrated to less expensive, slower media (tape.). A restore of the most recent backups is quicker if the backup data is still on disk (see also the discussion of sub-daily backups above).

Tape rotation methods

Using a manual tape rotation scheme leads to the inefficient use of tape media, since these methods usually require more tapes. Some manual schemes have been devised to reduce the number of tapes while providing long-term backup retention. However, these are often prone to error and inefficient due to the high requirement for manual intervention.

It is very important to use storage management software that keeps track of which tapes should be in an off-site vault and which must be on-site. Tivoli Storage Manager is an excellent example of software that reduces media use while providing strong data protection.

Off-site backups

The most important aspect of tape backups in terms of Business Recovery is that the tape media must be stored at an off-site storage facility. The backup media is of no use if it is also damaged or destroyed in the event of a disaster.

Based on the tier definitions in the chapter “Tier levels of Business Continuity solutions” in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547, you can

decide which off-site backup methodology best suits your off-site backup requirements. Basically Tiers 1-4 are achieved in most cases. With solutions like the IBM VTS PtP or TS7740 Grid, Tiers 5-7 can also be achieved for tape data.

4.1.4 Tape backup and recovery software

An important aspect of tape backup and recovery is the software element. For an overview of IBM software available, see Chapter 12, “Storage management software” on page 391.

Here is a brief list of items to consider when selecting backup and recovery software:

- ▶ How are the backup methods described in 4.1.3, “Tape backup methodologies” on page 176 implemented in the software?
 - Do the methods fit with your requirements?
 - What are the consequences regarding the usage of resources (Virtual Tape Systems, tape drives, cartridges, robots)?
- ▶ Does the overall functionality fit with your requirements?
- ▶ How does the software handle the management of its backup data and tapes? Does it have a central database for this purpose?
- ▶ Does the software support multiple storage hierarchies such as disk and tape?
- ▶ Does the backup software also include features for Disaster Recovery. How many of these functions are automated in the case of a disaster?
- ▶ Make sure your backup software has an approved plug-in for your databases and application servers. Certain databases and applications have specific backup requirements that must be addressed by the software you purchase. For more details see Chapter “Databases and applications: high availability options” in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.
- ▶ An important feature of backup and recovery software is that cross-platform support is provided. Unless it is absolutely necessary, it is best to avoid multiple software solutions. This is more costly, and the complexity of recovery increases the more solutions you implement.
- ▶ Make sure your software and tape hardware are compatible and certified to work together.
- ▶ Look for automated media-rotation schemes. For security and to achieve the maximum life span of your tape medium, you must use, rotate, and retire it regularly. If you have a large number of tapes, automation of this process is vital.
- ▶ Make sure to consider both ease-of-use and installation of the backup and recovery software, since your comfort level with the software is important. Confidence is essential when backing up and restoring critical data.
- ▶ Backup software should also accurately report data that was not backed up during a process, so that it can be revisited by the automated process or manually by the administrator when the file access is freed by the user.
- ▶ What are the security features of the software?
- ▶ How are the administration tasks set up?
- ▶ Your backup and recovery software should allow for scalability. As storage requirements increase, your requirement for more tapes, and possibly backup servers, can grow. The software must allow for the addition of hardware, as well as modification to the entire backup process and topology, if necessary.
- ▶ It is crucial for large companies to have software that can support the operation of multiple tape drives simultaneously from the same server.

Many backup and recovery software packages offer hierarchical storage management of data. These packages are designed to minimize storage costs while optimizing performance. They achieve this by combining multiple storage media such as magnetic disk, optical disk, and tape into a single logical unit, and transparently migrating data between media based on frequency of access.

They temporarily migrate data that has not been accessed for some time to a secondary tape storage medium. Many solutions allow administrators to manage data and servers on a case by case basis allowing them to set rules for the migration scheme.

4.2 IBM DB2 for z/OS backup and recovery options

This section discusses how to back up and restore DB2 for z/OS.

4.2.1 SET LOG SUSPEND

As discussed in Chapter 6 “Planning for Business Continuity in a heterogeneous IT environment” in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547, there are a number of methods available for performing backup tasks. The essence of traditional tape based recovery is moving data to tape on some set schedule and then moving those tapes to another location for storage and, eventually, for recovery (though the locations for storage and recovery might be in separate places). Creating these tapes still requires planning, though.

It is critical to ensure that all of the data to be used in the recovery is consistent to a set point in time. In order to make this so, at some point the DB2 related volumes (including the DB2 Catalog, DB2 Directory, User Data, BSDS, active logs, ICF catalogs, SMP data sets, and program libraries) all have to be at a single point in time and moved to tape. Doing so online could involve a significant impact to production applications. In order to avoid using the active data for this purpose, one option is to use FlashCopy.

Using FlashCopy gives us a consistent and restartable copy of data that can be used for the backup process, rather than using the live data. Using live data in for backup would still require some means of maintaining a single point of consistency and could involve a lengthy outage.

In order to avoid this long outage, but still maintain a consistent point for recovery in the FlashCopy, we use the **SET LOG SUSPEND** command. This stops all of the updates within a DB2 subsystem. This interruption only has to be in place for as long as it takes to establish the FlashCopy relationship between volumes which should be somewhere between seconds to a few minutes. The data is not consistent when the **SET LOG SUSPEND** is issued, but becomes so at a subsequent DB2 Restart.

After these bitmaps are complete (through the Establish process), a **SET LOG RESUME** command can be issued, allowing the updates to the data base to continue.

DB2 V8 Backup System utility

This utility, introduced in DB2 V8, allows the full back up described in the previous section. After the data has been defined to one SMS copy pool and the active logs/BSDS to another copy pool, this utility automates the process you had to perform manually: submitting and automatically parallelizing the volume FlashCopies, ending the job when all the establish phases are complete, and doing all this without taking the DB2 log write latch. In DB2 V8 and z/OS V1.5, the SMS copy pools can only be directed to DASD. It is the user's responsibility to dump the copy pools to tape and take them to offsite storage.

DB2 V9 Backup System enhancements

With z/OS V1.8 and DFSMS capabilities, DB2 V9 can direct the Backup System utility to dump the copy pools directly to tape for transport. When Backup System is performed with copy pools on DASD, a second Backup System can be submitted at a time more convenient for the user to dump the DASD copy pools to tape. This introduces ease of use to the V8 Backup System, as the V8 Backup System did in turn for **-SET LOG SUSPEND**.

4.2.2 FlashCopy Manager

Another option for establishing the FlashCopy relationships across the entire environment, but without the requirement of a **SET LOG SUSPEND** command is the FlashCopy Manager offering from STG Lab Services. The FlashCopy Manager code issues a low level “Freeze and Run” command across the storage environment, preserving consistency and issuing the commands to establish FlashCopy immediately.

More information about FlashCopy Manager can be found in 3.6.1, “FlashCopy Manager overview” on page 167. See the note in the Restore section in 4.2.3, “Backing up and restoring data in DB2 for z/OS” for restrictions concerning applying additional logs when you use this method.

4.2.3 Backing up and restoring data in DB2 for z/OS

Now let us consider the actual backup and restore processes.

Backup

The tape usage itself is handled by DFSMSdss™ rather than tools in DB2 itself. These tools are first used to dump the FlashCopy data to tape through the DUMP command. The tapes can then be shipped to the recovery site once per week (dependent on the recovery guidelines of the specific environment). Additional log data can then be sent more frequently using Automatic Tape Libraries and vaulting procedures if such equipment and procedures are available.

Restore

The process of moving data from tape back to disk is, likewise, handled through DFSMSdss. In this case it occurs through the **RESTORE** command. If desired, it is possible to use **RESTORE SYSTEM LOGONLY** in order to recover to a specified point in time.

By using this command, we can use log archives, which might have been shipped more frequently. The **RESTORE SYSTEM LOGONLY** command then uses the archive logs to apply additional transactions beyond that which is stored in the database.

Note: **RESTORE SYSTEM LOGONLY** cannot be performed with the FlashCopy Manager solution. This is because the starting point for the Log Apply phase of recovery is only initialized by **- SET LOG SUSPEND** or by the Backup System utility.

For more information about using DB2 UDB for z/OS in a backup and restore segmentation, see *Disaster Recovery with DB2 UDB for z/OS*, SG24-6370.

4.3 IBM Tivoli Storage Manager overview

IBM Tivoli Storage Manager protects an organization's data from failures and other errors. By managing backup, archive, space management and bare-metal restore data, as well as compliance and disaster-recovery data in a hierarchy of offline storage, the Tivoli Storage Manager family provides centralized, automated data protection. Thus, Tivoli Storage Manager helps reduce the risks associated with data loss while also helping to reduce complexity, manage costs, and address compliance with regulatory data retention requirements.

Since it is designed to protect a company's important business information and data in case of disaster, the Tivoli Storage Manager server should be one of the main production systems that is available and ready to run for recovery of business data and applications.

Tivoli Storage Manager provides industry-leading encryption support through integrated key management and full support for the inbuilt encryption capability of the IBM System Storage TS1120 Tape Drive.

This section provides the Tivoli Storage Manager solutions in terms of Business Continuity and Disaster Recovery. There are six solutions to achieve each BC tier:

- ▶ BC Tier 1: IBM Tivoli Storage Manager manual off-site vaulting
- ▶ BC Tier 2: IBM Tivoli Storage Manager manual off-site vaulting with a hotsite
- ▶ BC Tier 3: IBM Tivoli Storage Manager electronic vaulting
- ▶ BC Tier 4: IBM Tivoli Storage Manager with SAN attached duplicates
- ▶ BC Tier 5: IBM Tivoli Storage Manager clustering
- ▶ BC Tier 6: IBM Tivoli Storage Manager running in a duplicate site

This section also covers these additional Tivoli Storage Manager products for integrated solution capabilities:

- ▶ IBM Tivoli Storage Manager for Copy Services - Data Protection for Exchange
- ▶ IBM Tivoli Storage Manager for Advanced Copy Services, with the following modules:
 - Data Protection for IBM Disk Storage and SAN Volume Controller for mySAP™ with DB2
 - Data Protection for IBM Disk Storage and SAN Volume Controller for mySAP with Oracle®
 - Data Protection for IBM Disk Storage and SAN Volume Controller for Oracle
 - DB2 UDB Integration Module and Hardware Devices Snapshot Integration Module
 - Data Protection for ESS for Oracle
 - Data Protection for ESS for mySAP Oracle
 - Data Protection for ESS for mySAP DB2 UDB

4.3.1 IBM Tivoli Storage Manager solutions overview

These solutions provide protection for enterprise business systems.

Tier level and positioning within the System Storage Resiliency Portfolio

IBM Tivoli Storage Manager solutions support Business Continuity from Tier 1 to Tier 6, as shown in Figure 4-1. These solutions achieve each tier by using hardware, software and autonomic solutions.

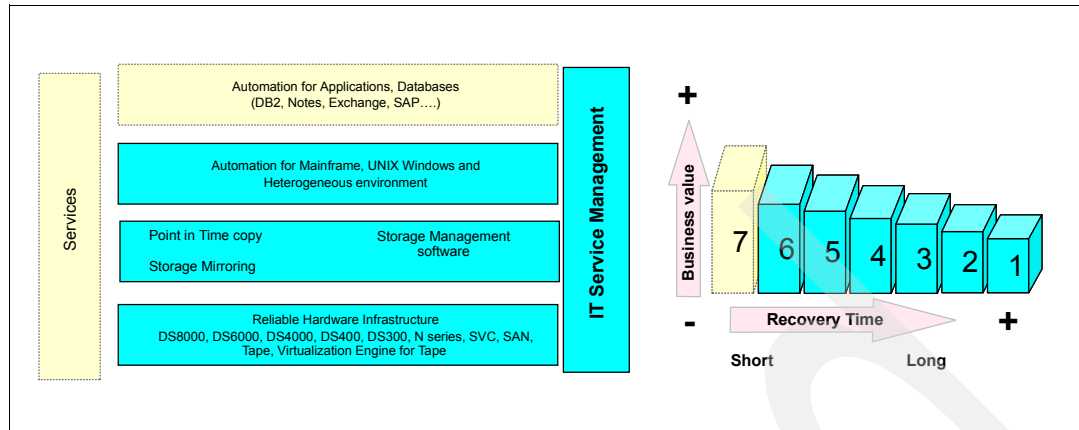


Figure 4-1 Tier level and positioning within the Resiliency Family

Solution description

These solutions enable IBM Tivoli Storage Manager system to achieve Business Continuity for Tier 1 to Tier 6. The solutions provide the ability to minimize the Recovery Time Objective (RTO) and the Recovery Point Objective (RPO) for the client's Tivoli Storage Manager system.

From BC Tier 1 to Tier 3, the Tivoli Storage Manager BC solutions use features such as Disaster Recovery Manager and server-to-server communication protocol to support tape vaulting and electronic vaulting to an off-site location.

From BC Tier 4 to Tier 6, data storage replication and clustering service are implemented on Tivoli Storage Manager systems. Integration with clustering technology comes into play. Tivoli Storage Manager systems have the ability to provide high availability and rapid data backup and restore service for your enterprise business systems.

Solution highlights

The highlights of these solutions are:

- ▶ Continuity of backup and recovery service, meet Service Level Agreement commitments
- ▶ Reduced risk of downtime of Tivoli Storage Manager system
- ▶ Increased business resiliency and maintaining competitiveness
- ▶ Minimized Recovery Time Objective (RTO) of Tivoli Storage Manager system
- ▶ Minimized Recovery Point Objective (RPO) of Tivoli Storage Manager system

Solution components

The components of these solutions include:

- ▶ IBM Tivoli Storage Manager server
- ▶ IBM Tivoli Storage Manager product features and functions:
 - IBM Tivoli Storage Manager Disaster Recovery Manager
 - IBM Tivoli Storage Manager server-to-server communication
 - IBM Tivoli Storage Manager for Copy Services - Data Protection for Exchange
 - IBM Tivoli Storage Manager for Advanced Copy Services
- ▶ IBM System Storage DS6000, DS8000, SAN Volume Controller
- ▶ IBM High Availability Cluster Multi-Processing (HACMP)
- ▶ Microsoft Cluster Server (MSCS)

Additional information

For additional information about these solutions, you can:

- ▶ Contact your IBM representative.
- ▶ See the IBM Redbooks, *IBM Tivoli Storage Management Concepts*, SG24-4877, and *Disaster Recovery Strategies with Tivoli Storage Management*, SG24-6844.
- ▶ Visit this Web site:

<http://www.ibm.com/software/tivoli/products/storage-mgr/>

4.3.2 IBM Tivoli Storage Manager solutions in detail

This section provides information about solutions and strategies to enable the Tivoli Storage Manager system to achieve the specific BC tiers and meet BC business requirements.

Tier 0: No off-site data

BC Tier 0 is defined as a single site data center environment having no requirements to back up data for implement a Business Continuity Plan at a different site. See Figure 4-2 for an illustration.

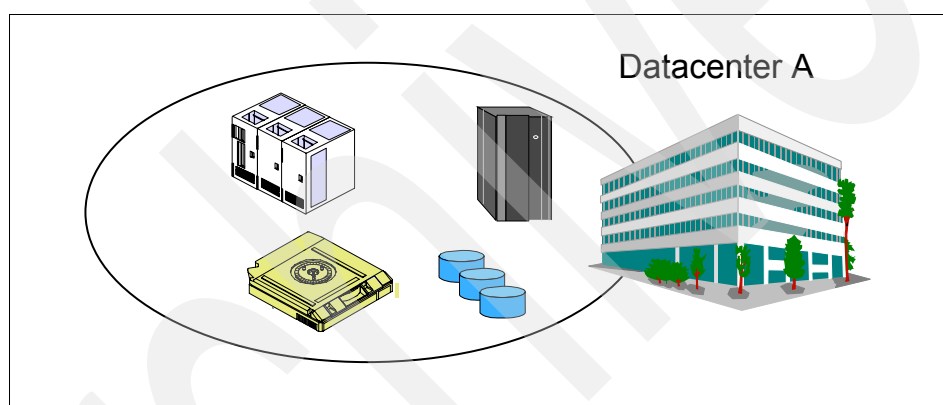


Figure 4-2 BC Tier 0 - Tivoli Storage Manager, no off-site data

For this BC tier, there is no saved information, no documentation, no backup hardware, and no contingency plan. There is therefore no Business Continuity capability at all. Some businesses still reside in this tier. While they might actively make backups of their data, these backups are left onsite in the same computer room, or are only infrequently taken offsite due to lack of a rigorous vaulting procedure. A data center residing on this tier is exposed to a disaster from which they might never recover their business data.

Typical length of time for recovery

The typical length of time for recovery is unpredictable. In many cases, complete recovery of applications, systems, and data is never achieved.

Strategies and operations

The strategies and operations for this solution are as follows:

- ▶ Normal Tivoli Storage Manager based onsite backups
- ▶ No off-site strategy
- ▶ No ability to recover from a site disaster except to rebuild the environment

Strategies and operations description

Because BC Tier 0 contains no off-site strategy, the Tivoli Storage Manager system provides backup and restore service for the production site only.

Tier 1: Tivoli Storage Manager backups with manual off-site vaulting

BC Tier 1 is defined as having a Business Continuity Plan (see Chapter 3, “Business Continuity planning, processes, and execution” in the *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547), where data is backed up and stored to a centralized location. Copies of these backups are then manually taken off-site, as shown in Figure 4-3. Some recovery requirements have been determined, including application and business processes. This environment might also have established a recovery platform, although it does not have a site at which to restore its data, nor the necessary hardware on which to restore the data, for example, compatible tape devices.

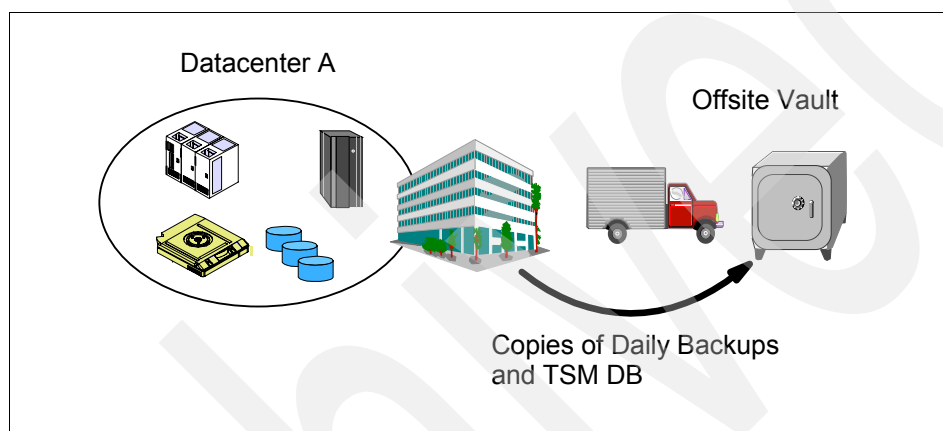


Figure 4-3 BC Tier 1 - Tivoli Storage Manager manual off-site vaulting

Because vaulting and retrieval of data is typically handled by couriers, this tier is described as the Pickup Truck Access Method (PTAM). PTAM is a method used by many sites, as this is a relatively inexpensive option. It can be difficult to manage and difficult to know exactly where the data is at any point. There is probably only selectively saved data. Certain requirements have been determined and documented in a contingency plan.

Recovery depends on when hardware can be supplied, or when a building for the new infrastructure can be located and prepared.

While some companies reside on this tier and are seemingly capable of recovering in the event of a disaster, one factor that is sometimes overlooked is the recovery time objective (RTO). For example, while it might be possible to eventually recover data, it might take several days or weeks. An outage of business data for this period of time is likely to have an impact on business operations that lasts several months or even years (if not permanently).

Typical length of time for recovery

With this solution, the typical length of time for recovery is normally more than a week.

Strategies and operations

The strategies and operations for this solution are as follows:

- ▶ Use of Disaster Recovery Manager (DRM) to automate the Tivoli Storage Manager server recovery process and to manage off-site volumes
- ▶ Recommended: Creation of a Business Continuity Plan and careful management of off-site volumes

- ▶ Manual storage pool vaulting of copies of data with Tivoli Storage Manager server environment
- ▶ A strategy that must include manual vaulting of copies of Tivoli Storage Manager database, volume history information, device configuration information, Business Continuity Plan file and copy pools for storage at an off-site location

Strategies and operations description

BC Tier 1 requires off-site vaulting. The following Tivoli Storage Manager components are sent to off-site:

- ▶ Tivoli Storage Manager storage pool copies
- ▶ Tivoli Storage Manager database backup
- ▶ Tivoli Storage Manager configuration files
 - Volume history file
 - Device configuration file

In this tier, we recommend that clients use IBM Tivoli Storage Manager Disaster Recovery Manager (DRM) to automate off-site volume management and the Tivoli Storage Manager recovery process. The solution diagram is shown in Figure 4-4.

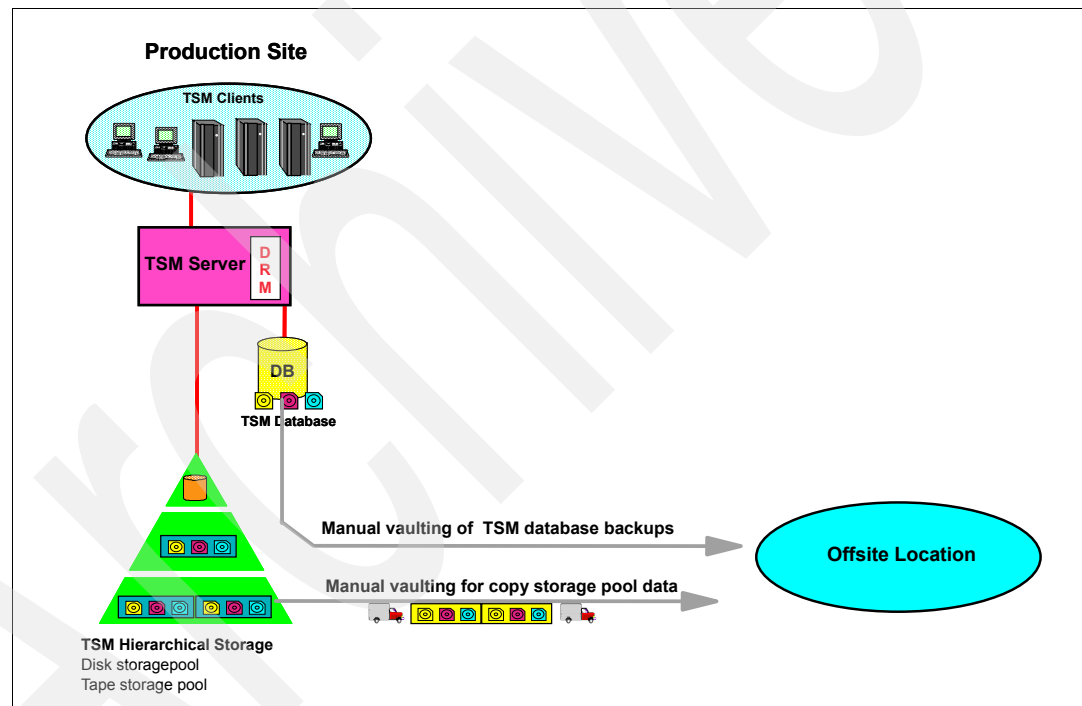


Figure 4-4 BC Tier 1 - Tivoli Storage Manager manual off-site vaulting - solution diagram

When recovery is required, Tivoli Storage Manager system can be recovered by using the off-site tape volumes, off-site database backup volumes, off-site configuration files, and the Business Continuity Plan file with the Tivoli Storage Manager DRM process. Then the Tivoli Storage Manager clients can restore their backup data from the new Tivoli Storage Manager system.

Tier 2: Tivoli Storage Manager off-site manual vaulting with a hotsite

BC Tier 2 encompasses all the requirements of BC Tier 1 (off-site vaulting and recovery planning), plus it includes a hotsite. The hotsite has sufficient hardware and a network infrastructure able to support the installation's critical processing requirements. Processing is considered critical if it must be supported on hardware existing at the time of the disaster. As shown in Figure 4-5, backups are being taken, copies of these are created, and the copies are being stored at an off-site storage facility. There is also a hotsite available, and the copy of the backups can be manually transported there from the off-site storage facility in the event of a disaster.

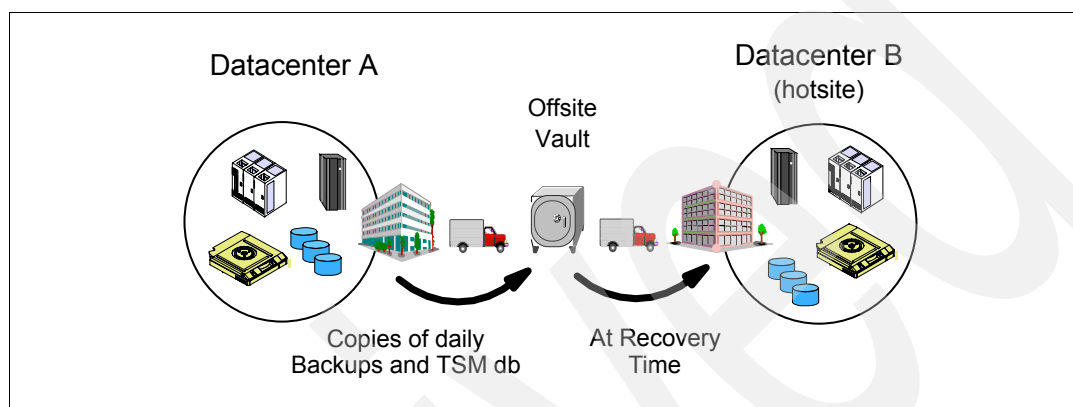


Figure 4-5 BC Tier 2 - Tivoli Storage Manager off-site manual vaulting with a hotsite, copies of daily backup

Tier 2 installations rely on a courier (PTAM) to get data to an off-site storage facility. In the event of a disaster, the data at the off-site storage facility is moved to the hotsite and restored on to the backup hardware provided. Moving to a hotsite increases the cost but reduces the recovery time significantly. The key to the hotsite is that appropriate hardware to recover the data (for example, a compatible tape device) is present and operational.

Typical length of time for recovery

With this solution, the typical length of time for recovery is normally more than a day.

Strategies and operations

The strategies and operations for this solution are as follows:

- ▶ Use of DRM to automate the Tivoli Storage Manager server recovery process and to manage off-site volumes
- ▶ Manual vaulting of copies of Tivoli Storage Manager server's database backup and storage pool copies
- ▶ Tivoli Storage Manager server installed at both locations
- ▶ Vaulting Tivoli Storage Manager database backup, volume history information, device configuration information, Business Continuity Plan file and storage pool copies at an off-site location
- ▶ Consideration given to using Tivoli Storage Manager server-to-server communications to enable enterprise configuration, enterprise event logging, event monitoring, and command routing

Strategies and operations description

BC Tier 2 requires off-site vaulting with a hot site; the implementation of this tier is based on Tier 1 implementation (see Figure 4-3). In this tier, an additional IBM Tivoli Storage Manager server should be installed at the hot site. We can use Tivoli Storage Manager server-to-server communication to enable enterprise configuration, enterprise event logging, event monitoring, and command routing. The solution is shown in Figure 4-6.

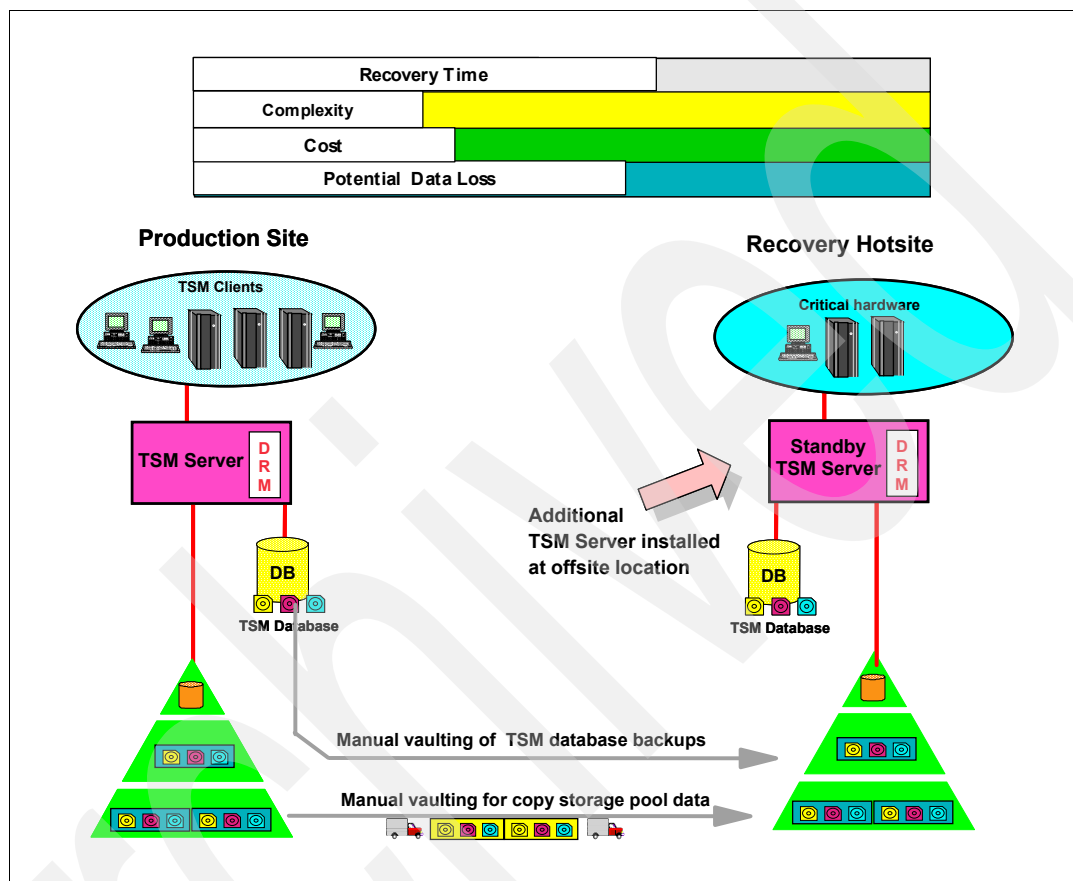


Figure 4-6 BC Tier 2 - Tivoli Storage Manager off-site manual vaulting with a hot site - solution diagram

When the recovery process is required, the Tivoli Storage Manager recovery process is run by using the Disaster Recovery Manager on the prepared Tivoli Storage Manager server. The recovery process uses the off-site tape volumes, off-site database backup volumes, off-site configuration files, and Business Continuity Plan file with the Tivoli Storage Manager DRM process. Then the Tivoli Storage Manager clients can restore their data from the new Tivoli Storage Manager system.

Tier 3: Tivoli Storage Manager electronic vaulting

BC Tier 3 encompasses all of the components of BC Tier 2 (off-site backups, Business Continuity Plan, hot site). In addition, it supports electronic vaulting of the backup of the Tivoli Storage Manager database and storage pools. Electronic vaulting consists of electronically transmitting the backups to a secure facility, thus moving business-critical data off-site faster. This is accomplished via Tivoli Storage Manager's virtual vault capability. The receiving hardware must be physically separated from the primary site. As shown in Figure 4-7, backups of the Tivoli Storage Manager database and the entire set or a subset of the storage pools can also optionally be made to physical media and manually moved to an off-site storage facility.

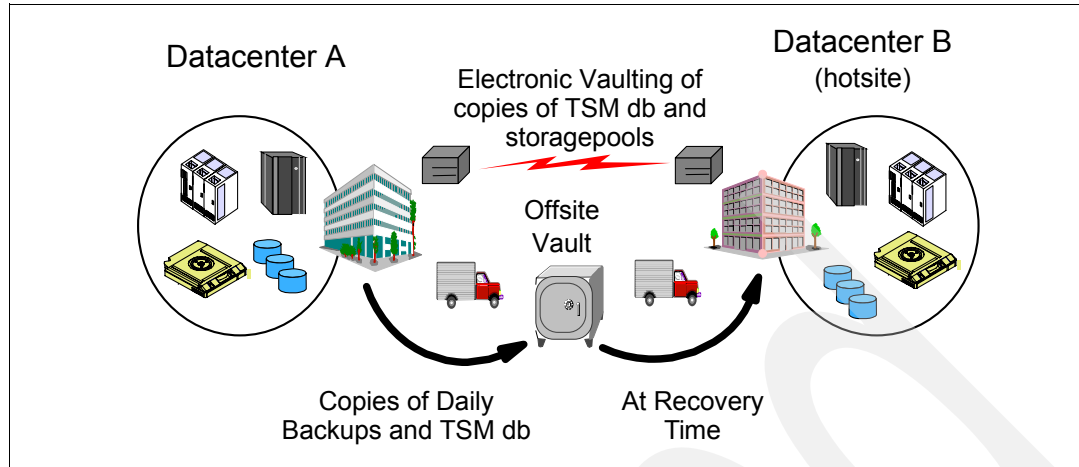


Figure 4-7 BC Tier 3 - Tivoli Storage Manager with electronic vaulting

The hot site is kept running permanently, thereby increasing the cost. As the critical data is already being stored at the hot site, the recovery time is once again significantly reduced. Often, the hot site is a second data center operated by the same firm or a Storage Service Provider.

Typical length of time for recovery

With this solution, the typical length of time for recovery is normally about one day.

Strategies and operations

The strategies and operations for this solution are as follows:

- ▶ Tivoli Storage Manager virtual volumes over TCP/IP for storing Tivoli Storage Manager entities (Tivoli Storage Manager database backups, primary and copy storage pool backups, DRM plan files) on remote target server
- ▶ Tivoli Storage Manager server installed at both locations
- ▶ Use of DRM to automate the Tivoli Storage Manager server recovery process and to manage off-site volumes.
- ▶ Use of Tivoli Storage Manager server-to-server communication to enable enterprise management features.
- ▶ Optionally, manually making vault copies of Tivoli Storage Manager database backup, volume history information, device configuration information, Business Continuity Plan file and storage pools copies at an off-site location

Strategies and operations description

Tivoli Storage Manager lets a server (a *source* server) store the results of database backups, export operations, storage pool operations, and a DRM plan on another server (a *target* server). The data is stored as *virtual* volumes, which appear to be sequential media volumes on the source server but which are actually stored as archive files on a target server.

Virtual volumes can be:

- ▶ Database backups
- ▶ Storage pool backups
- ▶ Data that is backed up, archived, or space managed from client nodes
- ▶ Client data migrated from storage pools on the source server
- ▶ Any data that can be moved by EXPORT and IMPORT commands
- ▶ DRM plan files

Clients can decide to send all of this information or just a subset of it via FTP to the target server. This decision in part is based on the amount of bandwidth available. The data they choose not to send over the network can be taken off-site manually. See Figure 4-8.

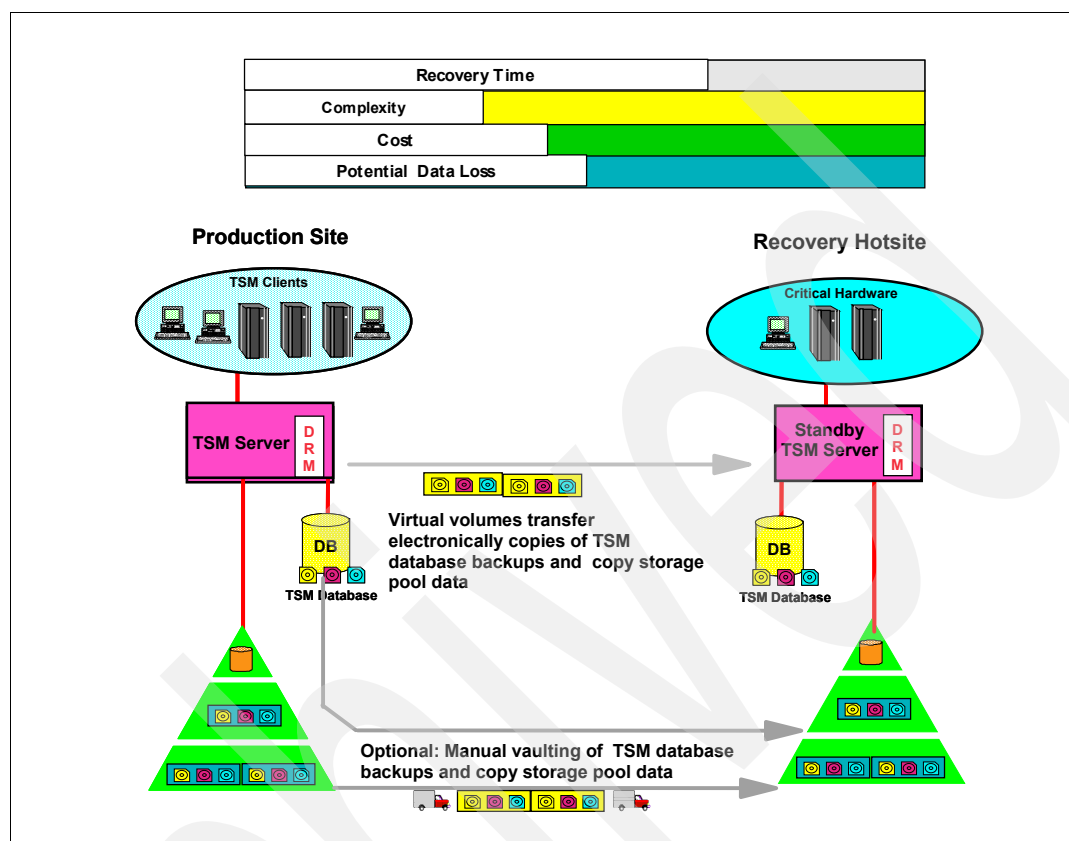


Figure 4-8 BC Tier 3 - Tivoli Storage Manager with electronic vaulting - solution diagram

BC Tier 4: Tivoli Storage Manager with SAN attached duplicates

BC Tier 4 is defined by two data centers with SAN attached storage to keep mirrors of the Tivoli Storage Manager database and log. Separate Tivoli Storage Manager storage pools are located on storage residing in the two locations. SAN technology is critical to facilitate the Tivoli Storage Manager database and log mirroring and Tivoli Storage Manager simultaneous data writes occurring to local and remote devices. See Figure 4-9. The hotsite also has backups of the Tivoli Storage Manager database storage pools stored either as virtual volumes or physical backup tapes to protect against database corruption and ensure the ability to restore the Tivoli Storage Manager server to a different point-in-time.

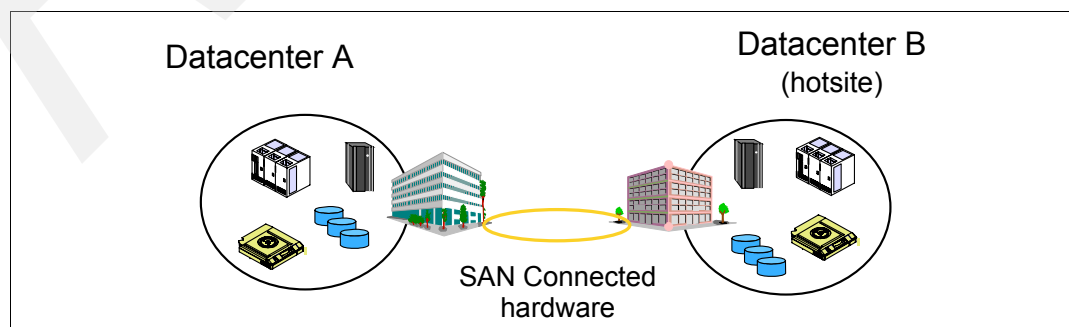


Figure 4-9 BC Tier 4 - Tivoli Storage Manager with duplicate SAN attached hardware

Typical length of time for recovery

With this solution, the typical length of time for recovery is usually up two to four hours.

Strategies and operations

The strategies and operations for this solution are as follows:

- ▶ Tivoli Storage Manager database and log volumes mirrored sequentially on SAN hardware at a different location from the primary volumes
- ▶ Using Tivoli Storage Manager's simultaneous copy storage pool write to copy data to a Tivoli Storage Manager storage pool located on SAN hardware that is located at a different locations from the primary storage pool volume
- ▶ Tivoli Storage Manager server code installed at the second location but not running
- ▶ Using DRM to automate the Tivoli Storage Manager server recover process and to manage off-site volumes
- ▶ Using Tivoli Storage Manager server-to-server communications to enable enterprise management features
- ▶ Vault copies of Tivoli Storage Manager database backups and storage pool copies, either done manually (BC Tier 2) or with virtual volumes (BC Tier 3) daily

Strategies and operations description

BC Tier 4 requires a SAN infrastructure to allow mirroring of the Tivoli Storage Manager database and log (using Tivoli Storage Manager's mirroring capability) to off-site storage. The SAN attached storage pool is written to concurrently with the local storage pool, using Tivoli Storage Manager's simultaneous copy storage pool write. The Tivoli Storage Manager database and storage pools are still backed up daily, and either electronically or manually vaulted off-site to protect against corruption to the Tivoli Storage Manager database, and to allow for the server to be restored to a previous time. See Figure 4-10.

Another Tivoli Storage Manager should be installed at the off-site location, with access to the database and log mirror and to the copy storage pools. However, this Tivoli Storage Manager server is not started, unless the primary Tivoli Storage Manager server fails. The two Tivoli Storage Manager servers should never be run at the same time, as this might corrupt the database and log.

When using the Tivoli Storage Manager's simultaneous copy storage pool write, the user should turn on the **copy continue = no** option. The simultaneous write creates a copy of the data to the primary storage pool and to a secondary storage pool at the initial time of write. If the writes occur at different speeds, the performance is gated by the slower write.

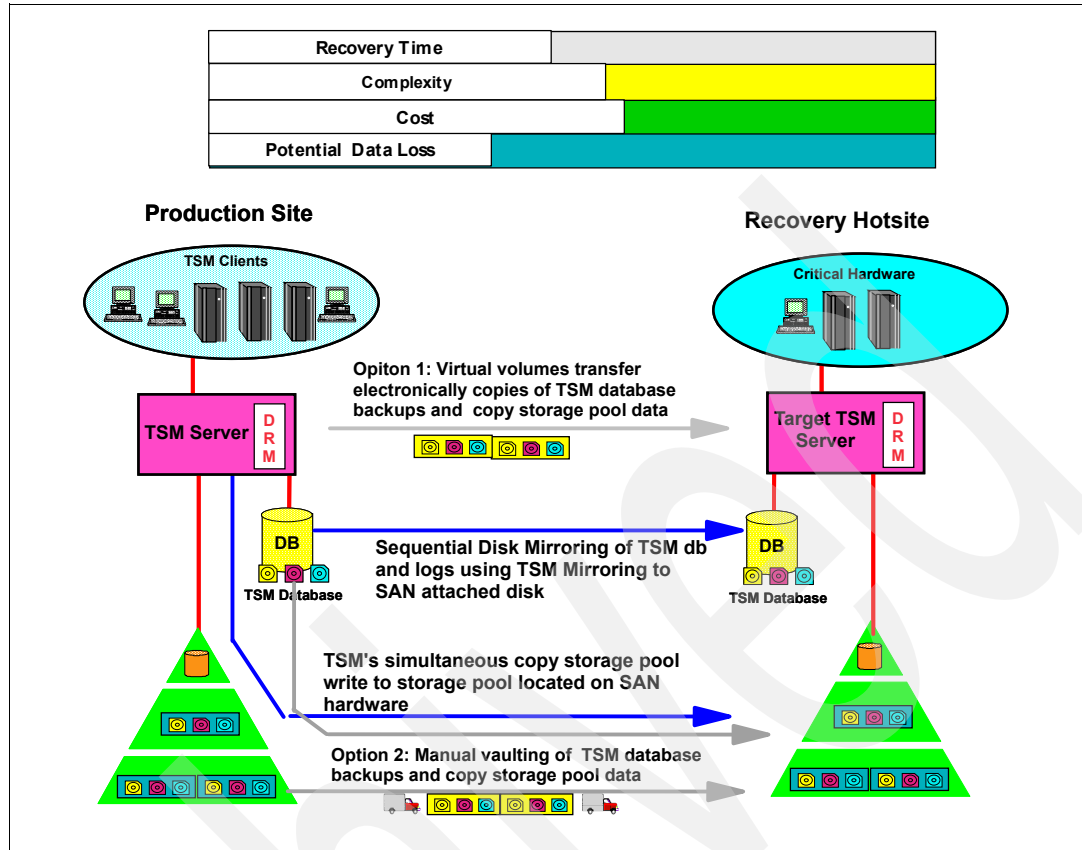


Figure 4-10 BC Tier 4 - Tivoli Storage Manager with SAN attached database, log mirrors, and copy storage pools

BC Tier 5: IBM Tivoli Storage Manager clustering

BC Tier 5 is defined by two data centers utilizing clustering technology to provide automated failover and failback capabilities for key applications. The key to this setup is clustering applications like HACMP and MSCS to facilitate the Tivoli Storage Manager server switching over to another server. The hotsite also has Tivoli Storage Manager database backups and copies of the Tivoli Storage Manager storage pools stored either as virtual volumes or physical backup tapes to protect against database corruption and ensure the ability to restore the Tivoli Storage Manager server back to a different point-in-time. See Figure 4-11.

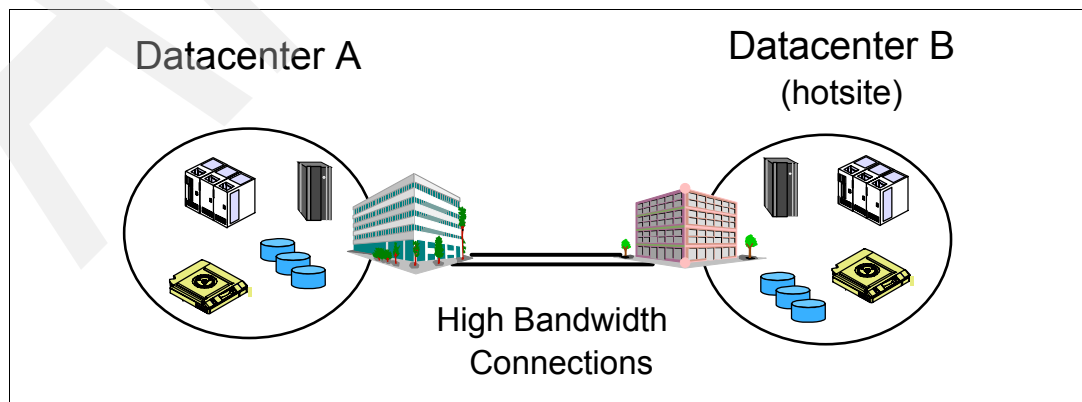


Figure 4-11 BC Tier 5 - Tivoli Storage Manager clustering

Typical length of time for recovery

With this solution, the typical length of time for recovery is normally a few minutes.

Strategies and operations

The strategies and operations for this solution are as follows:

- ▶ Utilizing HACMP or MSCS to cluster the Tivoli Storage Manager server and clients
- ▶ Installing Tivoli Storage Manager server at the second location in either an active-active, or active-passive cluster
- ▶ Utilizing Tivoli Storage Manager's SCSI device failover for SCSI tape libraries, or consider using SAN attached storage pools
- ▶ Using DRM to automate the Tivoli Storage Manager server recover process and to manage off-site volumes
- ▶ Using Tivoli Storage Manager server-to-server communications to enable enterprise management features
- ▶ Vault daily copies of Tivoli Storage Manager database backups and storage pool copies, done either manually (BC Tier 2) or with virtual volumes (BC Tier 3)

Strategies and operations description

Careful planning and testing is required when setting up a clustering environment. After the cluster environment is created, Tivoli Storage Manager can utilize the technology to perform automatic failovers for the Tivoli Storage Manager server and clients (including HSM and Storage Agents). See Figure 4-12. Clustering is covered in great detail in the Tivoli Storage Manager manuals and in the IBM Redbook, *IBM Tivoli Storage Manager in a Clustered Environment*, SG24-6679.

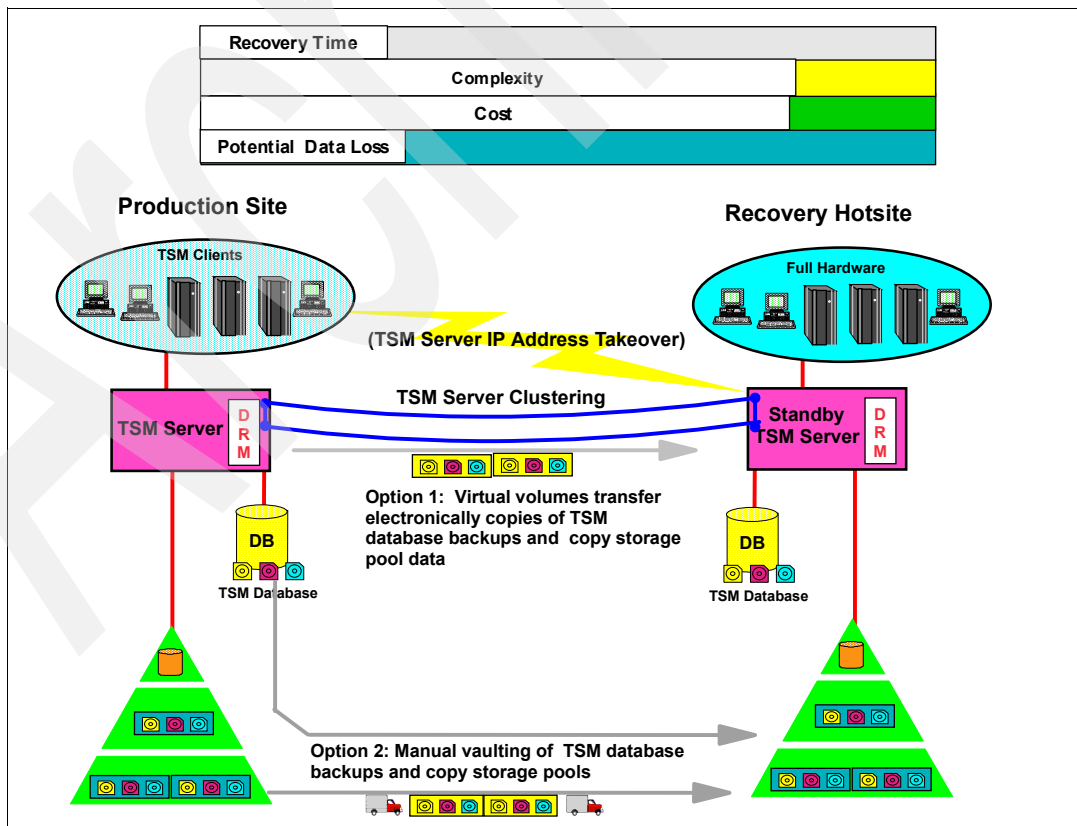


Figure 4-12 BC Tier 5 - Clustered Tivoli Storage Manager servers and clients

BC Tier 6: IBM Tivoli Storage Manager running in a duplicate site

BC Tier 6 encompasses zero data loss and immediate, automatic transfer to the secondary platform. The two sites are fully synchronized via a high-bandwidth connection between the primary site and the hotsite. The two systems are advanced coupled, allowing automated switchover from one site to the other when required. See Figure 4-13.

In this tier there are two independent Tivoli Storage Manager setups. One is dedicated to backing up the data on the primary site, the other is dedicated to backing up the data on the other site. Copies of each location's Tivoli Storage Manager database and storage pools have to be taken off-site daily either through virtual volumes or manually.

This is the most expensive BC solution as it requires duplicating at both sites all of the hardware and applications, and then keeping the data synced between the sites. However, it also offers the speediest recovery by far.

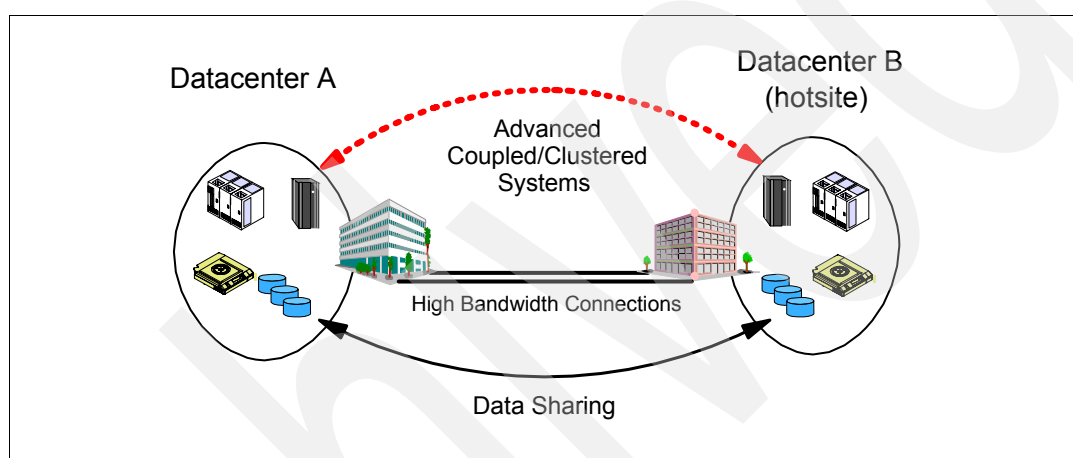


Figure 4-13 BC Tier 6 - Dual production Tivoli Storage Manager servers running in a zero data loss environment

Typical length of time for recovery

With this solution the typical length of time for recovery is normally almost instantaneous.

Strategies and operations

The strategies and operations for this solution are as follows:

- ▶ Independent Tivoli Storage Manager servers installed at both locations and backing up that location's data.
- ▶ Tivoli Storage Manager virtual volumes over TCP/IP connection to allow storage of Tivoli Storage Manager entities (Tivoli Storage Manager database backups, primary and copy storage pool backups, DRM plan files) on remote target server.
- ▶ Using DRM to automate the Tivoli Storage Manager server recover process and to manage off-site volumes.
- ▶ Using Tivoli Storage Manager server-to-server communications to enable enterprise management features.
- ▶ Optionally manually vault copies of Tivoli Storage Manager database backup, volume history information, device configuration information, Business Continuity Plan file and storage pools copies at an off-site location.

Strategies and operations description

The data at the primary and off-site location is fully synchronized utilizing a high-bandwidth connection. The two systems are advanced coupled, allow an automated switchover from one site to the other when required. A Tivoli Storage Manager setup is, however, installed at each location and run independently creating dual production systems. Each Tivoli Storage Manager server is protected using BC Tier 2 or BC Tier 3 technology. See Figure 4-14 on page 195.

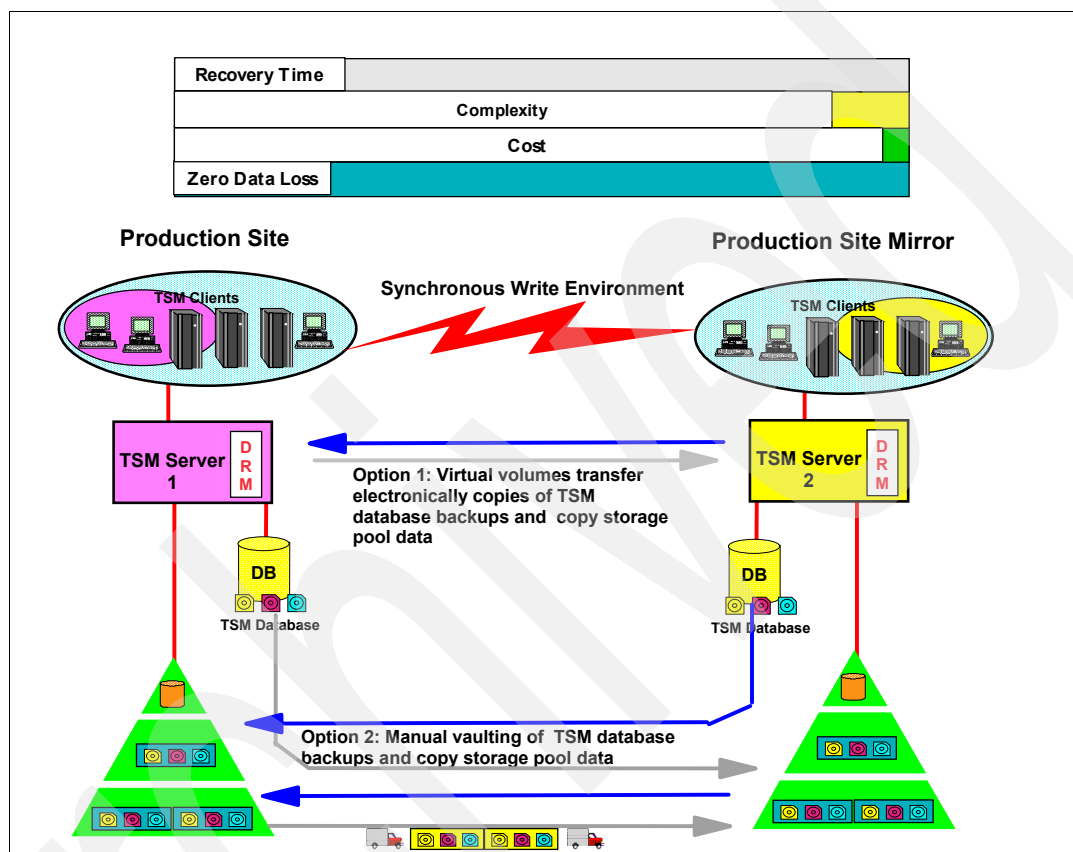


Figure 4-14 BC Tier 6 - Dual production Tivoli Storage Manager servers running in a zero data loss environment

4.3.3 Tivoli Storage Manager for Copy Services: Data Protection for Exchange

These days, e-mail systems play a key role in the business success. The business is severely impacted if the e-mail service is down even the rest of the production services are up. Consequently, keeping the e-mail servers available has become a critical business concern.

With these constraints, there are some requirements that have to be addressed, such as:

- ▶ Fast recovery
- ▶ Fast backups — “Zero Impact” high-performance backups
- ▶ Intelligent management (of these backups)

Addressing these demands, IBM Tivoli Storage Manager for Copy Services provides an enhanced backup and recovery solution which integrates VSS based snapshot capabilities with Tivoli Storage Manager for Mail — specifically the Data Protection for Microsoft Exchange component.

The foregoing seamless integration support can be deployed to the IBM System Storage DS6000, DS8000, SVC, DS4000, and N series, as well as any storage that is VSS-compliant.

Tivoli Storage Manager for Copy Services is a BC Tier 4 solution (see Figure 4-15).

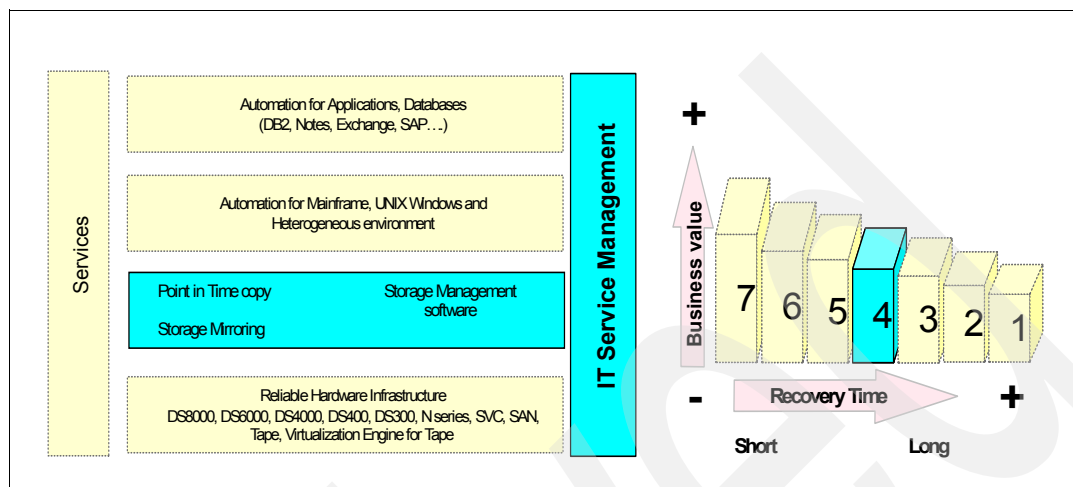


Figure 4-15 BC Tier 4 graph for Tivoli Storage Manager for Copy Services

Tivoli Storage Manager for Copy Services offers new options to implement highly efficient backup, while eliminating the backup-related impact on the performance of the Exchange production server.

Note: Detailed information about Tivoli Storage Manager for Copy Services, can be found in *Using IBM Tivoli Storage Manager to Back Up Microsoft Exchange with VSS*, SG24-7373.

Data Protection for Exchange

Data Protection for Exchange performs online backups and restores of Microsoft Exchange Server databases (standalone or MSCS clustered) to Tivoli Storage Manager storage.

When Tivoli Storage Manager for Copy Services is installed together with Data Protection for Exchange, you can also perform *online snapshot backups* to local *shadow* (disk) volumes, using VSS. These snapshot backups can be retained on the local shadow volumes, and also can be backed up to Tivoli Storage Manager server storage. To reduce the overhead of backup operations on the Exchange server, you can choose to have the backup to Tivoli Storage Manager performed by another server with access to the shadow volumes — this is referred to as an offloaded backup. If a restore is required, you can choose to restore from either local snapshot volumes, or from Tivoli Storage Manager server storage.

Features of Tivoli Storage Manager for Copy Services

Using Tivoli Storage Manager for Copy Services, you can:

- ▶ Back up Exchange Server 2003 databases (running on Windows Server® 2003) using Microsoft Volume Shadow Copy Service (VSS) technology.
- ▶ Perform a VSS backup to the Tivoli Storage Manager server from an alternate system instead of the production system (offloaded backup).
- ▶ Restore VSS backups that reside on local shadow volumes using file-level copy mechanisms.

- ▶ Restore VSS backups that reside on local shadow volumes using hardware-assisted volume-level copy mechanisms (Instant Restore).
- ▶ Utilize Tivoli Storage Manager policy-based management of VSS snapshot backups.
- ▶ Have a single GUI for performing VSS and non-VSS backup, restore, and query operations.
- ▶ Have a single command line interface for performing VSS and non-VSS backup, restore, and query operations.

Figure 4-16 summarizes the components of a Tivoli Storage Manager for Copy Services with Data Protection for Exchange solution providing VSS backup restore services.

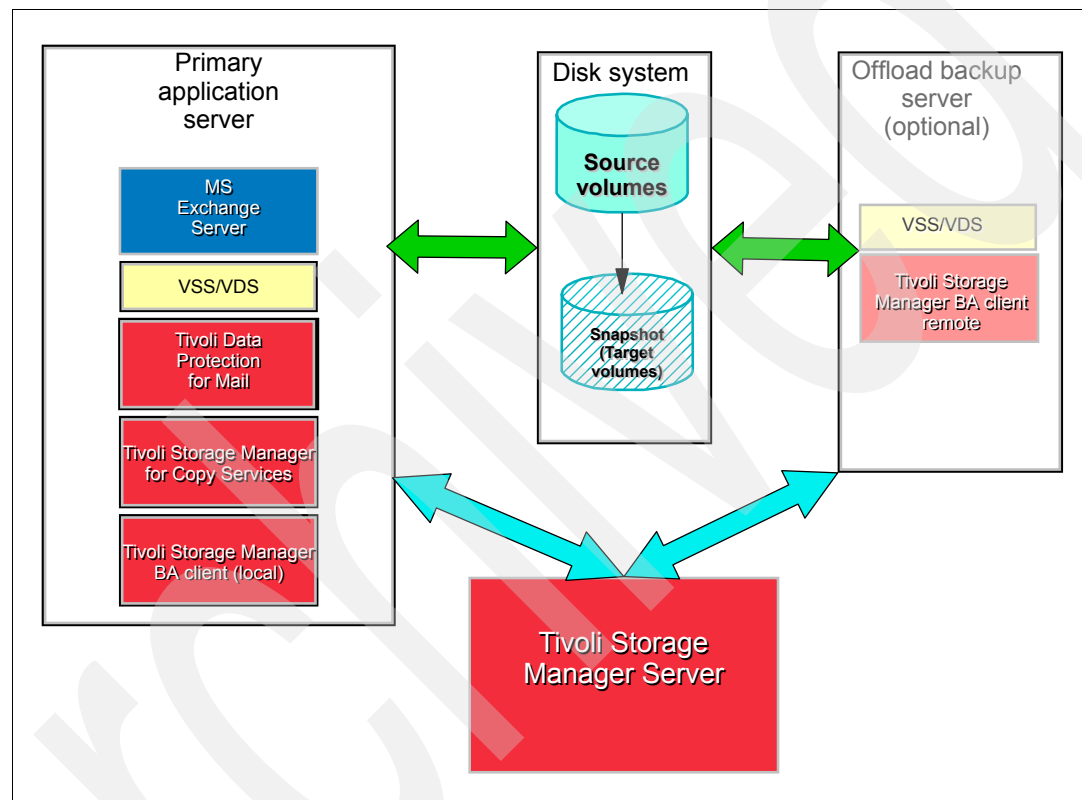


Figure 4-16 Data Protection for Exchange with Tivoli Storage Manager for Copy Services integration

Volume Shadow Copy Service

The Volume Shadow Copy (VSS) service provided with Microsoft Windows Server 2003 is an enhanced Storage Management feature that provides a framework and an API to create consistent point-in-time copies of data known as shadow copies.

The VSS service enables the interoperability of third-party storage management programs, business programs, and hardware providers in order to create and manage shadow copies.

Microsoft Virtual Disk Service

Microsoft Windows 2003 Server Virtual Disk Service (VDS) provides a single interface for management and configuration of multi vendor direct attached and SAN-based storage.

VDS is a set of APIs that provides a single interface for multivendor storage management. Each hardware vendor develops its own VDS hardware provider in order to translate the general purpose VDS APIs into their specific hardware instructions.

VDS uses two sets of providers to manage storage devices:

- ▶ Built-in VDS software providers that manage disks and volumes at the operating system level
- ▶ Hardware providers supplied by the hardware vendor that manage their specific hardware

The Virtual Disk Service is used for managing LUNs on hardware storage devices; managing disks and volumes; and managing end-to-end storage operations.

Tivoli Storage Manager for Copy Services integrates with VSS and any supported disk storage system in order to easily exploit the snapshot on the hardware side and manage the LUN allocation.

VSS with Tivoli Storage Manager

Tivoli Storage Manager can perform VSS backup of Exchange databases using Tivoli Storage Manager for Mail Data Protection for Exchange together with Tivoli Storage Manager for Copy Services.

Types of VSS Exchange backup and restore

You can use Tivoli Storage Manager for Copy Services to perform the following types of VSS (snapshot) **backups**.

VSS backup to LOCAL

This type of backup creates a persistent shadow copy (snapshot) of the Exchange database and log files or the volumes to VSS-capable disks. An integrity check is performed on the backup. The persistent shadow copy data remains on these disks based on the specific policy settings for VSS snapshot backups, and the expiry policy is handled by the Tivoli Storage Manager server. You can use VSS backup to LOCAL to create more frequent copies of your Exchange database, than you would have using traditional non-VSS methods, since this snapshot process takes seconds instead of potentially hours.

VSS backup to TSM

This type of backup creates a non-persistent shadow copy (snapshot) of the Exchange database and logs to VSS-capable disks. The snapshot is then backed up to Tivoli Storage Manager by the production server. After the backup is complete to Tivoli Storage Manager, the snapshot is deleted. The Tivoli Storage Manager backup remains in the storage pools based on special policy setting for VSS snapshots.

VSS backup to BOTH

When you specify backup destination BOTH, then both of the previous backups are performed; a persistent snapshot backup to VSS disk, which is also sent to the Tivoli Storage Manager server.

VSS offloaded backup

In an offloaded backup, the data backup from the VSS snapshot to the Tivoli Storage Manager server is performed by another server, known as the offloaded backup server. The VSS snapshot operation is still performed by the production server; in this case, a transportable snapshot is made. After the snapshot is complete, the target volumes are made available to the alternate (offloaded) server so that it can do the Tivoli Storage Manager backup. This means that the production server is only impacted for the short amount of time it takes to make the snapshot. Offloaded backup is ideal for environments where the production server performance is critical and must be minimally impacted by backup. Use of the offloaded backup server is optional.

After you have made a VSS backup, you have the following options for **restore**.

VSS restore

This restores a VSS snapshot backup (of Exchange database files and log files) residing on Tivoli Storage Manager server storage to its original location. This is a restore of a *VSS backup to TSM*. The restore is done at a file level — that is, all the files in the backup are copied back individually.

VSS fast restore

A VSS fast restore restores a VSS snapshot backup residing on local shadow volumes. This would be a restore of a *VSS backup to LOCAL*. The restore is done at a file level — that is, all the files in the backup are copied back individually.

VSS Instant Restore

In a VSS Instant Restore the target volumes — containing a valid VSS snapshot, that is, a *VSS backup to LOCAL* — are copied back to the original source volumes using *hardware-assisted volume-level copy mechanisms* (such as FlashCopy). This is distinct from the file-level restore, which is performed by VSS restore and VSS fast restore.

After a VSS Instant Restore, the application can return to normal operations as soon as the hardware-assisted volume-level copy and the log replay is complete. Since the data is copied back at the hardware level, from the local shadow volumes, it is very much faster than a network restore from a backup server. Be aware that even though the data is restored relatively quickly, the transaction logs must still be replayed after the restore. Hence, this recovery time increases the time until the application is online again.

At the time of writing, VSS Instant Restore can only be done from VSS snapshots made on an IBM System Storage SAN Volume Controller.

Supported environments

Tivoli Storage Manager for Copy Services is supported in Windows 2003 and Microsoft Exchange 2003. For detailed information, see:

<http://www.ibm.com/support/docview.wss?rs=3042&context=SSRURH&uid=swg21231465>

4.3.4 Tivoli Storage Manager for Advanced Copy Services

Tivoli Storage Manager for Advanced Copy Services software provides online backup and restore of data stored in SAP, DB2 and Oracle applications by leveraging the copy services functionality of the underlying storage hardware. Using hardware-based copy mechanisms rather than traditional file-based backups can significantly reduce the backup/ restore window on the production server. Backups are performed by an additional server called the *backup server*, which performs the actual backup. Since the backup operation is offloaded to the backup server, the production server is free from nearly all the performance impact. The production server's processor time is dedicated for the actual application tasks, so application users' performance is not affected during backup.

Tivoli Storage Manager for Advanced Copy Services is used in conjunction with Tivoli storage other products to interact with the applications and perform the backup from the backup server to Tivoli Storage Manager. The products which it interfaces with are Tivoli Storage Manager for Enterprise Resource Planning (Data Protection for mySAP), Tivoli Storage Manager for Databases (Data Protection for Oracle), and the inbuilt Tivoli Storage Manager interfaces for DB2 UDB.

The following Tivoli Storage Manager for Advanced Copy Services modules are currently available:

- ▶ Data Protection for IBM Disk Storage and SAN Volume Controller for mySAP with DB2 UDB — FlashCopy integration for mySAP with DB2 on SVC, DS6000, DS8000
- ▶ Data Protection for IBM Disk Storage and SAN Volume Controller for mySAP with Oracle — FlashCopy integration for mySAP with Oracle on SVC, DS6000, DS8000
- ▶ Data Protection for IBM Disk Storage and SAN Volume Controller for Oracle — FlashCopy integration for Oracle on SVC, DS6000, DS8000
- ▶ DB2 UDB Integration Module and Hardware Devices Snapshot Integration Module — FlashCopy integration for DB2 on ESS, SVC, DS6000, DS8000
- ▶ Data Protection for ESS for Oracle — FlashCopy integration for Oracle on ESS
- ▶ Data Protection for ESS for mySAP — FlashCopy integration for mySAP with DB2 or Oracle on ESS

In this section we use the abbreviated term Data Protection for FlashCopy as a generic term for all these products, specifying either mySAP Oracle, mySAP DB2, DB2, or Oracle where we have to be more specific.

Tivoli Storage Manager for Advanced Copy Services was previously known as Tivoli Storage Manager for Hardware which was supported on the ESS only. It had the following modules:

- ▶ Data Protection for IBM ESS for mySAP with DB2
- ▶ Data Protection for IBM ESS for mySAP with Oracle
- ▶ Data Protection for IBM ESS for DB2
- ▶ Data Protection for IBM ESS for Oracle

Note: Tivoli Storage Manager for Hardware was withdrawn from marketing on June 16, 2006. It is replaced by Tivoli Storage Manager for Advanced Copy Services.

For more information about Tivoli Storage Manager for Advanced Copy Services, see *IBM Tivoli Storage Manager for Advanced Copy Services*, SG24-7474.

The Tivoli Storage Manager for Advanced Copy Services is a BC Tier 4 solution, as shown in Figure 4-17.

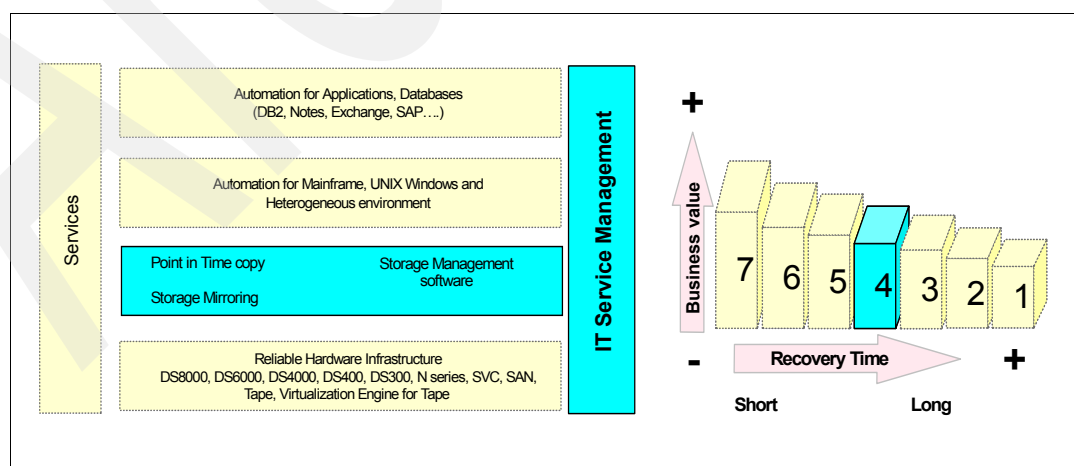


Figure 4-17 BC Tier 4 graph for Tivoli Storage Manager for Advanced Copy Services

Why use Tivoli Storage Manager for Advanced Copy Services?

In today's IT environments, database sizes are large and growing. Most of the servers operate 24 x 7 with very high uptime requirements. With large database size, the traditional direct to tape backup can last for hours, with significant impact on the production server's application performance due to the high I/O activity caused by backup. Faster tape technology cannot necessarily keep up with the shrinking backup windows. Restore time is also critical — restoring a very large database from tape can take too long — meaning too high an outage.

Many storage disk systems provide a snapshot function for a point-in-time copy. However, if this is used (in isolation) when the applications are running or online, the copied data is not in a consistent state for restore. To create a useful, restorable backup, you must have proper application knowledge to interact with the application, and put it in a proper state before performing the snapshot. Scripting is one way to achieve this, however the scripting task is complex, requiring detailed application knowledge, and testing and maintenance effort. A package solution, such as Tivoli Storage Manager for Advanced Copy Services, alleviates this.

In this section, we use the term FlashCopy, which is the IBM implementation of point-in-time snapshot, since at this time, only IBM disk systems are supported as the underlying hardware.

Major challenges for backup/restore

Here are some of the major challenges faced in today's environments:

- ▶ Application databases take a long time to back up.
- ▶ Application performance is impacted during the entire backup window.
- ▶ More archive logs get created during the large backup window causing difficulty in managing them. Also in the event of recovery, it takes time since more archive logs have to be applied.
- ▶ There is a large recovery time (that is, restore takes more time).
- ▶ Application knowledge is required to implement FlashCopy.
- ▶ Scripting is required to automate FlashCopy.

Tivoli Storage Manager for Advanced Copy Services overcomes these challenges, and provides the following benefits:

- ▶ Reduces backup time dramatically to a few seconds on the production server using FlashCopy services of the storage hardware.
- ▶ Application performance on the production server is not impacted, because the actual backup is done from the backup server.
- ▶ Since the backup window is much smaller, fewer archive logs are generated. This means that during recovery, fewer files have to be applied.
- ▶ Database restore can take a few seconds if done using the Flashback restore services of the storage hardware.
- ▶ It is “application-aware,” so consistent, restorable backups can be made.
- ▶ No scripting is required to do the FlashCopy, validate, and do backup. These functions are all integrated within the product.

Tivoli Storage Manager for Advanced Copy Services minimizes the impact on the database servers while allowing automated database backups to the Tivoli Storage Manager server. Tivoli Storage Manager for Advanced Copy Services employs a backup server that offloads the backup data transfer from the FlashCopy volumes to the Tivoli Storage Manager server.

Tivoli Storage Manager for Advanced Copy Services provides options to implement high efficiency backup and recovery of business critical databases while virtually eliminating backup related downtime, user disruption, and backup load on the production server. Data Protection for FlashCopy FlashBack restore functionality provides a fully automated tool for a fast restore of business critical databases.

Supported platforms

Tivoli Storage Manager for Advanced Copy Services (all components), supports AIX 5.2, AIX 5.3 in both 32 bit and 64 bit mode, and the following IBM disk systems:

- ▶ Enterprise Storage Server (ESS) Model 800
- ▶ DS6000 disk storage subsystem
- ▶ DS8000 disk storage subsystem
- ▶ SAN Volume Controller (SVC)

For detailed product support information, see:

<http://www.ibm.com/support/docview.wss?rs=3043&context=SSRUS7&uid=swg21231464>

Data Protection for mySAP (DB2 UDB)

Tivoli Storage Manager for Advanced Copy Services Data Protection for FlashCopy for mySAP (DB2 UDB) performs integrated FlashCopy backup of mySAP databases installed on DB2 UDB. It is well integrated with the DB2 administration utilities and the copy services of the underlying storage system.

Operating environment

Data Protection for FlashCopy for mySAP (DB2 UDB) requires a production server running mySAP with DB2 database on AIX operating system with one of the supported disk systems. The backup server must be another AIX server with access to the same disk system. The backup server backs up the FlashCopy'd data copied from production server to the Tivoli Storage Manager server. The Tivoli Storage Manager can be installed on the backup server, or on another server with connectivity to the production and backup servers. If a separate server is used for the Tivoli Storage Manager server, this can be on any supported operating system platform.

Data Protection for FlashCopy for mySAP (DB2 UDB) has a prerequisite of Tivoli Storage Manager for ERP — Data Protection for mySAP (DB2 UDB) to do the actual backup and restore. The DB2 client is also required on the backup server.

Data Protection for FlashCopy for mySAP with Oracle

Tivoli Storage Manager for Advanced Copy Services Data Protection for FlashCopy for mySAP (Oracle) performs integrated FlashCopy backup of mySAP databases installed on Oracle. It is well integrated with the mySAP DBA tools package BR*Tools and the copy services of the underlying storage system.

Operating environment

Data Protection for FlashCopy for mySAP (Oracle) requires a production server running mySAP with Oracle database on AIX operating system with one of the supported disk systems. The backup server must be another AIX server with access to the same disk system. The backup server backs up the FlashCopy'd data copied from production server to the Tivoli Storage Manager server. The Tivoli Storage Manager can be installed on the backup server, or on another server with connectivity to the production and backup servers. If a separate server is used for the Tivoli Storage Manager server, this can be on any supported operating system platform.

The entire backup is accomplished through the BR*Tools component **brbackup**, Data Protection for FlashCopy (**splitint**), and Data Protection for mySAP (**backint/prole**), working together.

Data Protection for FlashCopy for mySAP (Oracle) has a prerequisite of Tivoli Storage Manager for ERP — Data Protection for mySAP (Oracle) to do the actual backup and restore.

DB2 UDB Integration Module

The Tivoli Storage Manager for Advanced Copy Services DB2 UDB Integration Module, together with the Hardware Devices Snapshot Integration Module, performs integrated FlashCopy backup of DB2 databases.

Operating environment

DB2 UDB Integration Module and Hardware Devices Snapshot Integration Module require a production server running the DB2 UDB database on AIX operating system with one of the supported disk systems. The backup server must be another AIX server with access to the same disk system. The backup server backs up the FlashCopy'd data copied from production server to the Tivoli Storage Manager server. The Tivoli Storage Manager can be installed on the backup server, or on another server with connectivity to the production and backup servers. If a separate server is used for the Tivoli Storage Manager server, this can be on any supported operating system platform.

DB2 UDB Integration Module and Hardware Devices Snapshot Integration Module are supported in a DB2 multi-partitioned environment, as shown in Figure 4-18. In this way, the database application and backup workload can be distributed for better performance.

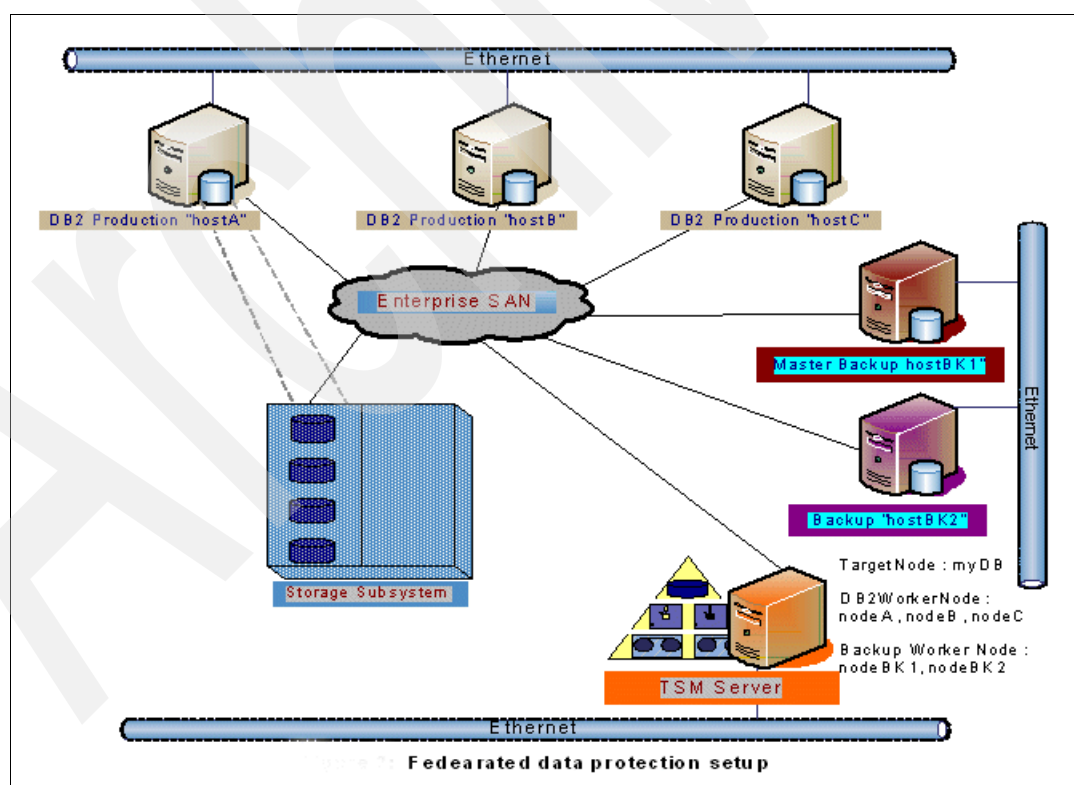


Figure 4-18 DB2 multi partition environment with multiple backup host

Data Protection for FlashCopy for Oracle

Tivoli Storage Manager for Advanced Copy Services Data Protection for FlashCopy for Oracle performs integrated FlashCopy backup of Oracle databases.

Operating environment

Data Protection for FlashCopy for Oracle requires a production server running the Oracle database on AIX operating system with one of the supported disk systems. The backup server must be another AIX server with access to the same disk system. The backup server backs up the FlashCopy'd data copied from production server to the Tivoli Storage Manager server. The Tivoli Storage Manager can be installed on the backup server, or on another server with connectivity to the production and backup servers. If a separate server is used for the Tivoli Storage Manager server, this can be on any supported operating system platform.

Data Protection for FlashCopy for Oracle has a prerequisite of Tivoli Storage Manager for Databases — Data Protection for Oracle to do the actual backup and restore.

Data Protection for mySAP

Tivoli Storage Manager for ERP Data Protection for mySAP is an intelligent client/server program which manages backup and restore of mySAP databases to Tivoli Storage Manager. Data Protection for mySAP lets you manage backup storage and processing independently of normal mySAP operations. Data Protection for mySAP and Tivoli Storage Manager provide reliable, high performance, repeatable backup and restore process that system administrators can manage large databases more efficiently.

Data Protection for mySAP (DB2) does backup and restores of data blocks using the DB2 vendor API as shown in Figure 4-19.

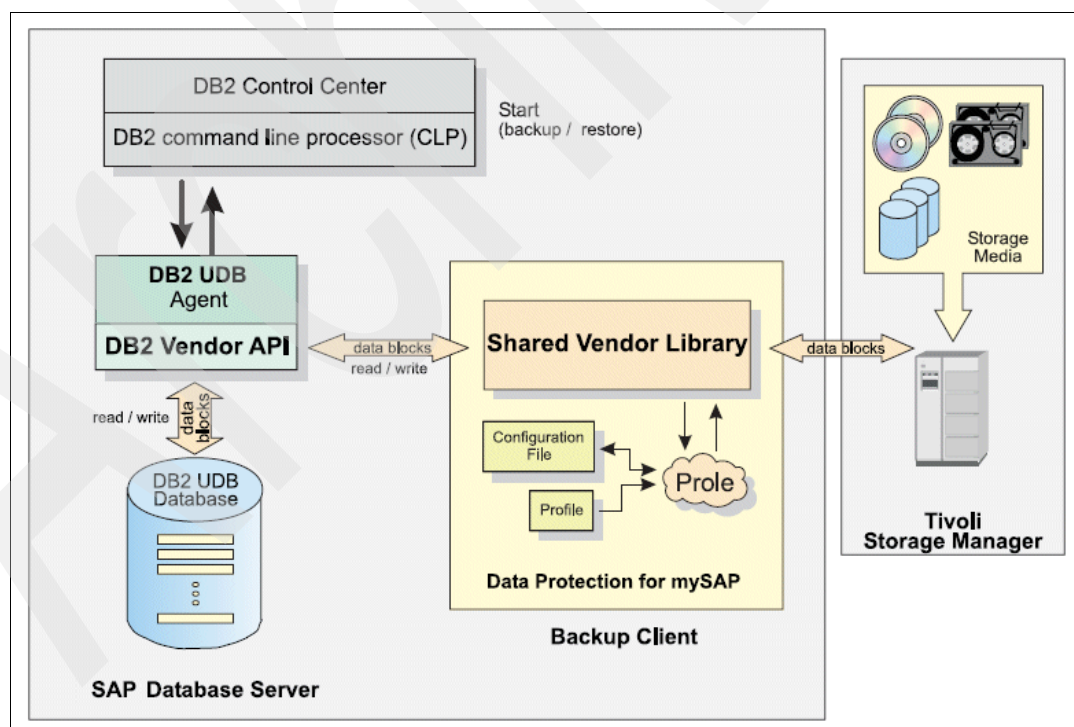


Figure 4-19 Data Protection for mySAP (DB2) overview

Data Protection for mySAP (Oracle) allows system administrators to follow SAP procedures using integrated mySAP database utilities. An overview of Data Protection for mySAP (Oracle) is shown in Figure 4-20.

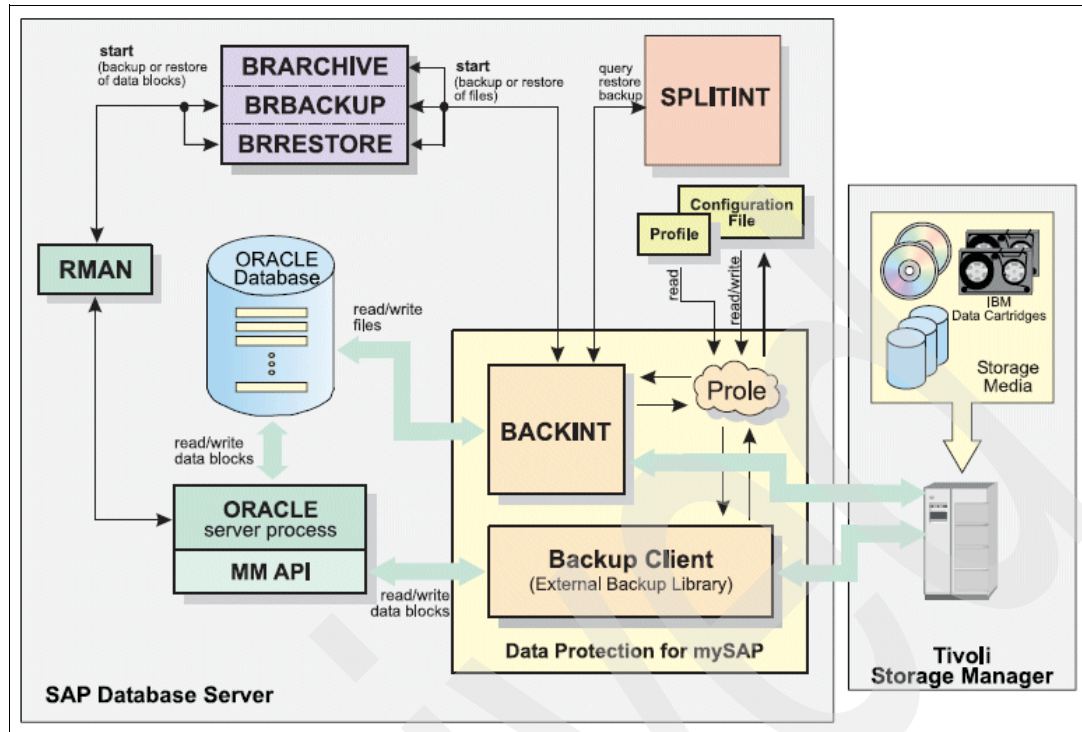


Figure 4-20 Data Protection for mySAP (Oracle) overview

Other SAP files, such as executables and configuration files, must be backed up using Tivoli Storage Manager standard backup-archive client for file backup.

4.3.5 Further information

For more information about products, solutions, and implementation, we recommend that you refer to the following IBM Redbooks and Web sites:

- IBM Redbooks:
 - *Disaster Recovery Strategies with Tivoli Storage Management*, SG24-6844
 - *IBM Tivoli Storage Management Concepts*, SG24-4877
 - *IBM Tivoli Storage Manager Version 5.3 Technical Guide*, SG24-6638

- Web sites:

<http://www.ibm.com/software/tivoli/products/storage-mgr/>
<http://www.ibm.com/storage/th/disk/ds6000/>
<http://www.ibm.com/storage/th/disk/ds8000/>
http://www.ibm.com/servers/aix/products/ibmsw/high_avail_network/hacmp.html
http://www.ibm.com/servers/aix/products/ibmsw/high_avail_network/hageo_georm.html

4.3.6 Summary

At an enterprise software level, Tivoli Storage Manager policy must meet overall business requirements for data availability, data security, and data retention. Enterprise policy standards can be established and applied to all systems during the policy planning process. At the systems level, RTO and RPO requirements vary across the enterprise. Systems classifications and data classifications typically delineate the groups of systems and data along with their respective RTO/RPO requirements.

In view of Business Continuity, enhancement of the Tivoli Storage Manager system to support the highest BC tier is a key for continuity for the backup and restore services. The tier-based IBM Tivoli Storage Manager solutions given above are guidelines to improve your current Tivoli Storage Manager solution to cover your business continuity requirements based on the BC tier. These solutions allow you to protect enterprise business systems by having a zero data lost on business backup data and have continuity of backup and recovery service.

4.4 IBM Data Retention 550

This section discusses the requirement for storage archiving with the IBM System Storage DR550 and DR550 Express, that can address data retention and other regulatory compliance requirements.

More detail on the overall business, legal, and regulatory climate, which is the underlying driving force behind the growth in retention-managed data, can be found in *Understanding the IBM System Storage DR550*, SG24-7091.

4.4.1 Retention-managed data

Beyond laws and regulations, data often has to be archived and managed simply because it represents a critical company asset.

Examples of such data include contracts, CAD/CAM designs, aircraft build and maintenance records, and e-mail, including attachments, instant messaging, insurance claim processing, presentations, transaction logs, Web content, user manuals, training material, digitized information, such as check images, medical images, historical documents, photographs, and many more.

The characteristics of such data can be very different in their representation, size, and industry segment. It becomes apparent that the most important attribute of this kind of data is that it has to be retained and managed, thus it is called *retention-managed data*.

Retention-managed data is data that is written once and is read rarely (sometimes never). Other terms abound to describe this type of data, such as reference data, archive data, content data, or other terms that imply that the data cannot be altered.

Retention-managed data is data that has to be kept (retained) for a specific (or unspecified) period of time, usually years.

Retention-managed data applies to many types of data and formats across all industries. The file sizes can be small or large, but the volume of data tends to be large (multi-terabyte to petabytes). It is information that might be considered of high value to an organization; therefore, it is retained near-line for fast access. It is typically read infrequently and thus can be stored on economical disk media such as SATA disks; depending on its nature, it can be migrated to tape after some period.

It is also important to recognize what does not qualify as retention-managed data. It is not the data that changes regularly, known as *transaction data* (account balance, inventory status, and orders today, for example). It is not the data that is used and updated every business cycle (usually daily), nor is it the backup copy of this data. The data mentioned here changes regularly, and the copies used for backup and business continuity are there for exactly those purposes, meaning backup and business continuity. They are there so that you can restore data that was deleted or destroyed, whether by accident, a natural or human-made disaster, or intentionally.

4.4.2 Storage and data characteristics

When considering the safekeeping of retention-managed data, companies also have to consider storage and data characteristics that differentiate it from transactional data.

Storage characteristics of retention-managed data include:

- ▶ Variable data retention periods: These are usually a minimum of a few months, with no upper limit.
- ▶ Variable data volume: Many clients are starting with 5 to 10 TB of storage for this kind of application (archive) in an enterprise. It also usually consists of a large number of small files.
- ▶ Data access frequency: Write-once-read-rarely or read-never is used (see data life cycle in the following list).
- ▶ Data read/write performance: Write handles volume; read varies by industry and application.
- ▶ Data protection: There are pervasive requirements for non-erasability, non-rewritability, and destructive erase (data shredding) when the retention policy expires.

Data characteristics of retention-managed data include:

- ▶ Data life cycle: Usage after capture, 30 to 90 days, and then near zero, is typical. Some industries have peaks that require access, such as check images in the tax season.
- ▶ Data rendering after long-term storage: Ability to view or use data stored in a very old data format (such as after 20 years).
- ▶ Data mining: With all this data being saved, we think there is intrinsic value in the content of the archive that could be exploited.

4.4.3 IBM strategy and key products

Regulations and other business imperatives, as we just briefly presented, stress the requirement for an Information Lifecycle Management process and tools to be in place.

The unique experience of IBM with the broad range of ILM technologies and its broad portfolio of offerings and solutions can help businesses address this particular requirement and provide them with the best solutions to manage their information throughout its life cycle.

IBM provides a comprehensive and open set of solutions to help. IBM has products that provide content management, data retention management, and sophisticated storage management, along with the storage systems to house the data.

To specifically help companies with their risk and compliance efforts, the IBM Risk and Compliance framework is another tool designed to illustrate the infrastructure capabilities required to help address the myriad of compliance requirements. Using the framework, organizations can standardize the use of common technologies to design and deploy a compliance architecture that might help them deal more effectively with compliance initiatives.

Important: The IBM offerings are intended to help clients address the numerous and complex issues relating to data retention in regulated and non-regulated business environments. Nevertheless, each client's situation is unique, and laws, regulations, and business considerations impacting data retention policies and practices are constantly evolving. Clients remain responsible for ensuring that their information technology systems and data retention practices comply with applicable laws and regulations, and IBM encourages clients to seek appropriate legal counsel to ensure their compliance with those requirements. IBM does not provide legal advice or represent or warrant that its services or products ensure that the client is in compliance with any law.

For more detailed information about the IBM Risk and Compliance framework, visit:

<http://www.ibm.com/software/info/openenvironment/rcf/>

Key products of the IBM data retention and compliance solutions are IBM System Storage Archive Manager and IBM DB2 Content Manager, along with any required disk-based and tape-based storage:

- ▶ IBM DB2 Content Manager for Data Retention Compliance is a comprehensive software platform combining IBM DB2 Content Manager, DB2 Records Manager, and DB2 CommonStore, and services that are designed to help companies address the data retention requirements of SEC, NASD, and other regulations. This offering provides archival and retention capabilities to help companies address retention of regulated and non-regulated data, providing increased efficiency with fast, flexible data capture, storage, and retrieval.
- ▶ IBM System Storage Archive Manager offers expanded policy-based data retention capabilities that are designed to provide non-rewritable, non-erasable storage controls to prevent deletion or alteration of data stored using IBM System Storage Archive Manager. These retention features are available to any application that integrates with the open IBM System Storage Manager API.

Key technologies for IBM data retention and archiving solutions include:

- ▶ IBM System Storage DS4000 with Serial Advanced Technology Attachment (SATA) disk drives to provide near-line storage at an affordable price. It is also capable of Enhanced Remote Mirroring to a secondary site.
- ▶ Write Once Read Many (WORM) media technology (in supported IBM tape drives and libraries). Once written, data on the cartridges cannot be overwritten (to delete the data, the tape must be physically destroyed). This capability is of particular interest to clients that have to store large quantities of electronic records to meet regulatory and internal audit requirements.

Figure 4-21 shows where the DR550 series, built on some of these key products and technologies, fits.

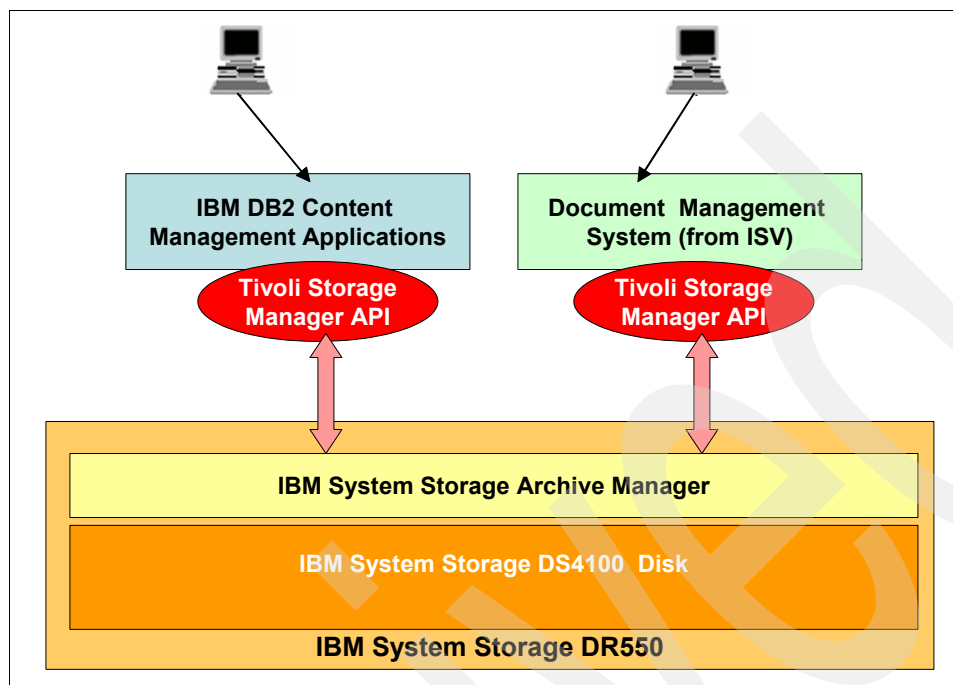


Figure 4-21 DR550 context diagram

The main focus of the IBM System Storage DR550 is to provide for a secure storage system, where deletion or modification of data is completely disallowed except through a well-defined retention and expiration policy.

The IBM System Storage DR550 is the repository for regulated business information. It does not create the information. A complete solution includes applications that gather information, such as e-mail and instant messaging, content management applications, such as IBM DB2 Content Manager or other document management systems from independent software vendors that can index, archive, and later present this information to compliance reviewers, and finally storage systems that retain the data in accordance with regulations.

4.4.4 IBM System Storage DR550

IBM System Storage DR550 (Figure 4-22), one of the IBM Data Retention offerings, is an integrated offering for clients that have to retain and preserve electronic business records. It is designed to help store, retrieve, manage and retain regulated and non-regulated data. In other words, it is not just an offering for compliance data, but can also be an archiving solution for other types of data.

Integrating IBM System p5™ servers (using POWER5™ processors) with IBM System Storage hardware products and IBM System Storage Archive Manager software, this system is specifically designed to provide a central point of control to help manage growing compliance and data retention requirements.

The DR550 brings together off-the-shelf IBM hardware and software products. The hardware comes already mounted in a secure rack. The software is preinstalled and to a large extent preconfigured. The system's compact design can help with fast and easy deployment and incorporates an open and flexible architecture.



Figure 4-22 DR550 dual node with console kit

4.5 System z backup and restore software

There are very well established products and methods to back up System z environments, within the DFSMS family of products. We simply summarize these here and refer the reader to IBM Redbooks such as *Z/OS V1R3 and V1R5 DFSMS Technical Guide*, SG24-6979.

4.5.1 DFSMSdss

The primary function of DFSMSdss is to move and copy data. It can operate at both the logical and physical level and can move or copy data between volumes of like and unlike device types. DFSMSdss can make use of the following two features of the DS6000 and DS8000:

- ▶ **FlashCopy:** This is a point-in-time copy function that can quickly copy data from a source location to a target location.
- ▶ **Concurrent Copy:** This is a copy function that generates a copy of data while applications are updating that data.

DFSMSdss does not communicate directly with the disk system to use these features, this is performed by a component of DFSMSdftp, the System Data Mover (SDM).

4.5.2 DFSMSHsm

Hierarchical Storage Manager (DFSMSHsm) is a disk storage management and productivity tool for managing low-activity and inactive data. It provides backup, recovery, migration, and space management functions as well as full function disaster recovery. DFSMSHsm improves disk use by automatically managing both space and data availability in a storage hierarchy. DFSMSHsm can also be useful in a backup/restore situation.

At a time specified by the installation, DFSMSHsm checks to see whether data sets have been updated. If a data set has been updated then it can have a backup taken. If a data sets are damaged or accidentally deleted, then it can be recovered from a backup copy. There can be more than one backup version, which assists in the recovery of a data set which has been damaged for some time, but this has only recently been detected.

DFSMSHsm also has a feature called Fast Replication that invokes FlashCopy for volume-level replication.

4.5.3 z/VM utilities

VM utilities for backup and restore include:

- ▶ DDR (DASD Dump and Restore), a utility to dump, copy, or print data which resides on z/VM user minidisks or dedicated DASDs. The utility can also be used to restore or copy DASD data which resides on z/VM user tapes
- ▶ DFSMS/VM
- ▶ z/VM Backup and Restore Manager

Refer to your z/VM operating system documentation for more information.

4.6 Solution examples of backup, restore, and archive

In order to illustrate the difference of BC solution taking part in the BC tiers of Tier 1, 2, and 3, this session describes some solution examples of backup, restore, and archive.

The BC solutions cover the BC Tier 1, 2, and 3, as follows:

- ▶ BC Tier 1 — Manual off-site vaulting:
 1. Native tape backup
 2. Shipping backup tape to remote vault for restore in case of disaster
- ▶ BC Tier 2 — Manual off-site vaulting with a hotsite:
 1. Native tape backup
 2. Shipping backup tape to remote hotsite
 3. Restore data at the hotsite, which is readily for quick recovery
- ▶ BC Tier 3 — IBM Tivoli Storage Manager electronic vaulting:
 1. Native tape backup with electronic vaulting
 2. Backup tape are readily at the remote hotsite, without physical media transportation
 3. Restore data at the hotsite, which is readily for quick recovery

4.6.1 BC Tier 1 — Manual off-site vaulting

In this section we present an example of this solution.

IBM Tivoli Storage Manager Manual off-site vaulting

This solution is shown in Figure 4-23.

1. Native tape backup with:

- Standard LTO3 / TS1120 tape
- Encrypted TS1120 tape

2. Native tape archive with:

- Standard LTO3 / TS1120 tape
- LTO3 WORM tape
- Encrypted TS1120 Tape

3. Archive to IBM System Storage DR550 locally

4. Ship backup and archive tapes to remote vault for restore in case of Disaster

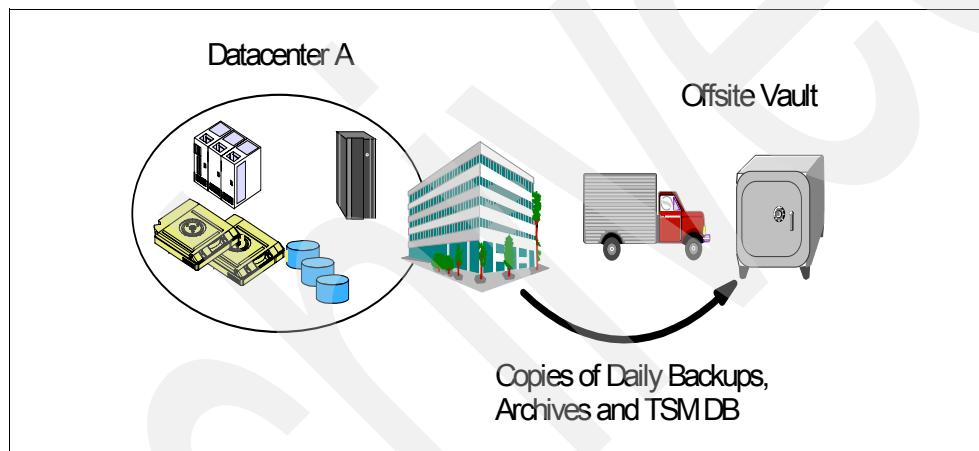


Figure 4-23 IBM Tivoli Storage Manager Manual off-site vaulting — BC Tier 1

Note: More information about Worm Tape and Encryption Tape can be found in Chapter 13, “Tape and Business Continuity” on page 447.

4.6.2 BC Tier 2: Solution example

In this section we present an example of this solution.

IBM Tivoli Storage Manager Manual off-site vaulting with hotsite

This solution is shown in Figure 4-24.

1. Native tape backup with:

- Standard LTO / TS1120 tape
- Encrypted TS1120 tape

2. Native tape archive with:

- Standard LTO / TS1120 tape
- LTO WORM tape
- Encrypted TS1120 Tape

3. Archive to IBM System Storage DR550 locally
4. Ship backup and archive tapes to remote vault
5. Restore the backup data at hot site for quick recovery in case of Disaster

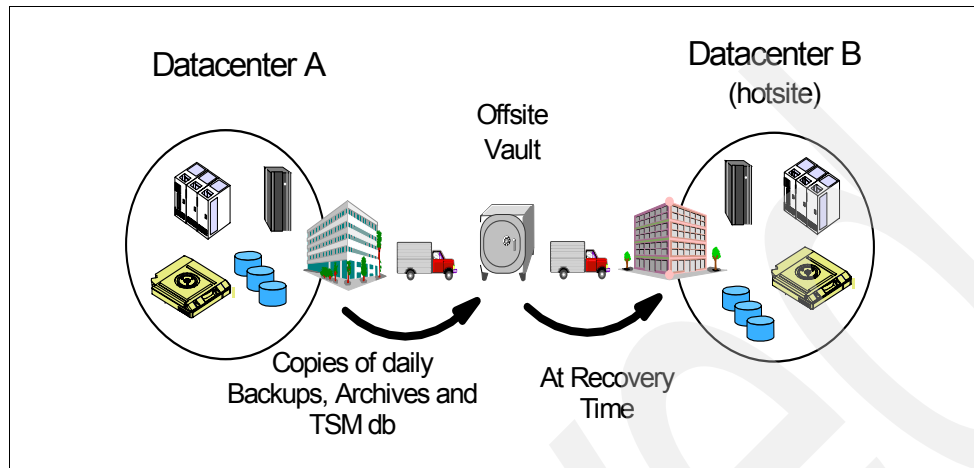


Figure 4-24 IBM Tivoli Storage Manager Manual offsite vaulting with hot site - BC Tier 2

4.6.3 BC Tier 3: Solution example

In this section we present an example of this solution.

IBM Tivoli Storage Manager with electronic vaulting

This solution is shown in Figure 4-25.

1. Tape backup to hot site with Electronic Vaulting:

- Standard LTO / TS1120 tape at hot site
- Encrypted TS1120 tape at hot site

2. Tape archive with to hot site with Electronic Vaulting:

- Standard LTO / TS1120 tape
- LTO WORM tape
- Encrypted TS1120 Tape

3. Archive to IBM System Storage DR550 both locally and remotely with electronic vaulting
4. No necessity to ship tape media to remote site (better data currency at hot site)
5. Restore the backup data at hot site for quick recovery in case of disaster

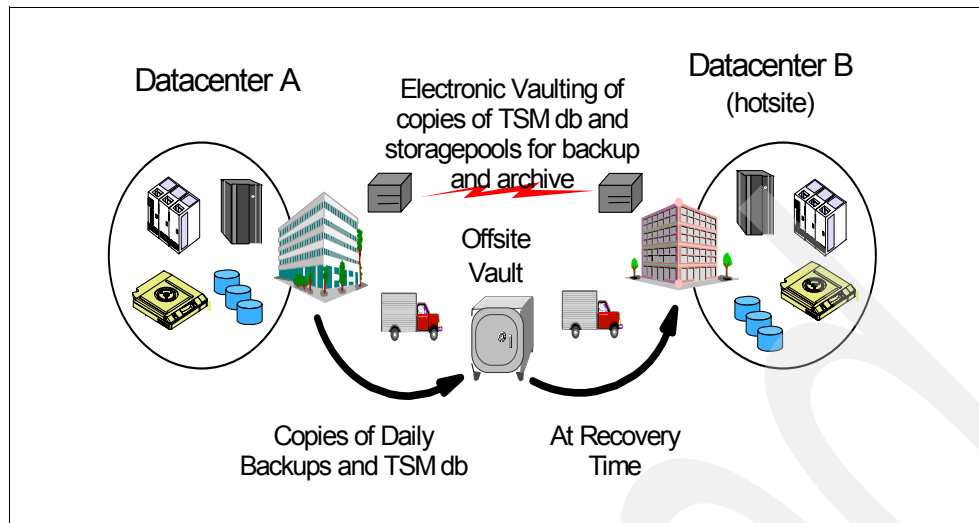


Figure 4-25 IBM Tivoli Storage Manager with electronic vaulting - BC Tier 3

4.7 Summary

Typically, the backup and restore segment has been positioned in Business Continuity Tier 1, Tier 2, and Tier 3.

In 4.3, “IBM Tivoli Storage Manager overview” on page 182, we showed that backup and restore can be applied in Business Continuity Tier 4, Tier 5 and Tier 6 as well.

Furthermore, archive data is also playing an important role for business continuity, which we have discussed in 4.4, “IBM Data Retention 550” on page 206.

The backup and restore can be flexibly implemented with Enterprise Data Management software such as Tivoli Storage Manager. Enterprise Data can be policy managed to achieve various business requirements and objectives on business continuity.

Snapshot and FlashCopy can be easily exploited with Tivoli Storage Manager for Copy Services and Tivoli Storage Manager for Advanced Copy Services.

Advanced WORM Tape or WORM Disk Technology can be easily exploited with Tivoli Storage Manager on storage pooling with WORM media.

Finally, with the availability of Encrypted Key Management capability of Tivoli Storage Manager, sensitive data which requires higher security protection can now be easily deployed on storage pooling of TS1120 tape. Backup and restore is making another step forward on business deployment with Information Technology.



Part 2

Business Continuity component and product overview

In this part of the book, we describe the IBM System Storage portfolio as well as other required elements for a Business Continuity solution.

We cover the following topics:

- ▶ Overview of Resiliency Family
- ▶ Storage Networking for IT business continuity
- ▶ DS6000, and DS8000, and ESS
- ▶ DS4000 series
- ▶ N series
- ▶ DS300 and DS400
- ▶ Virtualization products
- ▶ Storage Management software
- ▶ Tape and Disaster Recovery

Archived



Overview of IBM System Storage Resiliency Portfolio

In this chapter we give you an overview of how IBM System Storage organizes our IT Business Continuity solutions into the System Storage Resiliency Portfolio.

All IBM Business Continuity products, including servers, storage, software, and automation, networking, and services, are organized in a consistent and logical manner via this portfolio.

5.1 System Storage Resiliency Portfolio

IBM has organized its business continuity portfolio of IBM products, including servers, storage, software and automation, networking, and services in a consistent and logical manner via the System Storage Resiliency Portfolio.

In this architecture, the portfolio of IBM business continuity products are mapped according to their function, as shown in Figure 5-1.

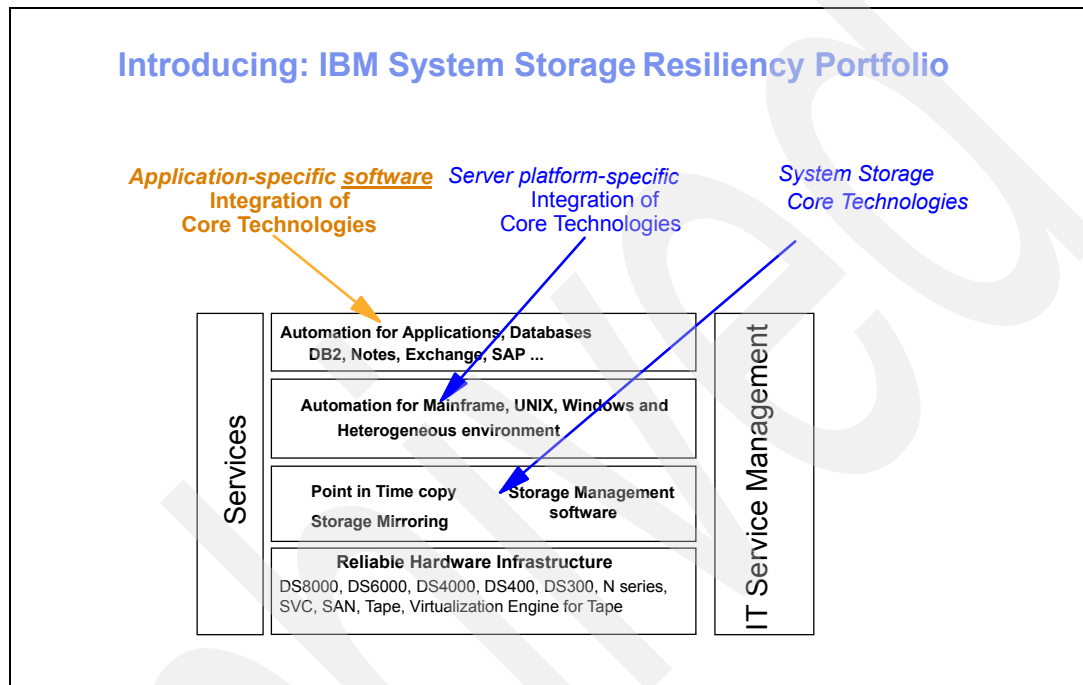


Figure 5-1 System Storage Resiliency Portfolio

In addition to the products being mapped into the System Storage Resiliency Portfolio, we also add *Services and Skills*, which are available through IBM Business Partners or IBM Global Services. These services and skills play a major role in defining and implementing necessary procedures such as component failure impact analysis, business impact analysis, a detailed definition of the required recovery time objectives, implementation of procedures, processes, problem management, change management, system management, monitoring, integration, and testing.

Updating necessary business procedures, processes, tools, problem and change management, in order to attain maximum effectiveness of Business Continuity products is provided via these services and skills, as well as providing experience to advise how to integrate them into the client environments.

The System Storage Resiliency Portfolio architecture also has a strategic purpose. The architecture is designed to provide foundation for future On Demand business continuity solutions. In order for future solutions to meet long range requirements in growth, scalability, interoperability across platforms and applications, with flexibility and efficiency, the System Storage Resiliency Portfolio architecture provides for an open standards-based set of *consistent interfaces*, between the System Storage Resiliency Portfolio layers. The System Storage Resiliency Portfolio architecture defines where and how these interfaces occur; the architecture and location of these interfaces is designed to provide future autonomic capability between products in the different layers.

An example objective for the autonomic System Storage Resiliency Portfolio architecture, is to support dynamic discovery and exploitation of new On Demand functionalities as they are introduced into the solution. For example, suppose that a FlashCopy point-in-time copy technology is introduced into a disk subsystem. The architected interfaces in the System Storage Resiliency Portfolio would allow an On Demand operating system to detect this new feature, and begin to dynamically use it. When backup requests are received by the On Demand operating system from an On Demand application layer, the application would also automatically receive the benefit of the fast point-in-time copy.

Let us now examine the product layers of the architecture in more detail.

5.1.1 Reliable hardware infrastructure layer

As discussed in the roadmap to IT Business Continuity, this layer provides the fault-tolerant and highly-available infrastructure (Figure 5-2).

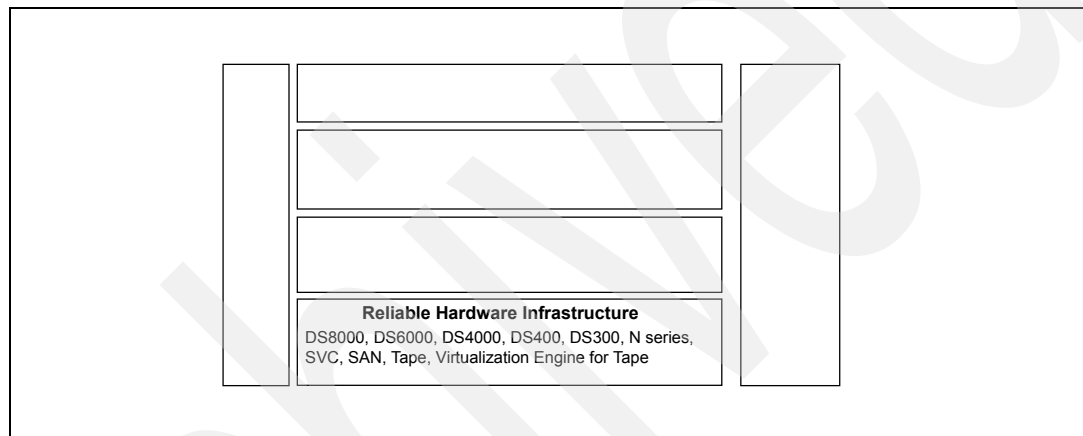


Figure 5-2 Reliable hardware infrastructure

Functionalities such as storage RAID, dual power supplies, dual internal controllers, redundant internal component failover, and so forth, all reside in this layer.

The following examples of IBM System Storage products, which reside in this layer, are designed to provide a robust reliability foundation layer:

- ▶ IBM DS Family disk systems: DS8000, DS6000, ESS, N series, DS4000, DS400, and DS300
- ▶ IBM Storage Virtualization products: SAN Volume Controller (SVC)
- ▶ IBM Storage Area Network: Cisco SAN switches, IBM b-type SAN switches, IBM m-type switches
- ▶ IBM TS Family tape system: Including drives, libraries and virtualization products, such as the TS3500 Enterprise Tape Library, TS7700 Virtual Tape server, and more
- ▶ IBM Servers: System z, System z Sysplex Coupling Facility, System p, System i, System x™, BladeCenter®, LPAR capability, and more.

5.1.2 Core technologies layer

Core technologies provide advanced copy functions for making point-in-time copies, remote replicated copies, file backup and restore, volume images, and other advanced replication services (Figure 5-3).

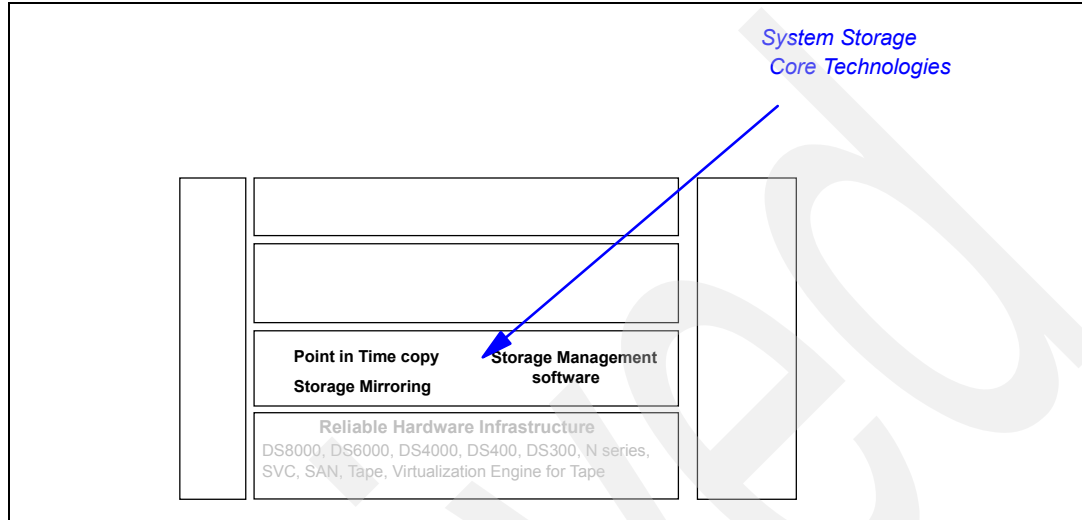


Figure 5-3 IBM System Storage advanced core technologies

Core technology functionality examples include (but are not limited to):

1. Nondisruptive point-in-time copies. In IBM disk storage, this is called FlashCopy.
2. Real-time synchronous storage mirroring at metropolitan distances: to make remote mirrors of important volumes or LUNs. In IBM storage, this is called Metro Mirror.
3. Remote storage mirroring capabilities at long distances: asynchronous copy technology on disk or tape subsystems. In IBM storage, this is called Global Mirror.
4. Storage Virtualization capabilities, via the IBM SAN Volume Controller.
5. Storage Management software, such as TotalStorage Productivity Center and Tivoli Storage Manager.

Specific product examples of core technologies on storage devices include (but are not limited to):

- ▶ FlashCopy: on DS8000, DS6000, ESS, DS4000, DS400, DS300, SAN Volume Controller (SVC) and N series
- ▶ Metro Mirror: on DS8000, DS6000, ESS, DS4000, SAN Volume Controller (SVC), N series and Virtual Tape Server
- ▶ Global Mirror: on the DS8000, DS6000, ESS, SAN Volume Controller, DS4000, N series, and Virtual Tape Server
- ▶ z/OS Global Mirror (Extended Remote Copy XRC): DS8000, ESS

Note: On IBM N series, the point-in-time copy function is called SnapShot. Synchronous local mirroring is called SyncMirror, and remote mirroring is called SnapMirror®.

Also within this layer is storage virtualization and storage management software for Business Continuity planning, enterprise-wide policy-based management of backup copies, and administration and management of the storage-provided replication services.

Specific product examples of these storage-based functions include (but are not limited to):

- ▶ Tivoli Continuous Data Protection for Files: for file-level protection using continuous data protection technology.
- ▶ Storage Virtualization capabilities via the IBM SAN Volume Controller.
- ▶ Tivoli Storage Manager capabilities that provide integrated invocation of point-in-time FlashCopy core technologies. These capabilities provide the foundation for online, backup of popular applications, and are further enhanced by additional Tivoli Storage Manager application-aware integration of software such as DB2, Oracle, SAP, WebSphere, Microsoft SQL Server, Microsoft Exchange, and Lotus® Domino®.
- ▶ N series integration software for integrated invocation of point-in-time Snapshot core technologies. These N series software products provide the foundation for point-in-time copies of the N series filer, and are further enhanced by application-aware, online, backup of popular applications include backup of Microsoft Exchange, Microsoft SQL Server, Oracle, and Lotus Domino.
- ▶ TotalStorage Productivity Center for management of the storage infrastructure, including SAN fabric, disk configuration, replication configuration and data/storage utilization.
- ▶ IBM DFSMS family of offerings for z/OS on System z servers.

5.1.3 Platform-specific integration layer

The next layer is the automation and integration of platform-specific commands and procedures to support use of the core technologies (Figure 5-4).

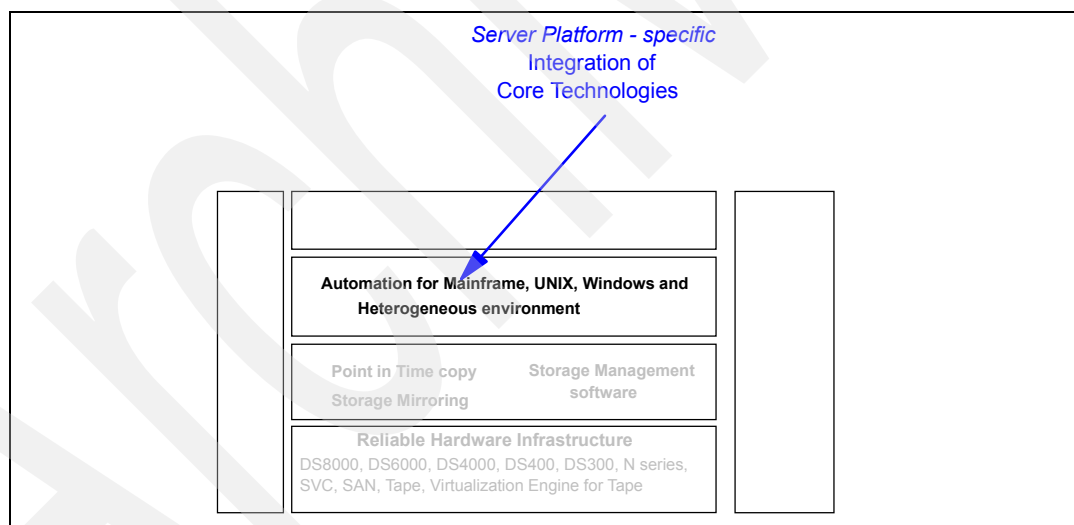


Figure 5-4 Server platform integration

This layer contains specific integrations by operating system. Currently, four categories of operating systems with integrated automation available are:

- ▶ System z
- ▶ System p
- ▶ Heterogeneous operating systems

Each specific category has specific products for that operating system. We provide a brief overview next; each of these solutions is described in more detail later in this book.

System z

Server integration and automation for business continuity core technologies are provided by the Geographically Dispersed Parallel Sysplex (GDPS) capabilities. GDPS is used by System z clients worldwide to provide automated management of Business Continuity and for automated management for high availability in Parallel Sysplex environments. GDPS supports Metro Mirror, z/OS Global Mirror, and Global Mirror.

GDPS HyperSwap Manager provides an automation capability for storage recovery tasks, without having to implement a full GDPS project of automation, including application restart, which is usually a considerably bigger project. GDPS HyperSwap is a new Rapid Data Recovery capability that allows I/Os to the mainframe to be dynamically switched between Metro Mirror sites. This can provide a System z data center-wide capability to execute unplanned or planned site switches in seconds. In addition to the unplanned outage nearly continuous System z operations recovery, this capability also enables planned outage avoidance; for example, performing a planned site switch in order to enable scheduled disk subsystem upgrades, maintenance, changes, without having to take the applications down.

All GDPS offerings are IBM Global Services offerings. More information about GDPS can be found in 2.1, “Geographically Dispersed Parallel Sysplex (GDPS)”.

AIX and System p

AIX High Availability Clustered Multi Processing eXtended Distance (HACMP/XD) is the high availability, clustered server support for System p processors and AIX applications. HACMP is a proven solution, providing AIX cluster failover, application restart, network takeover, workload management, and automated failback. HACMP/XD leverages Metro Mirror on supported IBM System Storage disk systems.

HACMP/XD can support cluster failover to a different site at metropolitan distances, and can restart applications using Metro Mirror copies of the primary disks. This combines with the System p cluster solution for high availability, using IBM System Storage as the mechanism for replication of AIX data to remote LUNs and the remote site.

More information about AIX HACMP/XD can be found in 2.4, “HACMP/XD” on page 71.

Heterogeneous open systems servers

For the heterogeneous server environment, the TotalStorage Productivity Center for Replication (TPC for Replication) is the IBM strategic disk replication management software, for providing an enterprise-wide disk mirroring control and management integration package for open systems data on IBM DS6000, DS8000, ESS, and SAN Volume Controller FlashCopy, Metro Mirror or Global Mirror. TPC for Replication manages the servers, the mirroring, and automates the insurance of a consistent recovery point for multiple servers's data. TPC for Replication also has extensive capabilities for testing and automation of the environment. TPC for Replication scales to support small and large sized configurations, which include DS8000, DS6000, ESS, or SAN Volume Controller FlashCopy, Metro Mirror, or Global Mirror environments.

More information about TPC for Replication is given in Chapter 3, “Rapid Data Recovery” on page 107 and 12.4.4, “IBM TotalStorage Productivity Center for Replication” on page 399.

5.1.4 Application-specific integration layer

This layer contains automation and integration packages that provide application-specific commands and procedures to support use of core technologies, and where necessary, interoperates with the server integration layer (Figure 5-5).

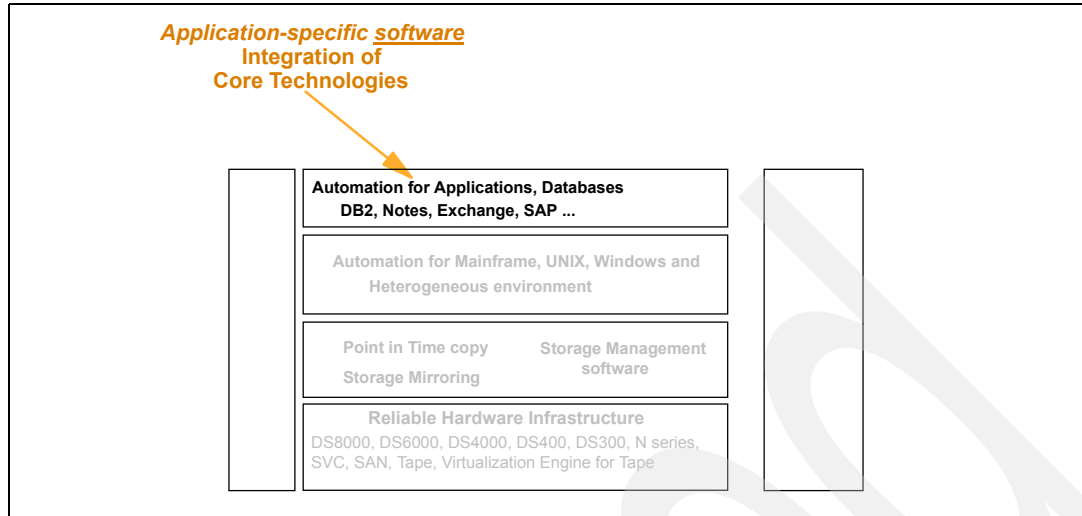


Figure 5-5 Application-specific integration

Application-specific integration examples are described in the following sections.

Tivoli Storage Manager application integration

This provides capabilities for application-aware, online, and backup of popular applications, including backup of DB2, Oracle, SAP, WebSphere, Microsoft SQL Server, Microsoft Exchange, and Lotus Domino. Examples include (but are not limited to):

- ▶ Tivoli Storage Manager for Databases
- ▶ Tivoli Storage Manager for Mail
- ▶ Tivoli Storage Manager for ERP

These are integrated solutions that combine System Storage FlashCopy core technology with an application-specific, end-to-end integration to make nondisruptive point-in-time backups and clones of the databases, mail servers, or ERP applications. All operations use the appropriate software interfaces to implement a fully automated replication process, which generates backups or clones of very large databases in minutes instead of hours. This solution eliminates the requirement for intermediate backups and can help address the client's key requirement for high application availability.

N series integration

This provides capabilities for application-aware, online, and backup of popular applications, including backup of Microsoft Exchange, Microsoft SQL Server, Oracle, and Lotus Domino. Examples include (but are not limited to):

- ▶ N series SnapManager® integration for Exchange, SQL Server, and Oracle
- ▶ N series Single Mailbox Recovery for Exchange

These are integrated solutions that combine N series Snapshot core technology with an application-specific, end-to-end integration to make nondisruptive point-in-time backups and clones of the databases, mail servers, and so on. All operations use the appropriate software interfaces to implement a fully automated replication process, which generates backups or clones of very large databases in minutes instead of hours. This solution eliminates the requirement for intermediate backups and can help address the client's key requirement for high application availability.

5.1.5 Summary

Services and skills tie all the technology components of the System Storage Resiliency Portfolio together, leading toward an end-to-end, automated business continuity solution (Figure 5-6).

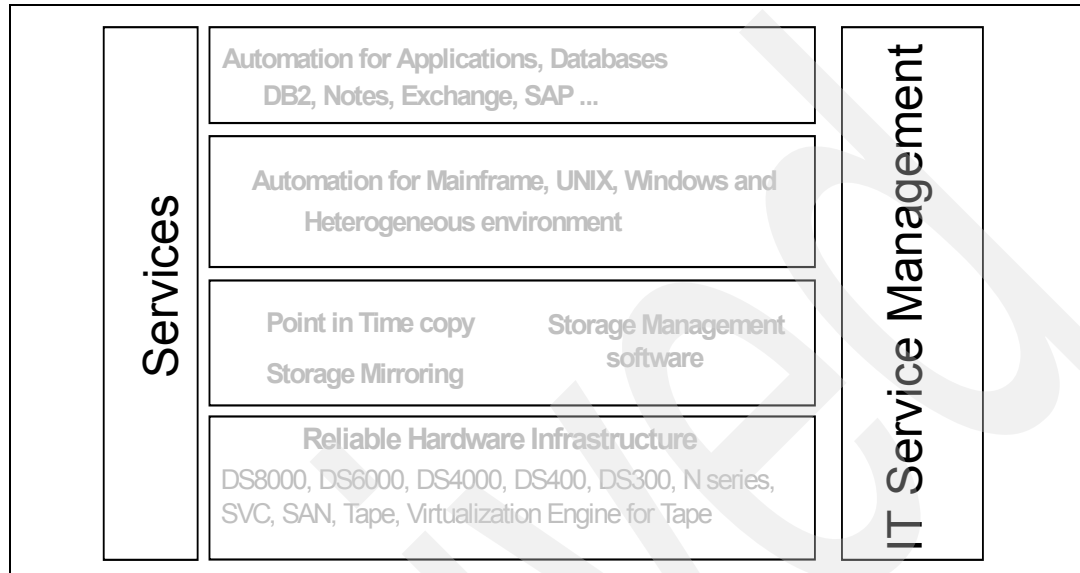


Figure 5-6 Services and IT Service Management capabilities to tie it all together

As stated earlier, the strategic value of the System Storage Resiliency Portfolio is the architecturally defined open standards-based *interfaces* between the Portfolio layers.

These interfaces provide points of technology interlock, exploitation, and discovery for future On Demand System Storage Resiliency Portfolio solutions.

System Storage Resiliency Portfolio is a family of IBM business continuity products from various product brand families and solutions, to help build a complete set of business continuity solutions that can provide:

- ▶ Protection for critical business data
- ▶ Improve business resiliency
- ▶ Lower daily operating costs

Using the core technologies and automation services described, the IBM System Storage Resiliency Portfolio has a spectrum of business continuity solutions, from basic backup and restore requirements, to a multi-site, almost real-time, disaster recovery implementation.

Clients can match their business continuity requirements with their budget constraints, choosing the most appropriate strategy and cost combination for their applications. Automation support helps decrease recovery time, can help eliminate errors in recovery, and can help ensure predictable recovery. System Storage Resiliency Portfolio is the architecture for the IBM Business Continuity products, solutions, and services.

All IBM Business Continuity products, including servers, storage, software and automation, networking, and services, are organized in a consistent and logical manner via this architecture.



Storage networking for IT Business Continuity

In this chapter we describe the main aspects of storage networking related to IT Business Continuity. We begin with an overview, and then discuss the IBM product portfolio, backup, and disaster recovery considerations for SAN, NAS, and iSCSI.

6.1 Storage networking overview

When Disaster Recovery and Business Continuity emerged as formal IT disciplines in the 1980s, the focus was on protecting the data center — the heart of a company's heavily centralized IT structure. This model began to shift in the early 1990s to distributed computing and client/server technology. At the same time, information technology became embedded in the fabric of virtually every aspect of a business. Computing was no longer something done in the background. Instead, critical business data could be found across the enterprise — on desktop PCs and departmental local area networks, as well as in the data center.

The storage networking technologies allow you to explore the potential for ensuring that your business can actually continue in the wake of a disaster.

Storage networking can do this by:

- ▶ Providing for greater operational distances
- ▶ Providing mirrored storage solutions for local disasters
- ▶ Providing three-site mirrored storage solutions for metropolitan disasters
- ▶ Providing failover support for local disasters
- ▶ Improving high availability clustering solutions
- ▶ Allowing selection of “best of breed” storage
- ▶ Providing remote tape vaulting
- ▶ Providing high availability file serving functionality
- ▶ Providing the ability to avoid space outage situations for higher availability

By enabling storage capacity to be connected to servers at a greater distance, and by disconnecting storage resource management from individual hosts, a storage networking solution enables disk storage capacity to be consolidated. The results can be lower overall costs through better utilization of the storage, lower management costs, increased flexibility, and increased control.

Options for connecting computers to storage have increased dramatically. Variations (and their associated acronyms) for storage networking seem to materialize out of thin air faster than they can be tracked. Understanding the technology basics is essential to making the best choices.

Storage networking is one of the fastest-growing technologies in the industry. This chapter describes the basics of IBM Storage Networking and should help you understand SAN, NAS, and iSCSI, as well as how to position them for disaster recovery solutions.

6.1.1 Storage attachment concepts

Now let us discuss the basic concepts for understanding the storage attachment alternatives:

- ▶ *Connectivity*: How processors and storage are physically connected. This includes the OSI layers model which is described in the appendix “Networking terminology tutorial” in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.
- ▶ *Media*: The type of cabling and associated protocol that provides the connection.
- ▶ *I/O protocol*: How I/O requests are communicated over the media.

It is how these three items are combined in practice that differentiates the various ways processors (hosts) and storage can be connected together. Essentially, the hosts are attached to the storage devices over a direct or network connection, and they communicate by way of an I/O protocol that runs on top of the physical media transport protocol.

Figure 6-1 shows a general overview of SAN, NAS, and iSCSI concepts.

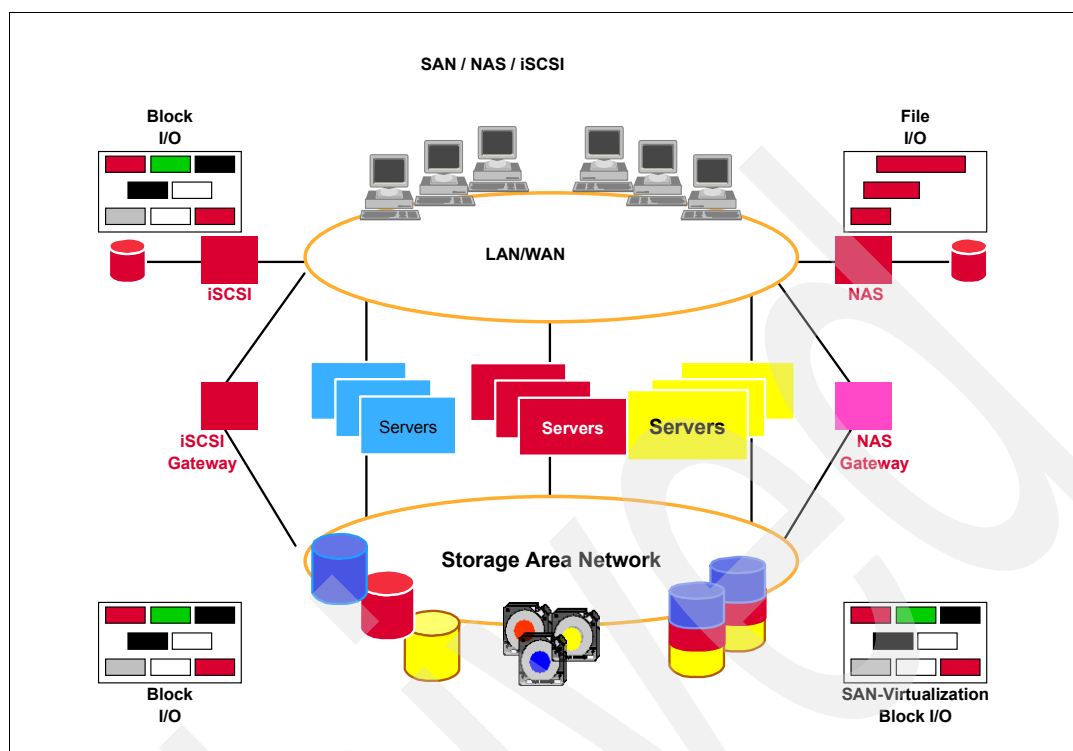


Figure 6-1 Storage networking - general overview of SAN, NAS, and iSCSI

A Storage Area Network is a dedicated Fibre Channel network for the connection of storage to servers. NAS and iSCSI are connected to the common LAN/WAN. Both use the existing TCP/IP network for the transport of data. In most cases this brings additional load to the network.

Figure 6-1 also shows the protocols used for I/O. You see that all mentioned techniques, except NAS, use block I/O. SAN (FCP and FICON), iSCSI, and also SAN virtualization use block I/O channel protocol. iSCSI uses TCP/IP and SAN uses Fibre Channel protocol as the transport layer. The original I/O of the application is not changed because the SCSI channel protocol is used. Whether you use an NAS appliance or your own NAS file server, you implement not only storage, but a complete server with the application *file services* on it. This will probably be a more complicated effort. You should also evaluate if and how well your applications run with an NAS solution.

Talking about the SAN normally means the Storage Network in open environments. Fibre Channel Protocol (FCP) as a channel technology for Open SAN is comparable to the introduction of FICON in mainframe environments.

FCP is SCSI over Fibre Channel, so the underlying protocol is still SCSI-3 channel protocol. This makes it very easy to change from SCSI native to Fibre Channel FCP connected devices. FICON is the name of the Fibre Channel protocol that transports z/OS channel programs over Fibre Channel. The data format on mainframe (ECKD™ - extended count key and data) is totally different than open systems (FB - Fixed block).

We could therefore totally separate these two areas, except that many storage systems such as the IBM DS6000 and DS8000 support both types of data: mainframe and open systems. Disk systems also support different connection types on the Host Adapters such as FICON, FCP and SCSI. Storage consolidation is used in both mainframe and open systems, and they can both use the same fibre network infrastructure with FCP and FICON. In this case the storage networking hardware supports both protocols.

How the virtualization products use a SAN is described in Chapter 11, “Storage virtualization products” on page 359.

Description of storage networking methods

In this section we describe several storage networking methods, starting with the block-level methods.

Storage Area Network (SAN)

SAN indicates storage which resides on a dedicated network, providing an any-to-any connection for processors and storage on that network. The most common media is Fibre Channel. The I/O protocol is SCSI.

iSCSI

Storage is attached to a TCP/IP-based network, and is accessed by block-I/O SCSI commands. iSCSI could be direct-attached or network-attached (DAS or SAN).

iSCSI gateway

An iSCSI gateway, or storage router, enables the connection of IP hosts to FC-connected storage devices. The storage router provides IP access to storage over an existing common LAN infrastructure or over a logically, or physically separate, IP network.

Then there is file-level networking

Network Attached Storage (NAS)

An NAS device is attached to a TCP/IP-based network (LAN or WAN), and accessed using CIFS (Common Interface File System) and NFS (Network File System), specialized I/O protocols for file access and file sharing. CIFS protocol is typically used in Windows-based environments, while NFS protocol is typically used in UNIX-based environments.

An NAS device is sometimes also called a *file server*, *filer*, or *NAS appliance*. It receives an NFS or CIFS request over a network and by its internal processors translates that request to the SCSI block-I/O commands to access the appropriate device only visible to the NAS product itself.

NAS gateway

An NAS gateway is a device with an internal processor but without integrated storage. Instead, the NAS device connects to storage by direct attachment or by a SAN. This term is most meaningful when there is a choice of the disk storage to attach to the gateway.

The various storage networking alternatives are summarized in Table 6-1.

Table 6-1 Comparison of storage networking techniques

Processor-storage connection	Block or file	Media	I/O protocol	Capacity sharing	Data sharing
SAN	Block	Fibre Channel is most common, with Ethernet for geographical SAN extension emerging.	SCSI (FCP)	Yes Yes, it requires specialized software such as clustering software	
		FCIP and iFCP are also used	SAN extension over IP for metro and global mirroring		
NAS/ NAS Gateway	File	Ethernet	NFS, CIFS	Yes	Yes
iSCSI/ iSCSI Gateway	Block	Ethernet	SCSI	Yes	Yes, it requires specialized software such as clustering software

IBM uses SAN storage as the basis for the infrastructure for all storage network products. Therefore we find NAS gateways, support of iSCSI-Gateways, and the IBM SAN Volume Controller, as well as functions located within the switches (iSCSI blades, SAN Volume Controller blades) having their storage functionality within a SAN. The advantage of separating storage data from the connecting protocol is more flexibility and, in the case of Business Continuity, having consolidated data is a good foundation for the solution.

6.2 Storage Area Network (SAN)

This section gives you more background information about SAN, including a short overview about the products and how to design SANs with disaster recovery considerations.

6.2.1 Overview

A SAN is a dedicated network for storage devices and the processors that access those devices. SANs today are generally built using Fibre Channel technology, but the concept of a SAN is independent of the underlying network.

I/O requests to disk storage on a SAN are called block I/Os because, just as for direct-attached disks, the read and write I/O commands identify a specific device (disk drive or tape drive) and, in the case of disks, specific block (sector) locations on the disk.

The major potential benefits of a SAN can be categorized as follows:

► **Access:**

Compared to SCSI, a SAN delivers longer distances between hosts and storage, higher availability and improved performance. Also, a larger number of hosts can be connected to the same storage device compared to typical built-in device attachment facilities.

A 4-Gbps SAN can connect devices at the following distances:

- 150 m with short-wavelength SFPs (850nm)
- 10 km with long-wavelength SFPs (1310nm)
- Up to around 100 km using WDM technologies (CWDM and DWDM), or many hundreds of km when also using amplification and regeneration of the fiber signals
- Above 100 km, preferably using a WAN protocol (SDH/SONET, IP) and FCIP or iFCP gateways or other supported channel extensions.

► **Consolidation and resource sharing:**

By enabling storage capacity to be connected to servers at a greater distance, and by disconnecting storage resource management from individual hosts, a SAN enables disk storage capacity to be consolidated. Consolidation means the replacement of multiple independent storage devices by fewer devices that support capacity sharing. Data from disparate storage systems can be combined onto large, enterprise class shared disk arrays, which can be located at some distance from the servers.

The capacity of these disk arrays can be shared by multiple servers, and users can also benefit from the advanced functions typically offered with such subsystems. This can include RAID capabilities, remote mirroring, and instantaneous data replication functions, which might not be available with smaller, integrated disks. The array capacity can be partitioned, so that each server has an appropriate portion of the available data storage.

Available capacity can be dynamically allocated to any server requiring additional space. Capacity not required by a server application can be re-allocated to other servers. This avoids the inefficiency associated with free disk capacity attached to one server not being usable by other servers. Extra capacity can be added, in a nondisruptive manner. Storage on a SAN can be managed from a single point of control.

Controls over which hosts can see which storage, called *SAN partitioning* and *LUN masking*, usually are implemented:

- SAN partitioning allows for segmentation of the switched fabric. Various methods are available on switches and directors from different vendors. For example, McDATA provides SAN LPARs, Cisco, and McDATA provide Virtual SANs (VSANs); zoning is available from all SAN product vendors. Zoning remains the standard, basic approach for segregating SAN ports from one another. Partitioning can be used to instigate a barrier between different environments so that only the members of that partition can communicate within the partition, and all other attempts from outside are rejected.
- LUN masking is another approach to securing storage devices between various hosts. A LUN is a logical unit number assigned to a logical disk image within a disk storage system. LUN masking allows administrators to dictate which hosts can see which logical disks. It is an access control mechanism that runs in the disk storage system. LUN masking and SAN Zoning are important access control mechanisms that make disk storage consolidation possible in heterogeneous environments. Disk storage consolidation can lower overall costs through better utilization of disk storage, it can lower management costs and increased flexibility. Consolidated storage can be an important starting point for Business Continuity.

► **Protection:**

LAN-free backup solutions offer the capability for clients to move data directly to tape using the SAN. The less common *server-free backup* solutions allow data to be read directly from disk to tape (and tape to disk), bypassing the server, and eliminating the processor overhead.

► **Data Sharing:**

The term *data sharing* is sometimes interpreted to mean the replication of files or databases to enable two or more users, or applications, to concurrently use separate copies of the data. The applications concerned can operate on different host platforms.

A SAN can ease the creation of such duplicate copies of data using facilities such as remote mirroring. Data sharing can also be used to describe multiple users accessing a single copy of a file. This could be called *true data sharing*.

In a homogeneous server environment, with appropriate application software controls, multiple servers can access a single copy of data stored on a consolidated storage subsystem. If attached servers are heterogeneous platforms (for example, with a mix of UNIX, Linux, and Windows), sharing of data between such unlike operating system environments is more complex.

The SAN advantage in enabling enhanced data sharing can reduce the requirement to hold multiple copies of the same file or database. This reduces duplication of hardware costs to store such copies or the requirement to transfer copies of data between servers over the network. It also enhances the ability to implement cross enterprise applications, such as e-business, which might be inhibited when multiple data copies are stored.

Because it uses a specialized network usually based on Fibre Channel, the initial cost to implement a SAN is higher than for DAS and might be higher than for NAS. SANs require specialized hardware and software to manage the SAN and to provide many of its potential benefits.

Additionally, an organization must add new skills to manage this sophisticated technology. However, an analysis can justify the cost due to the long-term lower total cost of ownership compared to an alternate connectivity approach.

6.2.2 Types of SAN implementations

SANs are implemented for many basic reasons, or a combination of these, as described in the following sections.

Simplification

Environments that currently have no experience or limited experience with SANs are candidates for simplification or consolidation. Although this could conceivably include smaller environments that have only a few server types, it could also scale up to large enterprises which have numerous server types and storage dedicated to each through direct access connections.

By implementing a SAN infrastructure with one of these switches or directors, we can remove some of the complexity and increase the efficiency of the storage environment.

For example, in Figure 6-2, each server is using its own storage. Although they can communicate via TCP/IP for certain applications, the environment is complex and each storage system must be administered and maintained individually.

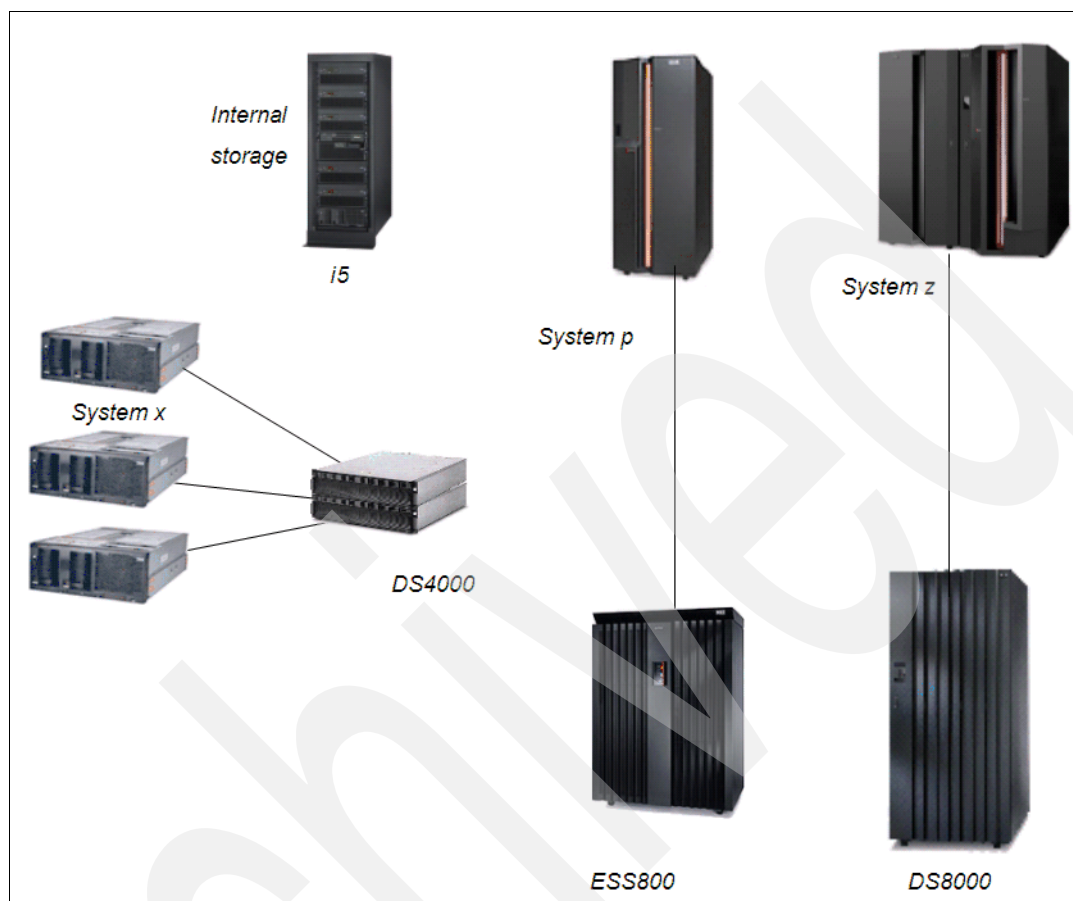


Figure 6-2 Multiple Server environments each with their own storage and no SAN

Moving on to the storage environment shown in Figure 6-3, we can see that it is greatly simplified. Rather than each storage environment existing as an island, we have one unified storage environment. Disk systems can be used for multiple server types depending on differences in performance and availability requirements, and management of the entire storage environment can be done more easily since they are all joined on the same network.

This infrastructure simplification can be the first step towards enabling your environment for a Business Continuity solution.

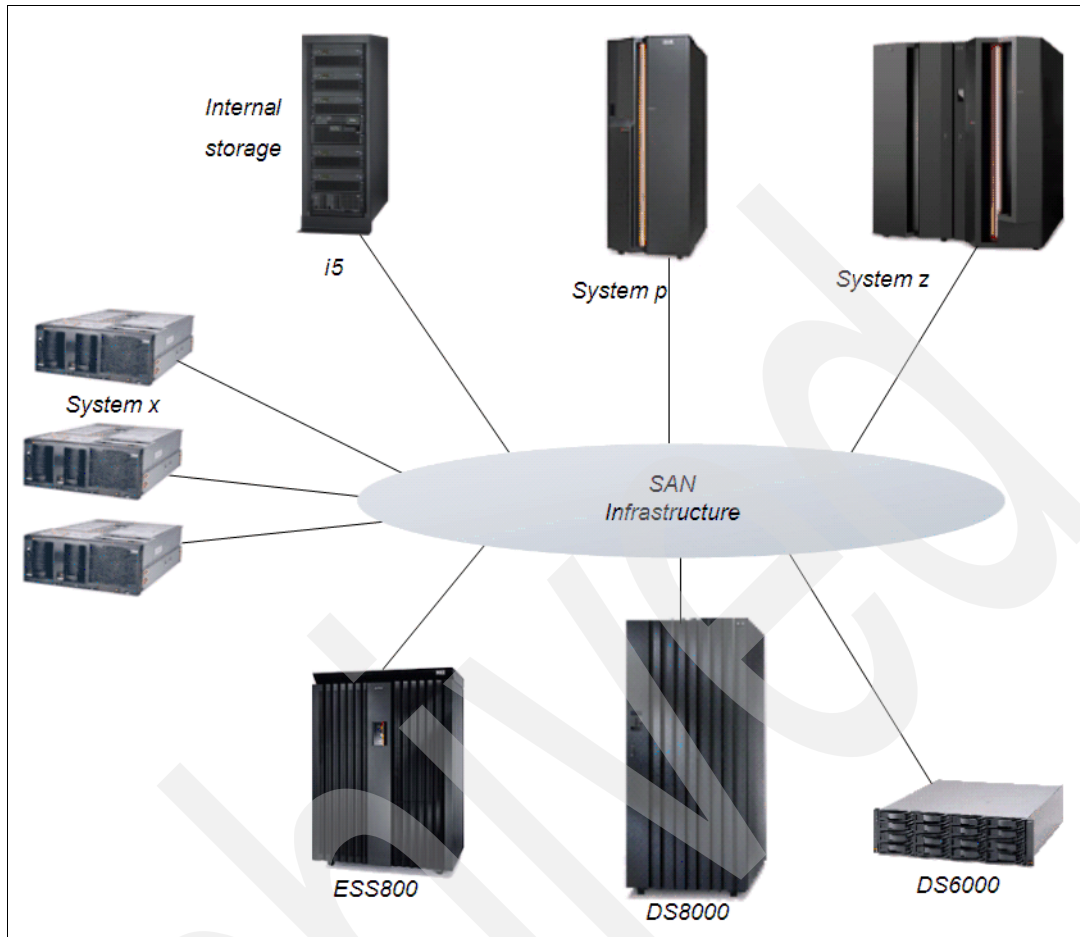


Figure 6-3 Simplified environment using a SAN infrastructure and IBM System Storage

Metro Continuous Availability

Environments that are interested in using their SAN for more than just managing and organizing their storage infrastructure within their data center, but who would also like to plan for storage networks and data access that span high speed Gigabit Ethernet (GbE), can make use of SAN equipment designed for Metro Area Solutions.

These solutions allow administrators to work on exploiting or creating their own Metropolitan Area Network (Figure 6-4). Once established, such networking can be used to enable a variety of Business Continuity technologies. These can include metropolitan area clusters (such as those provided in HACMP/XD as detailed in Chapter 2, “Continuous Availability” on page 9.

Additionally, they can be used for mirroring of data over IP connections. This could be using any form of replication including storage based block level mirroring or server based Logical Mirroring.

Note: Although Wavelength Division Multiplexing (WDM) can be used for SAN extension in the metropolitan area, it is covered in more detail in Chapter 8 Networking and inter-site connectivity options in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.

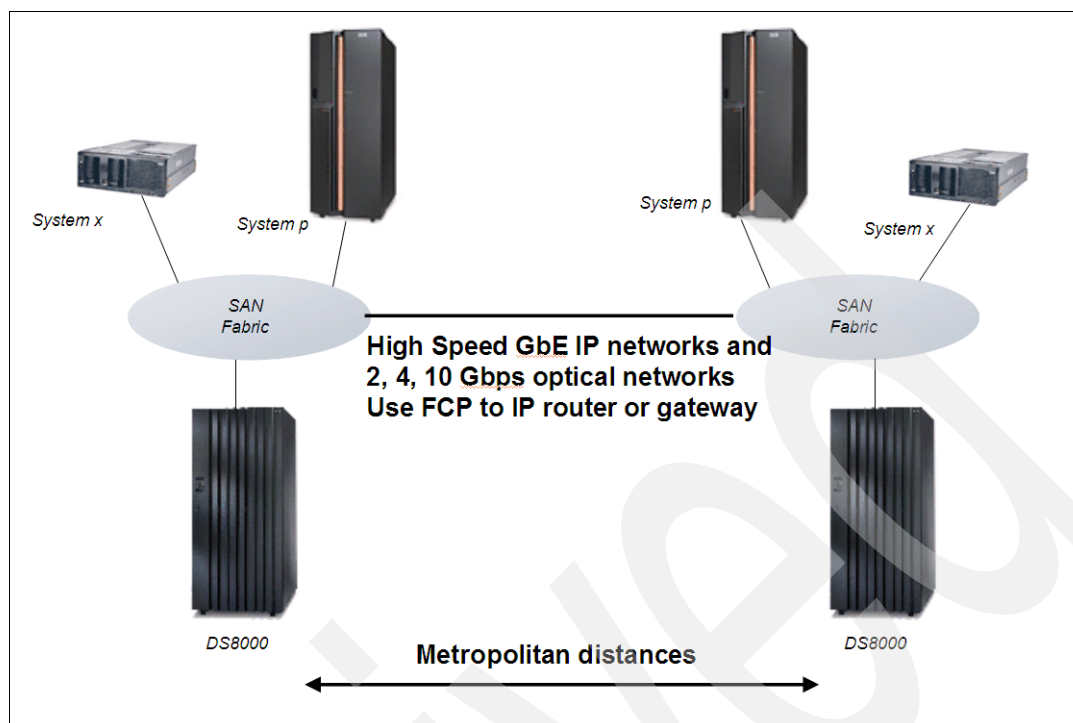


Figure 6-4 Metropolitan Area Continuous Availability SAN

Global Continuous Availability

This capability is for environments that require the capability to mirror data or access data over potentially any distance. Such environments use routers and directors capable of FCIP or iFCP in order to extend storage channel connections over TCP/IP networks to anywhere in the world.

6.2.3 Product portfolio

At the heart of a SAN infrastructure are the Fibre Channel switches or the Fibre Channel directors. Gateways and routers can be used to enable FCP traffic across IP networks at great distances.

IBM either OEMs or resells the products of the major vendors in this market. This gives the flexibility to use the best of breed products in a particular client situation.

Note: All noted models are those available at the time of this writing. Because of the changes occurring in this portion of the industry, we recommend that you check the Web sites indicated to make sure that the model you are interested in is still available, or to see what new models are available.

Entry level

These products are typically aimed at first time SAN buyers who require low cost and simplicity — usually with homogeneous, Windows, or Linux servers. IBM can offer IBM storage devices and software integrated into simple solutions at affordable prices with worldwide IBM and IBM Business Partner support. Figure 6-5 shows an overview of the entry switch portfolio.



Figure 6-5 Entry Level IBM Fibre Channel switch portfolio

IBM SAN 16M-2 and IBM SAN 16B-2

These are entry level switches from McDATA and Brocade, respectively. They can provide as few as 8 ports and up to 16 ports. Additionally, if more scalability is required, more switches can be added to the fabric.

They both support 4, 2, and 1 Gbps based on the compatibility of those devices.

In terms of continuance, these can be used to maintain high availability by using multiple HBAs on the servers and dual switch fabrics.

Cisco 9020

The Cisco 9020 is designed to handle traffic of 4, 2, or 1 Gbps in System x, p, or i environments. It is fully capable of creating higher availability environments by networking multiple switches together into a fabric.

IBM SAN 10Q-2

For System x only, the IBM SAN 10 Q-2, provides lower pricing with limited scalability. It provides a simple-to-use SAN infrastructure for IBM System x servers and IBM BladeCenter.

Midrange level

Midrange solutions provide additional capability, features, and benefits beyond the simple entry solutions. In enterprise installations, this type of switches are also often used at the edge of a core-edge-SAN, for example to tie in large numbers of servers without sacrificing expensive director ports (Figure 6-5).


IBM System Storage b-type (Brocade) switches	
 <p>IBM SAN32B-2 16, 24, 32 ports; 1, 2, 4 Gbps, FICON ibm.com/totalstorage/san/b-type</p>	 <p>IBM SAN64B-2 32, 48, 64 ports; 1, 2, 4 Gbps, ibm.com/totalstorage/san/b-type</p>
IBM System Storage m-type (McDATA) switches	
 <p>IBM SAN32M-2 16, 24, 32 ports; 1, 2, 4 Gbps, FICON ibm.com/totalstorage/san/m-type</p>	
Cisco MDS 9000 Fabric Switches	
 <p>Cisco MDS 9140 40 FC ports, 1, 2 Gbps www.ibm.com/storage/cisco</p>	 <p>Cisco MDS 9216A 16-64 FC ports: 2, 4 Gbps, integrate FICON, FCIP and iSCSI www.ibm.com/storage/cisco</p>

Figure 6-6 IBM System Storage Midrange SAN portfolio

Environments interested in this midrange tier of technology are cost conscious with limited technical skill who require affordable, scalable, and simple to manage solutions. Flexible pricing is also important. It can be offered with full fabric upgrade features and Ports On Demand features on the switches, which support a buy-and-grow strategy for the on demand business. Good performance is obviously required.

Higher availability switches are offered with dual, replaceable power supplies and hot code activation. Some of these clients can require more complex heterogeneous, mixed Windows and UNIX fabrics.

b-Type

Midrange computing solutions for b-Type SAN switches include two different Brocade models:

- ▶ **IBM SAN 32B-2** is a midrange computing switch that can be sold in port counts of 16, 24, or 32. It is capable of supporting any of the IBM System Storage environments including x, p, i, and even z (FICON) making it appropriate even for smaller System z environments. All required 4 Gbps SFPs must be specifically ordered in this model.
- ▶ **IBM SAN 32B-2 Express** is for midrange computing environments that do not require FICON support. In addition, it comes with the minimum number of active 4 Gbps SFPs (16).

- **IBM SAN 64B-2** is available with port counts of 32, 48, and 64 at 4 Gbps. It is suitable for environments with any IBM Server from System i, p, x, or z. In addition to functioning as a typical switch, it is capable of extending the fabric, by way of an “extended fabric” feature which allows switches to be connected at a speed of 4 Gbps at a range of up to 100km.

m-Type

With McDATA, we have the following offering:

- **IBM SAN 32M-2** supports 16, 24, or 32 ports at up to 4 Gbps in Windows, UNIX, Linux, NetWare, i5/OS, and z/OS environments. When combined with a routing infrastructure, it is possible to use this switch to mirror storage data across an IP network even at extended ranges if Longwave SFPs are in use.

Cisco

With Cisco, we have the following offering:

- **MDS9216A** comes with 16 FCP ports at 2 Gbps (expandable up to 64, at either 2 or 4Gbps). It can incorporate modules with additional FC ports, Gigabit Ethernet ports, and multiprotocol ports for FCIP, iSCSI, and FICON.

Routers

Routers differ from switches because, in addition to their ability to direct traffic between devices, they are capable of interconnecting dissimilar networks — that is, allowing networks made by Cisco, Brocade, and McDATA to coexist without merging into a single fabric. As such, as computing environments grow, routers can be useful when it comes to joining the “SAN Islands” together. Routers can usually act as gateways, allowing FCP traffic across TCP/IP networks.

As storage consolidation becomes more prevalent, routers are useful in allowing equipment from different vendors to coexist — as new fabric equipment gradually replaces older models that are removed from production (Figure 6-7).

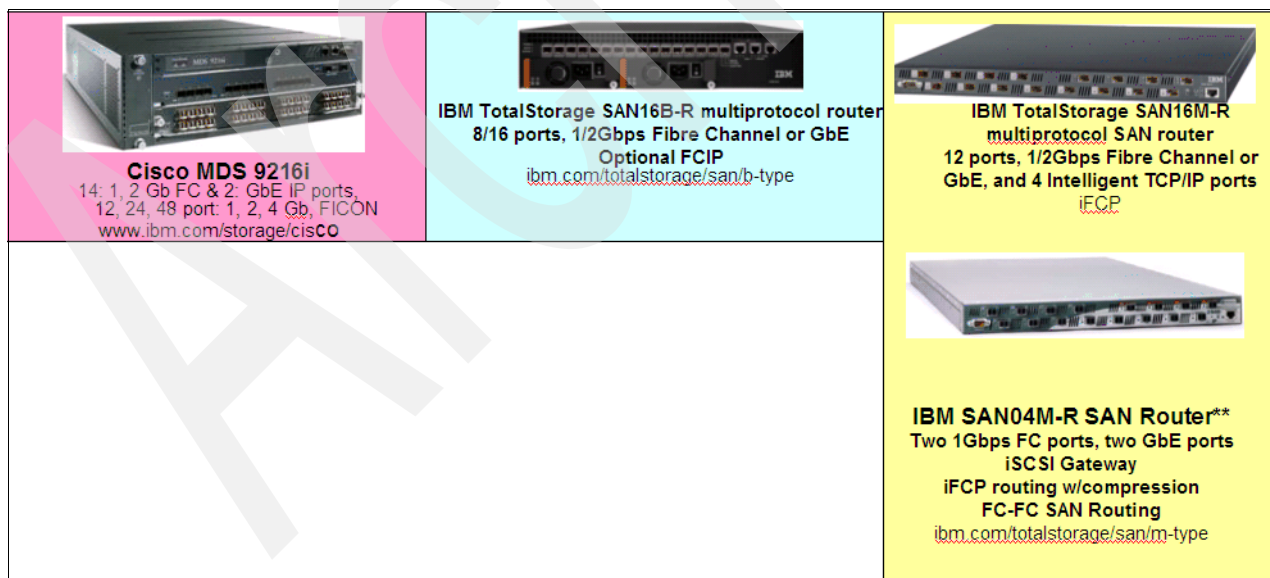


Figure 6-7 IBM System Storage Routers

Cisco

With Cisco, we have the following offering:

- ▶ **MDS9216i** comes with 14 FCP ports at 2 Gbps (expandable by up to 48 at 4 Gbps) and 2 Gigabit Ethernet ports at a speed of 4 Gbps for connectivity over IP or iSCSI remote networks. It also has Inter Switch Links (ISL) at 10 Gbps.

The MDS9216i can extend the SAN through Fibre Channel over Internet Protocol (FCIP) - this is useful Business Continuity environments where long distance disk mirroring is required.

The MDS9216i is compatible with environments using FCP, FICON, Gbe, or iSCSI.

Brocade

With Brocade, we have the following offering:

- ▶ **SAN18B-R** is the IBM Brocade Multiprotocol Router offering. It auto-detects and uses either 4, 2, or 1 Gbps FCP/FICON and GbE links as required. As with the Cisco 9216i, it can extend the SAN via FCIP for long distance disk mirroring.

McDATA

With McDATA, we have the following offerings:

- ▶ **SAN04M-R** comes with 2 FCP ports (1Gbps) and 2 Gigabit Ethernet ports.
- ▶ **SAN16M-R** comes with 12 FCP ports and 4 Gigabit Ethernet ports.

Unlike the routers by Brocade and Cisco, the McDATA routers make use of iFCP rather than FCIP. This is a subtle difference however the end goal is the same, and it provides support for extending the SAN and mirroring data at extended distances through this technology.

Enterprise level

An Enterprise SAN solution takes the midrange solution to a higher level of integration and value.

Many large enterprises have single supplier, homogeneous SAN islands. These typically separate departmental solutions in data centers, distribution centers, branch offices, and even in client and supply locations. There is a trend to integrate these SAN islands into an enterprise SAN for more advanced and manageable solutions.

iSCSI gateways and IP blades can connect hundreds of iSCSI servers to enterprise storage. Low cost Ethernet GbE adapters and IP switches enable many iSCSI servers to be attached to a small number of IP ports in the data center.

Multiprotocol IP and FC routers and blades use tri-rate SFP transceivers which support both IP and FC speeds.

The 4-Gbps switches are now common, and 10 Gbps blades are available for Inter Switch Links (ISLs).

Here are some enterprise director requirements for you to consider when creating enterprise solutions:

- ▶ High scalability up to 256 ports with logical partitions to isolate fabrics. This allows connectivity upgrades to support on demand business.
- ▶ High availability switch directors offered with built-in redundancy for concurrent repairs and upgrades without loss of performance.

- Enterprise clients who buy DS8000 and Enterprise tape demand the highest availability and scalability and value the advanced intelligence, which can help simplify the management of complex networks.
- In addition to heterogeneous Windows and UNIX servers, this space often includes System z FICON servers, too.

Figure 6-8 shows an overview of enterprise directors.

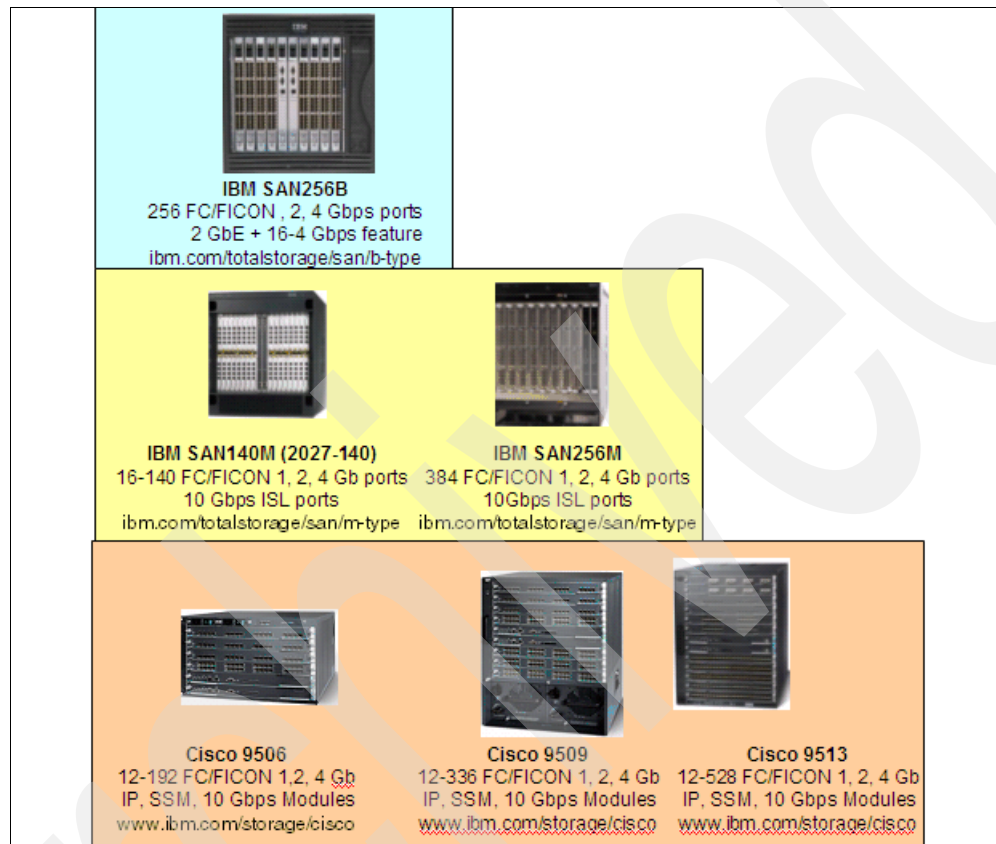


Figure 6-8 IBM System Storage Enterprise Directors

Brocade

With Brocade, we have the following offering:

- **SAN 256B** is the Brocade offering in the IBM System Storage Enterprise Directors. It is capable of handling up to 384 ports in a single domain with 16 or 32 ports per switch blade within the frame. Additionally, it can contain “router blades” with up to 16 FCP ports and 2 IP ports for SAN extension. The maximum speed of ports in the 256B is 4 Gbps, with 10 Gbps for ISLs.

Because it integrates routing into its frame, like the multiprotocol routers listed above, it is possible to merge SAN islands into a single fabric by using this director. Also, like the Brocade Multiprotocol Router, it can support FCIP in order to extend SANs and mirroring capabilities even to extended distances across an IP network.

McDATA

Note: Although McDATA has agreed to be purchased by Brocade at the time of writing, there is a commitment on the part of Brocade to continue supporting the McDATA Directors until a suitable replacement is available.

With McDATA, we have the following offerings:

- ▶ **SAN 256M** from McDATA scales from 64 to 256 ports, up to 4 Gbps with 10Gbps ISLs. Each of the switch modules in this Director can support up to 32 ports. This particular director does not support a router blade, however, so although the SAN can be extended by a distance of up to 1,000 km at up to 2 Gbps (high rates at shorter distances), it cannot make use of iFCP without an external router of some sort.
- ▶ **SAN 140M** is a smaller scale Enterprise Director from McDATA, which scales from 16 to 140 ports of FCP or FICON in open systems or System z environments.

Cisco

Note: All Cisco Directors also function as routers and as such can be used to join SAN islands into a single fabric

With Cisco, we have the following offerings:

- ▶ **MDS 9513** is an enterprise director with from 12 ports to 528 FCP or FICON ports. It can act as a router in order to join various SAN islands into one larger fabric. Modules for this director can come in 12, 24, or 48 port blades at 4 Gbps of bandwidth. Alternately, ISL blades are available with 4 ports at 10 Gbps.
As with other Cisco routers, the MDS 9513 can extend the network either through the ISLs or through Fibre Channel over IP (FCIP) in order to mirror data.
- ▶ **MDS 9506** is a smaller scale director for complex open systems environments. Although this router is not capable of handling FICON traffic, it works with FCIP, GbE, or iSCSI. It scales from 12 to 192 ports at 4, 2, or 1 Gbps.
As with other Cisco routers, it is capable of extending the SAN through its Gigabit Ethernet ports, either through iSCSI or FCIP.
Port modules can be 12, 24, or 48 ports.
- ▶ **MDS 9509** is a director for use in open systems environments supporting from 12 to 336 Fibre Channel ports at up to 4 Gbps. ISL ports are 10 Gbps. Port modules come in increments of 12, 24, or 48 ports.
This model director is, of course, capable of extending the network through its GbE ISLs for either iSCSI or FCIP connectivity.

6.2.4 Disaster recovery and backup considerations

The main considerations about SANs, regarding disaster recovery and backup procedures, are the benefits that a SAN can give to a company. In the following sections we talk about some areas where a company could have good results with a SAN. Some key aspects for Disaster Recovery and backup include:

- ▶ **Redundancy:**

The SAN itself delivers redundancy by providing multiple paths between servers furnished with multipathing software and storage. If a server or switch detects a failed connection, it can automatically route traffic through an alternate path without application awareness or downtime. When the problem is solved, the fallback, or restoration, of the original route can be similarly transparent. A well designed SAN's redundant connections also can be enhanced by *trunking*, which logically joins (in real time and when required) separate paths for higher throughput and automatic failover.

► **Distance:**

SANs facilitate robust, enterprise-wide remote mirroring by enabling copy strategies between servers and storage. Today a high-performance Fibre Channel switched fabric can connect multiple local remote volumes using a variety of campus area, metro area, or wide area configurations. Fibre Channel SANs optimize the backup process, delivering the infrastructure for functions such as LAN-free backup.

► **Recoverability:**

A key element of resilience is the time it takes to recover from small and large interruptions. A SAN delivers the infrastructure for data redundancy and distance. Such a tiered approach is immensely facilitated by the reach, connectivity, and management capabilities of a SAN. Manual invocation of such a tiered process would not only be costly and prone to errors further exacerbated by stress, but its invocation would not be as fast as the automated recovery enabled by a SAN.

SAN infrastructure supports, extends, speeds, and enables the following solutions for enterprise-class resilience before, during, and after disasters:

- Greater operational distances
- High availability clustering
- Alternate paths
- Synchronous mirroring
- Asynchronous mirroring
- High bandwidth
- Sophisticated manageable infrastructure
- Tape-device sharing
- Point-in-time copy between storage
- LAN-free, server-free and server-less tape back up
- Electronic tape vaulting
- Optimized tape restore
- Conventional tape restore

With data doubling every year, what effect does this have on the backup window? Backup to tape, and recovery, are operations that are problematic in the parallel SCSI or LAN based environments. For disk subsystems attached to specific servers, two options exist for tape backup. Either it must be done onto a server-attached tape subsystem, or by moving data across the LAN.

Providing tape drives to each server is costly, and also involves the added administrative overhead of scheduling the tasks, and managing the tape media. SANs allow for greater connectivity of tape drives and tape libraries, especially at greater distances. Tape pooling is the ability for more than one server to logically share tape drives within an automated library. This can be achieved by software management, using tools such as Tivoli Storage Manager, or with tape libraries with outboard management, such as the IBM 3494.

Backup using the LAN moves the administration to centralized tape drives or automated tape libraries. However, at the same time, the LAN experiences very high traffic volume during the backup or recovery operations, and this can be extremely disruptive to normal application access to the network. Although backups can be scheduled during non-peak periods, this might not allow sufficient time. Also, it might not be practical in an enterprise which operates in multiple time zones. SAN provides the solution, by enabling the elimination of backup and recovery data movement across the LAN. Fibre Channel's high bandwidth and multi-path switched fabric capabilities enables multiple servers to stream backup data concurrently to high speed tape drives. This frees the LAN for other application traffic.

LAN-free backup basics

The idea of *LAN-free backup* is to share the same backup devices across multiple backup clients directly over the SAN. With such a setup the client can perform backup directly to the backup device over a high speed SAN, compared to LAN backup, where all the traffic goes through one backup server. Utilizing LAN-free backup allows you to shorten backup windows, thus freeing up more time to perform other more important tasks.

This approach bypasses the network transport's latency and network protocol path length; therefore, it can offer improved backup and recovery speeds in cases where the network is the constraint. The data is read from the source device and written directly to the destination device.

Server-free backup basics

As in the LAN-free backup environment, SANs provide the infrastructure for *server-free backup*. Server-free backup is only possible in cases where the storage devices, such as disks and tapes, can talk to each other over a dedicated network. With server-free backup, the data is copied directly from SAN-attached disk to SAN-attached tape using SCSI-3 extended copy commands are used.

The overall advantage is the absence of backup I/O on the backup server as well as on the backup client. The data is transferred on a block-by-block basis and it is not under the control of the applications or databases.

Nevertheless, server-free backup has still not been widely deployed, mainly because of multiple hardware and software related dependencies. The advantage over LAN-free backup is very small, while LAN-free backup is a well accepted technique.

Server-less backup basics

By utilizing the any-to-any storage design, the SAN can be used to implement *server-less backup*. The SAN allows us to implement high speed data sharing across the various clients. With this we can share the data which has to be backed up among the production servers, and the servers performing backup tasks. In such a setup, the backup server can back up the production data, off-loading the processor usage from the production server, which is usually required to perform backup tasks.

SAN design considerations for Disaster Recovery

In the following sections we discuss the ideas of a dual SAN as well as the various distance options for a SAN.

It is important to understand midrange FC switch terminology when creating SAN solutions.

Business recovery solutions require longwave SFP transceivers that support up to 10 km (6 miles) and extended longwave SFP transceivers that support distances up to 80 km (48 miles). Single-mode fiber cables are required for longwave transceivers. These can be client provided for campus area networks (CAM) or telecom provided with leased *dark fiber* networks.

Because Business Continuity solutions require distances between sites, switches can be used to “fan-in” for device concentration to reduce the number of ISLs.

High availability designs require two ISLs for resiliency (loss of one ISL without loss of access).

Dual SAN

Consider that SAN switches and directors have different features and functions for availability. For more detailed information about the specific characteristics of each switch, refer to the IBM Redbook, *IBM TotalStorage: SAN Product, Design, and Optimization Guide*, SG24-6384 or to the IBM Web site:

<http://www.ibm.com/servers/storage/san/index.html>

For availability, the design of a dual SAN approach is the best way to start (Figure 6-9). We have two independent infrastructures that provide two independent paths of connection.

Note: Even when the physical paths are independent, you still have to check other dependencies of the components, dual path drivers, HBAs, switches, disk storage systems, tape, and so on. Of particular note, IBM does not recommend, or support, connecting both disk and tape devices to a single HBA.

Also, the support of different vendors' disk systems on the same server/HBA is normally not applicable because of the different device drivers. The IBM SAN Volume Controller is one possibility to reduce such dependencies.

Considering duplication of data for Disaster Recovery, we have two choices:

- ▶ Server-based / software-based data duplication using Volume Manager (such as AIX LVM, VERITAS Volume Manager) or applications-related replication features
- ▶ Storage system-based / hardware-based replication functions (storage system dependent)

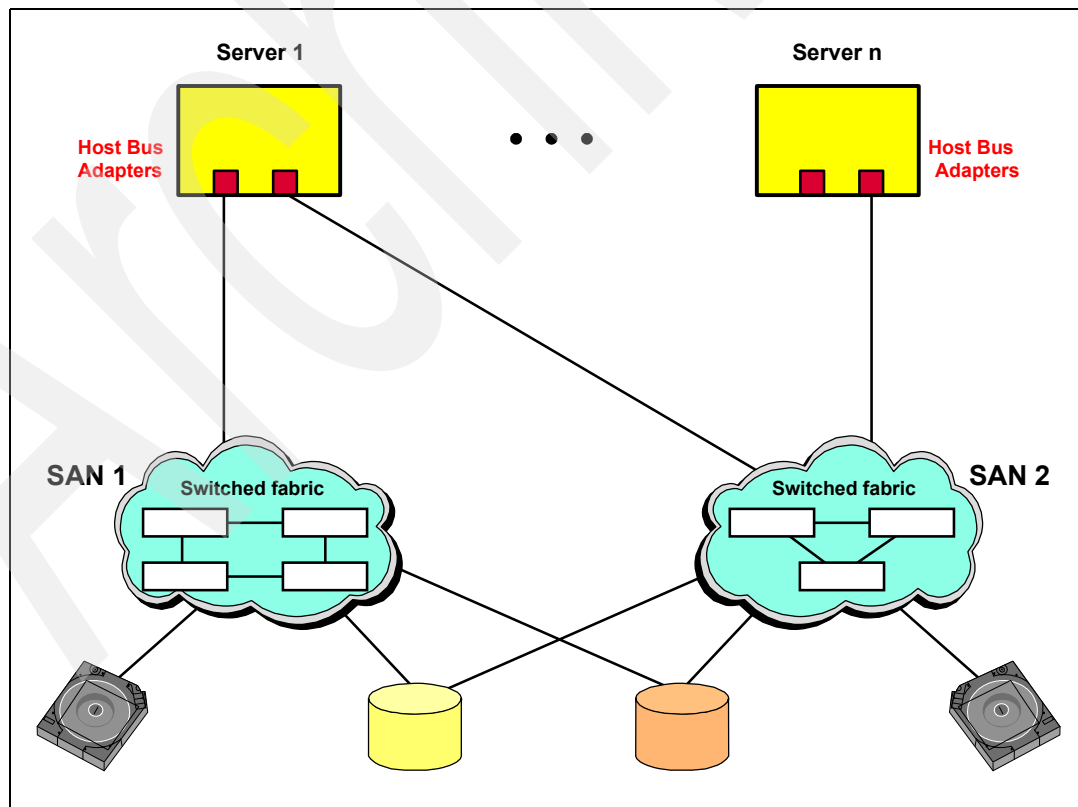


Figure 6-9 Dual SAN concept

SAN with intermediate distances

We have situations at client sites where we have several locations, but the distance between is either short (a few hundred metres) or intermediate (several kilometers). As there can be no public area between the sites, there is no problem to use an unrestricted amount of cables to connect these locations with each other.

If this is the case, we have a very flexible infrastructure for all methods of disaster recovery. The solutions can be set up easily dedicated to the requirements of different applications.

The solutions can include:

- ▶ Clusters spanning locations
- ▶ Replicating data on server levels (software based)
- ▶ Synchronous/asynchronous replication of data on the storage level (hardware based)

Note: If you start with a new SAN design today, multimode-cabling is sufficient within one location. The supported distance with 50 micron cables with 2 Gbps adapters is 300 m at 4 Gbps, it is reduced further to 150 m. So be aware that with the next generations of Fibre Channel, the distance for multimode will decrease again.

To be prepared for the near future, take this into consideration. Some single mode cabling might be a good choice, because the supported distance stays at 10 km.

SAN with long distances

There are two possibilities for using a SAN for long distances:

- ▶ With a dedicated dark fibre line, today we can go up to 150 km using DWDM (Dense Wave Division Multiplexer). Longer distances are possible via RPQ.
- ▶ Using channel extenders that interconnect Fibre Channel and TCP/IP networks, we can go with Fibre Channel over traditional networks.

To find which products are useful and proven by IBM for IBM disk-based replication, check the supported SAN extension products as listed in each disk system's interoperability matrix on the IBM Web site.

Be aware, that you must investigate a list of several different issues:

- ▶ Consider additional hardware such as optical repeaters to guarantee the signal intensity.
- ▶ Consider different vendors Fibre Channel transceivers (SFP - Small Form Factor Optical Transceivers) support different distances.
- ▶ Clarify the whole technical infrastructure.
- ▶ Consider distance support of the storage device.
- ▶ Consider distance support of the different switch vendors - including buffer credits available in the switches or directors; see *Introduction to SAN Distance Solutions*, SG24-6408 for more information.
- ▶ Consider data replication methods, including their dependencies and characteristics:
 - Storage based replication methods are one possibility.
 - Software based replication methods, such as: Logical Volume Manager, Application and operating system dependencies. Similar storage hardware probably has to be used even with software replication because of other dependencies, such as Multipath device driver issues.

Long distance considerations

Here is a list of questions to consider if you plan a SAN with DWDM over dark fibre or channel extenders over WAN:

- ▶ Which Telco/Service provider can provide the required bandwidth, the best service levels and of course the best price? Do they guarantee the service levels? Are they prepared for future growth?
- ▶ Does the Telco/Service provider really understand your disaster recovery requirements? Can they size, determine, and guarantee the appropriate bandwidth for your critical applications?

Note: Keep in mind that switching over networks is generally the longest delay in getting back up after a disaster occurs, especially in a complex environment. Be careful when you define the network switch over time in the contract with your service provider.

- ▶ Which network connectivity is the most cost effective in meeting your disaster recovery objectives (E1/E3, T1/T3, ATM, dark fiber, and so on)?
- ▶ What equipment is necessary (storage subsystems, Fibre Channel switches and directors, networking equipment, channel extenders or DWDM, encryption) and are they compatible with each other? What if you decide to change one of these components?
- ▶ If you decide to use replication / data mirroring, what are the restrictions for synchronous, asynchronous and cascading solutions?

Note: Be aware that you could have to go through a Request for Price Quotation (RPQ) process for getting some nonstandard configuration supported.

6.2.5 Additional information

For additional details about SAN topologies, design, and detailed component characteristics, refer to these IBM Redbooks:

- ▶ *Introduction to Storage Area Networks*, SG24-5470
- ▶ *Designing and Optimizing an IBM Storage Area Network*, SG24-6419
- ▶ *IBM TotalStorage: SAN Product, Design, and Optimization Guide*, SG24-6384
- ▶ *Introduction to SAN Distance Solutions*, SG24-6408
- ▶ *IBM SAN Survival Guide*, SG24-6143
- ▶ *IBM System Storage: Implementing an Open IBM SAN*, SG24-6116

You can also find further information by visiting the IBM Web site:

<http://www.ibm.com/servers/storage/san/>

6.3 Network Attached Storage (NAS)

This section discusses the Network Attached Storage (NAS) architecture, the products, and considerations for disaster recovery.

6.3.1 Overview

Storage devices that optimize the concept of file sharing across the network have come to be known as Network Attached Storage (NAS). NAS solutions utilize the mature Ethernet IP network technology of the LAN. Data is sent to and from NAS devices over the LAN using

TCP/IP. By making storage devices LAN addressable, the storage is freed from its direct attachment to a specific server and any-to-any connectivity is facilitated using the LAN fabric. In principle, any user running any operating system can access files on the remote storage device. This is done by means of a common network access protocol, for example, NFS for UNIX servers, and CIFS (SMB) for Windows servers.

A storage device cannot just attach to a LAN. It requires intelligence to manage the transfer and the organization of data on the device. The intelligence is provided by a dedicated server to which the common storage is attached. It is important to understand this concept. NAS comprises a server, an operating system (OS), plus storage which is shared across the network by many other servers and clients. So NAS is a device, rather than a network infrastructure, and shared storage is either internal to the NAS device or attached to it.

In contrast to the block I/O used by DAS and SAN, NAS I/O requests are called *file I/Os*. A file I/O is a higher-level type of request that, in essence, specifies the file to be accessed, an offset into the file (as though the file was a set of contiguous bytes), and a number of bytes to read or write beginning at that offset.

Unlike block I/O, there is no awareness of a disk volume or disk sectors in a file I/O request. Inside the NAS product (appliance), an operating system or operating system kernel tracks where files are located on disk, and issues a block I/O request to the disks to fulfill the file I/O read and write requests it receives.

Block I/O (raw disk) is handled differently. There is no OS format done to lay out a file system on the partition. The addressing scheme that keeps up with where data is stored is provided by the application using the partition. An example of this would be DB2 using its tables to keep track of where data is located rather than letting the OS do that job. That is not to say that DB2 cannot use the OS to keep track of where files are stored. It is just more efficient, for the database to bypass the cost of requesting the OS to do that work.

Using File I/O is like using an accountant. Accountants are good at keeping up with your money for you, but they charge you for that service. For your personal bank account, you probably want to avoid that cost. On the other hand, for a corporation where many different kinds of requests are made, having an accountant is a good idea. That way, checks are not written when they should not be.

An NAS appliance generally supports disk in an integrated package; tape drives can often be attached for backup purposes. In contrast to SAN devices that can usually also be direct-attached (such as by point-to-point Fibre Channel) as well as network-attached by SAN hubs and switches, an NAS device is generally only an NAS device and attaches only to processors over a LAN or WAN. NAS gateways, discussed later, offer some flexibility in combining NAS and SAN characteristics.

NAS has the following potential advantages:

- ▶ *Connectivity*. LAN implementation allows any-to-any connectivity across the network. NAS products can allow for concurrent attachment to multiple networks, thus supporting many users.
- ▶ *Resource pooling*. An NAS product enables disk storage capacity to be consolidated and pooled on a shared network resource, at great distances from the clients and servers that share it.
- ▶ *Exploitation of existing infrastructure*. Because NAS utilizes the existing LAN infrastructure, there are minimal costs of implementation.
- ▶ *Simplified implementation*. Because NAS devices attach to mature, standard LAN implementations, and have standard LAN addresses, they are typically extremely easy to install, operate, and administer.

- ▶ *Enhanced choice.* The storage decision is separated from the server decision, thus enabling the buyer to exercise more choice in selecting equipment to meet the business requirements.
- ▶ *Scalability.* NAS products can scale in capacity and performance within the allowed configuration limits of the individual system. However, this might be restricted by considerations such as LAN bandwidth constraints, and the requirement to avoid restricting other LAN traffic.
- ▶ *Heterogeneous file sharing.* Remote file sharing is one of the basic functions of any NAS product. Multiple client systems can have access to the same file. Access control is serialized by NFS or CIFS. Heterogeneous file sharing can be enabled by the provision of translation facilities between NFS and CIFS. File sharing allows you to reduce or eliminate data transfers of multiple copies on distributed hosts.
- ▶ *Improved manageability.* By providing consolidated storage, which supports multiple application systems, storage management is centralized. This enables a storage administrator to manage more capacity on a system than typically would be possible for distributed, directly attached storage.
- ▶ *Enhanced backup.* NAS system backup is a common feature of most popular backup software packages. Some NAS systems have also some integrated advanced backup and restore features which enable multiple point-in-time copies of files to be created on disk, which can be used to make backup copies to tape in the background.

Organizations can implement a mix of NAS, SAN, and DAS solutions; having one type of storage does not preclude you from using others.

NAS storage has the following characteristics:

- ▶ *Dedicated:* It is designed and pre-configured specifically for serving files. It performs no other application function.
- ▶ *Simple:* These appliances are typically pre-configured from the factory, making installation quick and easy.
- ▶ *Reliable:* High-level NAS storage has features such as hot swappable disk drives through RAID support, hot swappable power supplies, redundant disk drives and power supplies, and software that is *task-tuned* to minimize error potential.
- ▶ *Flexible:* It is heterogeneous. It often supports multiple file protocols. The most common file protocols supported by NAS appliances are NFS for UNIX, CIFS for Windows, NetWare for Novell, Apple File Protocol, HTTP and FTP.
- ▶ *Affordable:* Last but not least, NAS is affordable.

On the converse side of the storage network decision, you have to take into consideration the following factors regarding NAS solutions:

- ▶ *Proliferation of NAS devices.* Pooling of NAS resources can only occur within the capacity of the individual NAS system. As a result, in order to scale for capacity and performance, there is a tendency to grow the number of individual NAS systems over time, which can increase hardware and management costs.
- ▶ *Software overhead impacts performance.* TCP/IP is designed to bring data integrity to Ethernet-based networks by guaranteeing data movement from one place to another. The trade-off for reliability is a software intensive network design which requires significant processing overheads, which can consume more than 50% of available processor cycles when handling Ethernet connections. This is not normally an issue for applications such as Web-browsing, but it is a drawback for performance intensive storage applications.

- ▶ *Consumption of LAN bandwidth.* Ethernet LANs are tuned to favor short burst transmissions for rapid response to messaging requests, rather than large continuous data transmissions. Significant overhead can be imposed to move large blocks of data over the LAN. The maximum packet size for Ethernet is 1518 bytes. A 10 MB file has to be segmented into more than 7000 individual packets. Each packet is sent separately to the NAS device by the Ethernet collision detect access method. As a result, network congestion might lead to reduced or variable performance.
- ▶ *Data integrity.* The Ethernet protocols are designed for messaging applications, so data integrity is not of the highest priority. Data packets might be dropped without warning in a busy network, and have to be resent. Since it is up to the receiver to detect that a data packet has not arrived, and to request that it be resent, this can cause additional network traffic. With NFS file sharing there are some potential risks. Security controls can fairly easily be by-passed. This might be a concern for certain applications. Also the NFS file locking mechanism is not foolproof, so that multiple concurrent updates could occur in some situations.
- ▶ *Impact of backup/restore applications.* One of the potential downsides of NAS is the consumption of substantial amounts of LAN bandwidth during backup and restore operations, which might impact other user applications. NAS devices might not suit applications that require very high bandwidth. To overcome this limitation, some users implement a dedicated IP network for high data volume applications, in addition to the messaging IP network. This can add significantly to the cost of the NAS solution.

Building your own NAS

NAS devices support standard file access protocols such as NFS, CIFS, and sometimes others, that run over an IP network. These protocols were developed before dedicated NAS appliances existed, and are often implemented in software that runs on most client and server processors. So, in fact, anyone could build their own NAS device by taking a server of any size and installing NFS programming on it. The builder or integrator can use any disk products they want, even a single, internal disk for a small NAS, built using a low-cost desktop PC.

Building your own NAS means having flexibility in choosing the components and software that are used. But putting together your own NAS solution also has its downside. Here are some considerations when deciding between make or buy:

- ▶ The vendor's NAS appliance is tuned for storage-specific tasks, including the operating system and software pre-load, so performance is maximized as compared with that of a general purpose server.
- ▶ Much less time is required from IT resources to order and install an NAS solution than is required to purchase and assemble all of the necessary components and software themselves.
- ▶ You have assurance that the package works because the vendor has put it through a rigorous testing procedure, so IT resources do not get involved in de-bugging the system.
- ▶ Vendor support is provided for a wide range of auxiliary devices and software such as tape backup, SAN-attached storage, and backup software.
- ▶ The product is supported by the vendor via the Web, phone support, and an extensive service organization.
- ▶ The vendor provides a warranty for the product, thus creating financial protection for you.

NAS gateway

An NAS gateway provides the same functions and features as a conventional NAS system. It can use existing SAN storage resources instead of the integrated disk storage. This increases NAS scalability.

The disk storage is attached externally to the gateway, possibly sold separately, and can also be a stand-alone offering for direct or SAN attachment. The gateway accepts a file I/O request (using the NFS or CIFS protocols) and translates that to a SCSI block-I/O request to access the external attached disk storage. The gateway approach to file sharing offers the benefits of a conventional NAS system, with the additional potential advantages:

- ▶ You have an increased choice of disk types.
- ▶ There is an increased disk capacity scalability (compared to the capacity limits of most of the NAS appliances).
- ▶ It offers the ability to preserve and enhance the value of selected installed disk systems by adding file sharing.
- ▶ It includes the ability to offer file sharing and block-I/O on the same disk system.
- ▶ Disk capacity in the SAN can be shared (reassigned) among gateway and non-gateway use. So a gateway can be viewed as an NAS/SAN hybrid, increasing the flexibility and potentially lowering costs (versus capacity that might be under-utilized if it were permanently dedicated to an NAS appliance or to a SAN).

Figure 6-10 illustrates the two generic NAS configurations.

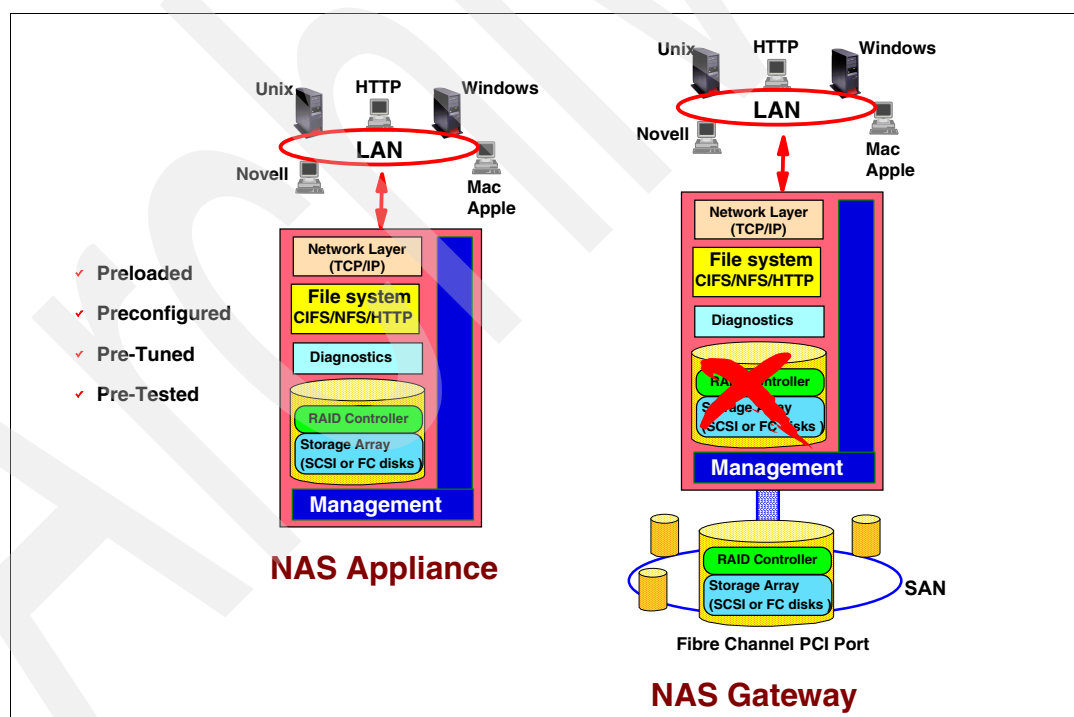


Figure 6-10 Network Attached Storage appliance and gateway solutions

More details on IBM NAS products are provided in Chapter 9, "IBM System Storage N series" on page 325.

6.4 iSCSI

Internet SCSI (iSCSI) is an IP-based standard for connecting hosts to data storage devices over a network and transferring data by carrying SCSI commands over IP networks. iSCSI supports Ethernet or Gigabit Ethernet interface at the physical layer, which allows systems supporting iSCSI interfaces to connect directly to standard Ethernet switches and IP routers.

When an operating system receives a request, it generates the SCSI command and then sends an IP packet over an Ethernet connection. At the receiving end, the SCSI commands are separated from the request, and the SCSI commands and data are sent to the SCSI controller and then to the SCSI storage device. iSCSI also returns a response to the request using the same protocol.

Some of the elements of iSCSI include:

- ▶ *Initiators.* These are the device drivers, that reside on the client, that route the SCSI commands over the IP network and provide access to the target device. These initiators drive the initiation of the SCSI request over TCP/IP. These initiators can either be software- or hardware-based. Dedicated iSCSI adapters can significantly improve performance and minimize the processor-overhead. The iSCSI device drivers for Windows versions can be downloaded from Microsoft, and IBM provides the AIX iSCSI driver.
- ▶ *Target software.* The target software receives SCSI commands from the IP network. It can also provide configuration support and storage management support.
- ▶ *Target hardware.* This can be a storage appliance that has imbedded storage in it as well as a gateway product.

iSCSI, Network File System (NFS), and Common Internet File System(CIFS) all allow storage access over the IP networking infrastructure. However, iSCSI enables block storage transfer, while NFS and CIFS transfer files. Typically, block level storage access offers superior performance for data intensive software applications and databases.

iSCSI devices can be attached directly using iSCSI-capable disk devices, or via an iSCSI gateway in front of non-iSCSI devices. The IBM iSCSI products offerings include:

- ▶ *IBM TotalStorage DS300.* This is an entry-level, cost effective iSCSI appliance for System x and BladeCenter servers and scales up to 4.2TB of physical storage capacity using new 300 GB Ultra320 SCSI drives.

See Chapter 10, “DS300 and DS400” on page 351.

- ▶ *IBM System Storage SAN18B-R multiprotocol router.*
For additional information, refer to 6.2.3, “Product portfolio”.
- ▶ *IBM System Storage SAN16M-R multiprotocol SAN router.*
For additional information, refer to 6.2.3, “Product portfolio”.
- ▶ *Cisco MDS series with IP Storage Services Module or Multiprotocol Services Module.*
For additional information, refer to 6.2.3, “Product portfolio”.

6.4.1 Business Continuity considerations

As we use TCP/IP networks, several possibilities come up with iSCSI as to where the data can be located. Depending on the application, the data can be accessed remotely, such as being replicated to two different locations.

iSCSI is a technology that promises to make long distance disaster recovery solutions more easily available, particularly to smaller companies. Being an IP network solution, its costs are typically lower than those related to other disaster recovery solutions. For example, if your primary site with an iSCSI target device is located in San Francisco, you could theoretically mirror all of your data to another iSCSI target device in Denver in real time. Or you could use the iSCSI gateways for realizing an electronic vaulting solution where two backed-up copies are made, one in the local site and the other in the remote site via existing IP networks.

The major technical issue that you would have to cross is getting a connection with enough bandwidth. Although iSCSI allows you to realize a disaster recovery solution with a reasonable budget, it can be very bandwidth intensive. Therefore you could require a high bandwidth and an expensive leased line. The amount of bandwidth that you require varies widely from company to company, depending on the amount of data that's being transmitted. Moreover in the electronic vaulting scenario the latency that accompanies the distance can limit the tape backup solution based on IP networks. The sequential nature of the tape backup I/O operations requires that the I/O for a single block must complete before the next block can be written. The greater the distance between the server and the tape drive, the longer this serial single-block operation takes.

6.4.2 Additional information

For more details, see the IBM Redbook, *SAN Multiprotocol Routing: An Introduction and Implementation*, SG24-7321, or visit the IBM Web site:

<http://www.ibm.com/servers/storage/>

6.5 Booting from the SAN

Many clients are considering booting directly from the SAN to avoid having internal hard drives in the servers and to extend the usage of the SAN infrastructure. This can sound attractive due to financial issues and logical from a technical view. Especially in a disaster tolerant infrastructure, robustness, ease of maintainability and reliable solutions are always preferred over highly sophisticated, but difficult to maintain implementations. Technically it is possible to boot most operating systems straight out of the SAN. However, we want to point out a few considerations to keep in mind before implementing this. We use an example of 10 servers that are directly SAN-attached via 2 Host Bus Adapters (HBAs) each.

Note that these considerations principally apply to Windows, UNIX, and Linux environments — mainframes have supported SAN (ESCON® and later FICON) booting for many years.

Operating system support

Check with the operating system vendor about support for booting from the SAN. Consider attaching the 10 servers to the SAN and enabling all of them. Expect that to have the following impact:

1. **Capacity:** Each server must have its own LUN for SAN Boot. This LUN must be capable of keeping the OS plus the adjacent Page Files. Assuming that a server requires 5 GB for the operating system itself and 20 GB for paging/swapping, then in our scenario we would implement 250 GB of SAN storage just to provide the platform for the basic OS installation on the servers.

2. **Performance:** In our example, with 10 servers, they each use their HBAs for both OS and paging I/O as well as for regular data access. Therefore the possibility of performance degradation is likely. Additional HBAs in the server might be an alternative (considering if the server can manage the number of adapters). However, we also have to consider that if we put 2 more adapters in each server, we would end up with an additional 20 more SAN switch ports.

All storage systems have a limited bandwidth. If this usage is shared across multiple partitions, in particular when operating system intensive I/O tasks such as paging and swapping are taken into account, then large amounts of performance bandwidth is used for these operating system purposes. Paging and swapping tend to fill up the storage system cache. Careful sizing and special cache usage settings for paging and swapping volumes in the storage systems are essential.

3. **Maintainability:** Especially when booting from the SAN, it is key to have a well documented SAN infrastructure and a team experienced in SAN management. Consider the consequences if one has to replace an HBA in one server and no one knows to which LUN it attaches or how the HBA should be customized for boot!

Serious considerations

With regard to booting from the SAN, there are still rumors that tell you how easy and simple it might be to avoid internal disk drives and boot from the SAN. Here are the most popular ideas and their dependencies:

- Easy server replacement:

Some users consider the advantage of booting from the SAN as a very simple way of server replacement in case the server hardware must be exchanged. However, this is only possible if the server hardware with all features such as BIOS and firmware matches the server systems to be replaced. Otherwise, unpredictable errors might occur, or the server simply might not work at all.

- Easy storage replacement:

Imagine you have the ten servers in our scenario attached and booting from a DS4700. Replacing the DS4500 with a DS8000, for example, would cause a reinstallation of all of the servers or would cause difficulties in the migration path.

Other considerations are with backup tasks, where you have to decide between LAN-free backup or network backup. In the case of LAN-free backup, it is quite likely that you will end up with additional HBA requirements just for the backup.

Summary

If technically possible, we think you should seriously consider using SAN boot in a disaster tolerant IT infrastructure, while keeping in mind the above concerns. The management of the SAN becomes increasingly critical and more complex as one starts to boot from the SAN. Very tight security is essential for the storage controllers and other SAN building blocks. Therefore the stakes become increasingly high.

Detailed documentation on the SAN infrastructure is essential — the full architecture description, including version information and all appropriate addressing details of the LUNs, adapters, ports, and cabling. Comprehensive documentation on the basic administration processes as well as the recovery/maintenance procedures is also highly advisable. Again, with SAN boot, one saves, in most cases, just the recommended three hard drives.



IBM System Storage DS6000, DS8000, and ESS

In this chapter we describe the enterprise disk systems available from IBM. We focus on the DS6000 and DS8000, and also provide some brief information about the Enterprise Storage System (ESS).

7.1 IBM System Storage DS6000 and DS8000

The IBM System Storage DS6000 and DS8000 are some of the key elements in the IT Business Continuity solutions described in this book.

The IBM TotalStorage Enterprise Storage Server (ESS 750/800) was the precursor for the DS6000 and DS8000, and much of the technology from the ESS has been re-used. Although it is no longer actively sold by IBM, the ESS remains widely deployed in IT environments.

- ▶ **Product names:** DS6000, DS8000
- ▶ **Tier level:** 7
- ▶ **Positioning with the System Storage Resiliency Portfolio:** Refer to Figure 7-1.

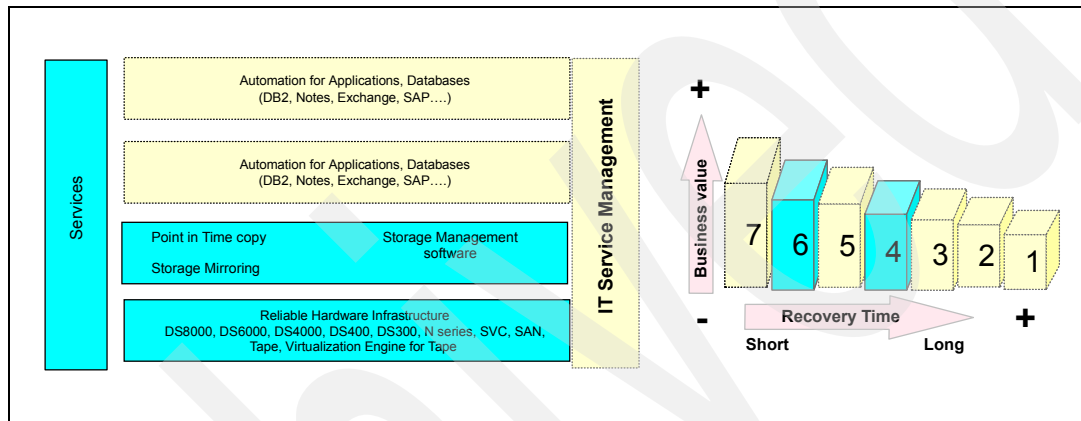


Figure 7-1 The positioning of the enterprise disk systems

- ▶ **Product descriptions:**

This chapter describes the IBM enterprise disk systems in the Hardware Infrastructure layer and how they relate to the components of the Core Technology layer.
- ▶ **Product highlights:**
 - IBM System Storage Hardware infrastructure
 - IBM System Storage Core technologies
- ▶ **Products and Advanced Copy Services:**
 - DS6000
 - DS8000
 - Supported Advanced Copy Services
 - Interfaces to configuration and Copy Services
 - Advanced Copy Services description

The System Storage continuum between the ESS, DS6000, and the DS8000 is shown in Figure 7-2.



Figure 7-2 A figure of the IBM System Storage continuum

7.2 The IBM System Storage DS6000 series

In this section we introduce the IBM System Storage DS6000 Series and its key features:

- ▶ Positioning
- ▶ Naming conventions
- ▶ Hardware
- ▶ Storage capacity
- ▶ Supported servers environment

7.2.1 Positioning

The DS6000 is an affordable enterprise class storage system that offers resiliency, performance and most of the key features of the IBM ESS in a rack mountable 3U-high unit. It is a lower cost alternative for a secondary remote mirror site or a test/development environment, which leverages the technical skill across the enterprise.

There is a single point of control with the IBM System Storage DS Storage Manager. The DS6000 is flexible and supports heterogeneous environments like IBM z/OS, IBM OS/400, IBM AIX, Linux, Microsoft Windows, HP-UX, SUN Solaris, and others.

7.2.2 DS6000 models

Table 7-1 summarizes the current DS6000 models. The 1750 522 and EX2 are the currently available controller and expansion frame. The 511 and EX1 are the originally announced models.

Table 7-1 DS6000 naming conventions

DS68000 controller	DS6000 expansion frame
Model 1750-511	Model 1750-EX1
Model 1750-522	Model 1750-EX2

7.2.3 Hardware overview

The DS6000 series consists of the DS6800, which has dual Fibre Channel RAID controllers with up to 16 disk drives in the enclosure. Capacity can be increased by adding up to eight DS6000 expansion enclosures.

Figure 7-3 shows the DS6800 disk system.



Figure 7-3 DS6800 disk system

DS6800 controller enclosure

IBM System Storage systems are based on a server architecture. At the core of the DS6800 controller unit are two active/active RAID controllers based on IBM industry leading PowerPC® architecture. By employing a server architecture with standard hardware components, IBM can always take advantage of best of breed components developed anywhere within IBM. Clients get the benefit of a very cost efficient and high performing storage system.

The DS6000 series controller's Licensed Internal Code (LIC) is based on the DS8000 series software, a greatly enhanced extension of the ESS software. 97% of the functional code of the DS6000 is identical to the DS8000 series, giving greater stability.

Table 7-2 summarizes the basic features of the DS6800.

Table 7-2 DS6000 hardware features

Feature	Value
Controllers	Dual Active
Max Cache	4 GB
Max Host Ports	8 Ports; 2Gbps FC/FICON
Max Hosts	1024
Max Storage / Disks	128 (up to 64 TB)
Disk Types	FC 10K: 146GB, 300 GB FC 15K: 73 GB, 146GB FATA 7.2K: 500GB
Max Expansion Modules	7
Max Device Ports	8

Feature	Value
Max LUNs	8192 (up to 2 TB LUN size)
RAID Levels	5, 10
RAID Array Sizes	4 or 8 drives
Operating Systems	Z/OS, i5/OS, OS/400, AIX, Solaris, HP-UAX, VMWare, Microsoft Windows, Linux
Packaging	3U - Controller and Expansion Drawers

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

We now give a short description of the main hardware components of the DS6000 Series:

- ▶ Processors
- ▶ Fibre Channel Arbitrated Loop connections to disks
- ▶ Disk drives
- ▶ Dense packaging
- ▶ Host adapters
- ▶ Expansion enclosures

The processors

The DS6800 utilizes two PowerPC processors for the storage server and the host adapters, respectively, and another PowerPC processor for the device adapter on each controller card. The DS6800 has 2 GB memory in each controller card for a total of 4 GB. Some part of the memory is used for the operating system and another part in each controller card acts as nonvolatile storage (NVS), but most of the memory is used as cache. This design to use processor memory makes cache accesses very fast.

Switched FC-AL subsystem

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

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

The disk drive module

The DS6800 controller unit can be equipped with up to 16 internal FC-AL disk drive modules, for up to 8 TB of physical storage capacity in only 3U (5.25") of standard 19" rack space.

Dense packaging

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

Host adapters

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

DS6000 expansion enclosure

The size and the front appearance of the DS6000 expansion enclosure (1750-EX1/EX2) is the same as the DS6800 controller enclosure. It can contain up to 16 disk drives.

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

Up to 7 DS6000 expansion enclosures can be added to a DS6800 controller enclosure. For connections to the previous and next enclosure, four inbound and four outbound 2 Gbps Fibre Channel ports are available.

7.2.4 Storage capacity

The minimum storage capability with eight 73 GB disk drive modules (DDMs) is 584 GB. The maximum storage capability with 16 500 GB DDMs for the DS6800 controller enclosure is 8 TB. To connect more than 16 disks, the DS6000 expansion enclosures allow a maximum of 128 DDMs per storage system and provide a maximum storage capability of 64 TB. Disk capacity can be added nondisruptively.

Every group of four or eight drives forms a RAID array that can be either RAID-5 or RAID-10. The configuration process requires that at least two spare drives are defined on each loop. In case of a disk drive failure or when the DS6000's predictive failure analysis comes to the conclusion that a disk drive might fail soon, the data on the failing disk is reconstructed on the spare disk.

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

Five disk types are available - all are Fibre Channel drives, except for the 500GB FATA drive:

- ▶ 73 GB 15k RPM
- ▶ 146 GB 10k RPM
- ▶ 146 GB 15k RPM
- ▶ 300 GB 10k RPM
- ▶ 500 GB 7.2k RPM

Therefore, a DS6800 can have from 584 GB (73 GB x 8 DDMs) up to 64 TB (500 GB x 128 DDMs).

Table 7-3 describes the capacity of the DS6800 with expansion enclosures.

Table 7-3 DS6800 physical capacity examples

Model	with 73 GB DDMs	with 146 GB DDMs	with 300 GB DDMs	with 500 GB DDMs
1750-522 (16 DDMs)	1.17 TB	2.34 TB	4.80 TB	7.50 TB
1750-522 + 3Exp (64 DDMs)	4.67 TB	9.34 TB	19.20 TB	32.00 TB
1750-522 + 7Exp (128 DDMs)	9.34 TB	18.69 TB	38.40 TB	64.20 TB

The DS6800 can have different types of DDMs in each enclosure (an intermixed configuration), except for 500GB FATA DDMs which cannot be intermixed with other types of DDMs in the same enclosure. For example, if the server enclosure has 73 GB DDMs and the expansion enclosure 1 has 300 GB DDMs, you can use the 73 GB DDMs for performance-related applications and the 300 GB DDMs for capacity-related applications.

Note: You cannot configure 500 GB DDMs with other different types of DDMs within an enclosure.

7.2.5 DS management console

The DS management console (MC) consists of the DS Storage Manager software, shipped with every DS6000 series system and a computer system to run the software. The DS6000 MC running the DS Storage Manager software is used to configure and manage DS6000 series systems. The software runs on a Windows or Linux system. This is not included with the DS6000 — it must be provided separately on a new or existing system. An additional MC can be provided for redundancy.

7.2.6 Supported servers environment

The DS6000 system can be connected across a broad range of server environments, including System Z, System p, System i, System x, and BladeCenter, as well as servers from Sun™, Hewlett-Packard, and other providers. You can easily split-up the DS6000 system storage capacity among the attached environments. This makes it an ideal system for storage consolidation in a dynamic and changing on demand environment.

For an up-to-date list of supported servers, refer to:

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

Highlights of the DS6000 series

The IBM System Storage DS6000 highlights include these:

- ▶ Two RAID controller cards
- ▶ Three PowerPC processors on each RAID controller card
- ▶ 4 GB of cache (2GB for each controller card)
- ▶ Two battery backup units (one for each controller card)
- ▶ Two AC/DC power supplies with imbedded enclosure cooling units
- ▶ Eight 2 Gbps device ports - for additional DS6000 expansion enclosures connectivity

- ▶ Two Fibre Channel switches for disk drive connectivity in each DS6000 series enclosure
- ▶ Eight Fibre Channel host ports that can be configured as pairs of FCP or FICON host ports. The host ports auto-negotiate to either 2 Gbps or 1 Gbps link speeds.
- ▶ Attachment of up to seven(7) DS6000 expansion enclosures
- ▶ Very small size, weight, and power consumption - all DS6000 series enclosures are 3U in height and mountable in a standard 19-inch rack

7.3 The IBM System Storage DS8000 series

In this section we introduce the IBM System Storage DS8000 Series and its key features:

- ▶ Positioning
- ▶ Naming convention
- ▶ Hardware
- ▶ Storage capacity
- ▶ Logical Partition
- ▶ Supported servers environment

7.3.1 Positioning

The IBM System Storage DS8000 is a high-performance, high-capacity series of disk storage systems, offering balanced performance that is up to 6 times higher than the previous IBM TotalStorage Enterprise Storage Server (ESS) Model 800. The capacity scales linearly from 1.1 TB up to 320TB.

The DS8000 supports heterogeneous environments like IBM z/OS, IBM OS/400, IBM AIX, Linux, UNIX, Microsoft Windows, HP-UX, SUN Solaris and others.

The DS8000 series is designed for 24x7 availability environments and provides industry leading remote mirror and copy functions to ensure business continuity.

The DS8000 supports a rich set of copy service functions and management tools that can be used to build solutions to help meet business continuance requirements. These include IBM Point-in-time Copy and Remote Mirror and Copy solutions.

Note: Remote Mirror and Copy was formerly known as Peer-to-Peer Remote Copy (PPRC).

You can manage Copy Services functions through the DS Command-Line Interface (CLI) called the IBM System Storage DS CLI and the Web-based interface called the IBM System Storage DS Storage Manager. The DS Storage Manager allows you to set up and manage the following data copy features from anywhere that network access is available.

7.3.2 DS8000 models

The DS8000 has six models: 921, 922, 9A2 and 931, 932, 9B2. The difference in models is the number of processors and the capability of storage system LPARs (see 7.3.5, "Storage system logical partitions (LPARs)"). You can also order expansion frames with the base frame. The expansion frame is either a model 92E or 9AE.

We summarize these specifications in Table 7-4.

Table 7-4 DS8000 models

	DS8100	DS8100 Turbo	DS8300	DS8300 Turbo	Expansion frame
9xy	Y=1	Y=1	Y=2	Y=2	Y=#
X=2	Non-LPAR 2-way		Non-LPAR 2-way		Non-LPAR Expansion
X=3		Non-LPAR 2-way		Non-LPAR 2-way	
X=A			LPAR 4-way		LPAR Expansion
X=B				LPAR 4-way	

The last position of the three characters indicates the number of 2-way processors on each processor complex (xx1 means 2-way and xx2 means 4-way, xxE means expansion frame (no processors) The middle position of the three characters means LPAR or non-LPAR model (x2x means non-LPAR model and xAx means LPAR model).

7.3.3 Hardware overview

Figure 7-4 shows a picture of the DS8000.



Figure 7-4 DS8000 disk system

The DS8000 hardware has been optimized for performance, connectivity, and reliability. Architecturally, the DS8000 series has commonality with the previous ESS models — 75% of the operating environment remains the same as for the ESS Model 800. This ensures that the DS8000 can leverage a very stable and well-proven operating environment, offering the optimum in availability, while incorporating superior performance levels.

Table 7-5 summarizes the current DS6000 hardware features.

Table 7-5 DS6000 hardware features

Feature	2-way	4-way
Server Processors	2-way Power5(+)	4-way Power5(+)
Cache	16 to 128 GB	32 to 256
Host Ports FICON (4GB/s) (4 ports per adapter) Fibre Channel (4GB/s) (4 ports per adapter) ESCON (2 ports per adapter)	8 to 64 8 to 64 4 to 32	8 to 128 8 to 128 8 to 64
Device Ports (4 ports per adapter)	8 to 32	8 to 64
Drives 73 GB (15K RPM) 146GB (10K 15K RPM) 300GB (10K RPM) 500GB (7.2K RPM)	16 to 384	16 to 640
Physical Capacity	1.2 to 192 TB	1.2 to 320TB
Number of Frames	1 to 2	1 to 3

In the following section we give a short description of the main hardware components:

- ▶ POWER5 (+) processor
- ▶ Internal fabric
- ▶ Switched Fibre Channel Arbitrated Loop connections to disks
- ▶ Disks available in different sizes and speeds
- ▶ 4Gb Host adapters
- ▶ Integrated management console

POWER5 processor technology

The DS8000 series exploits the IBM POWER5 technology, which is the foundation of the storage system LPARs. The DS8100 Model 921 uses 64-bit dual 2-way processor complexes and the DS8300 Model 922/9A2 uses 64-bit dual 4-way processor complexes. The POWER5 processors can have up to 256 GB of cache, which is up to 4 times as much as the ESS models.

Internal fabric

The DS8000 comes with a high bandwidth, fault tolerant internal interconnection, called the RIO2 (Remote I/O). It can operate at speeds up to 1 GHz and offers a 2 GB per second sustained bandwidth per link.

Switched Fibre Channel Arbitrated Loop (FC-AL)

The disk interconnection is switched FC-AL implementation. This offers a point-to-point connection to each drive and adapter, so that there are four paths available from the controllers to each disk drive.

Fibre Channel and FATA disk drives

The DS8000 offers a choice of Fibre Channel and FATA disk drives. There are 73 GB (15K RPM), 146 GB (10K and 15K RPM), 300 GB (10K RPM) FC and 500 GB (7.2K RMP) FATA disk drive modules (DDMs) available.

The 500 GB DDMs allow a single system to scale up to 320TB of capacity.

Host adapters

The DS8000 offers enhanced connectivity with the availability of four-port Fibre Channel/FICON host adapters. The 4 Gbps Fibre Channel/FICON host adapters, which are offered in long-wave and shortwave, can also auto-negotiate to 2 Gbps and 1 Gbps link speeds. This flexibility enables immediate exploitation of the benefits offered by the higher performance, 4 Gbps SAN-based solutions, while also maintaining compatibility with existing 2 Gbps and 1 Gbps infrastructures. In addition, the four adapter ports can be configured with an intermix of Fibre Channel Protocol (FCP) and FICON. This helps protect investment in fibre adapters, and provides the ability to migrate to new servers. The DS8000 also offers two-port ESCON adapters. A DS8000 can support up to a maximum of 32 host adapters, which provide up to 128 Fibre Channel/FICON ports.

Storage Hardware Management Console (S-HMC) for the DS8000

The DS8000 offers an integrated management console. This console is the service and configuration portal for up to eight DS8000s in the future. Initially there is one management console for one DS8000 storage subsystem. The S-HMC is the focal point for configuration and Copy Services management which can be done locally on the console keyboard or remotely via a Web browser.

Each DS8000 has an internal S-HMC in the base frame and you can have an external S-HMC for redundancy.

7.3.4 Storage capacity

The physical capacity for the DS8000 is purchased via disk drive sets. A disk drive set contains sixteen identical disk drives, which have the same capacity and the same revolution per minute (RPM). Disk drive sets are available in:

- ▶ 73 GB (15,000 RPM)
- ▶ 146 GB (10,000 RPM)
- ▶ 146 GB (15,000 RPM)
- ▶ 300 GB (10,000 RPM)
- ▶ 500GB (7,200 RPM)

Feature conversions are available to exchange existing disk drive sets, when purchasing new disk drive sets with higher capacity or higher speed disk drives.

In the first frame, there is space for a maximum of 128 disk drive modules (DDMs) and every expansion frame can contain 256 DDMs. So using the maximum of 640 500 GB drives gives a maximum capacity of 320 TB.

The DS8000 can be configured as RAID-5, RAID-10, or a combination of both. As a price/performance leader, RAID-5 offers excellent performance for many client applications, while RAID-10 can offer better performance for selected applications.

IBM Standby Capacity on Demand offering for the DS8000

Standby Capacity on Demand (Standby CoD) provides *standby* on demand storage for the DS8000, so that extra storage capacity can be accessed whenever the requirement arises. With Standby CoD, IBM installs up to 64 drives (in quantities of 16) in a DS8000. At any time, you can logically configure Standby CoD capacity for use, nondisruptively without requiring intervention from IBM. Upon logical configuration, you are charged for the capacity.

7.3.5 Storage system logical partitions (LPARs)

The DS8000 series provides *storage system* LPARs as a first in the industry. This means that you can run two completely segregated, independent, virtual storage images with differing workloads, and with different operating environments within a single physical DS8000 storage subsystem. The LPAR functionality is available in the DS8300 Model 9A2 and 9B2.

The LPAR implementation in the DS8000 is based on System p Virtualization Engine technology, which partitions the disk system into two virtual storage system images. So the processors, memory, adapters and disk drives are split between the images. There is a robust isolation between the two images via hardware and the POWER5 Hypervisor firmware.

Initially each storage system LPAR has access to:

- ▶ 50 percent of the processors
- ▶ 50 percent of the processor memory
- ▶ Up to 16 host adapters
- ▶ Up to 320 disk drives (up to 160TB of capacity)

With these separate resources, each storage system LPAR can run the same or different versions of microcode, and can be used for completely separate production, test, or other unique storage environments within this single physical system. This can enable storage consolidations, where separate storage subsystems were previously required, helping to increase management efficiency and cost effectiveness.

Supported environments

The DS8000 series offers connectivity support across a broad range of server environments, including IBM z/OS, IBM OS/400, IBM AIX, Linux, UNIX, Microsoft Windows, HP-UX, SUN Solaris and others — over 90 supported platforms in all. This rich support of heterogeneous environments and attachments, along with the flexibility to easily partition the DS8000 series storage capacity among the attached environments, can help support storage consolidation requirements and dynamic, changing environments.

For an up-to-date list of supported servers, go to:

<http://www.ibm.com/servers/storage/disk/ds8000/interop.html>

Highlights of DS8000 series:

The IBM System Storage DS8000 highlights include these:

- ▶ Delivers robust, flexible, and cost-effective disk storage for mission-critical workloads
- ▶ Helps to ensure exceptionally high system availability for continuous operations

- ▶ Scales to 320 TB and facilitates unprecedented asset protection with model-to-model field upgrades
- ▶ Supports storage sharing and consolidation for a wide variety of operating systems and mixed server environments
- ▶ Helps increase storage administration productivity with centralized and simplified management
- ▶ Provides the creation of multiple storage system LPARs, that can be used for completely separate production, test, or other unique storage environments
- ▶ Provides the industry's first four year warranty

7.4 The IBM TotalStorage ESS 800

Here we briefly describe the IBM TotalStorage ESS 800 and its key features:

- ▶ Naming conventions
- ▶ Hardware
- ▶ Storage capacity
- ▶ Supported servers environment

7.4.1 ESS800 models

Table 7-6 summarizes the ESS models.

Table 7-6 ESS models

ESS 800 controller	ESS 800 expansion frame	ESS 750 controller
Model 2105-800	Model 2105-800 feature 2110	Model 2105-750

7.4.2 Hardware overview

Historically, the ESS offered exceptional performance, extraordinary capacity, scalability, heterogeneous server connectivity, and an extensive suite of advanced functions to support mission-critical, high-availability, multi-platform environments. The ESS set a new standard for storage servers back in 1999 when it was first available from IBM, and it subsequently evolved into the F models and finally the third-generation ESS Model 750 and Model 800, keeping up with the pace of the clients' requirements by adding more sophisticated functions to the initial set, enhancing the connectivity options, and enhancing performance.

The main ESS family hardware components are:

- ▶ Processor technology
- ▶ SSA loop connection to disks
- ▶ SSA attached disks
- ▶ Host adapter
- ▶ ESS master console

Processor technology

Each of the two active clusters within the ESS contains fast symmetrical multiprocessor (SMP). There are two processor options available:

- ▶ Standard processor feature
- ▶ Turbo processor feature (model 800 only)

Model 750 supports a 2-way processor, while the Model 800 supports a 4-way or 6-way processor (turbo feature).

SSA loop connection to disks

The ESS uses Serial Storage Architecture (SSA) for interconnection to the drives.

SSA attached disk

The ESS provides integrated caching and support for the attached disk drive modules (DDMs). The DDMs are attached through an SSA interface. Disk storage on an ESS is available in modules that contain eight DDMs — these are called disk eight packs.

Model 750 supports 72.8 GB and 145.6 GB 10 000 RPM drives, that can be intermixed and configured as RAID-5 or RAID-10.

Model 800 supports 18.2, 36.4, 72.8 GB disks at 10000 RPM and at 15000 RPM. In addition, it supports 145.6 GB DDMs at 10000 RPM.

Host adapters

The Model 800 supports a maximum of 16 host adapters. These can be an intermix of the following host adapter types and protocols:

- ▶ SCSI adapters
- ▶ Fibre Channel adapters, for support of Fibre Channel protocol (FCP) and fibre connection (FICON) protocol
- ▶ Enterprise Systems Connection Architecture® (ESCON) adapters

The Model 750 supports a maximum of 6 host adapters. These can be an intermix of the following host adapter types and protocols:

- ▶ Fibre Channel adapters, for support of FCP and FICON protocol
- ▶ ESCON adapters

Storage capacity

The physical capacity for the ESS family is via disk drive sets. A disk drive set contains 8 identical disk drives, which have the same capacity and the same revolution per minute (RPM). Disk drive sets are:

- ▶ For Model 750:
 - 73 GB (10,000 RPM)
 - 146 GB (10,000 RPM)
- ▶ For Model 800:
 - 18 GB (10,000 and 15,000 RPM)
 - 37 GB (10,000 and 15,000 RPM)
 - 73 GB (10,000 and 15,000 RPM)
 - 146 GB (10,000 RPM)

The ESS can be configured as RAID-5, RAID-10, or a combination of both.

ESS Model 750 supports up to eight disk eight-packs and up to 4.659 TB of physical capacity.

ESS Model 800, with an expansion enclosure, can provide the following data storage capacity:

- ▶ With 18.2 GB homogeneous DDMs, the maximum capacity is 7.06 TB.
- ▶ With 36.4 GB homogeneous DDMs, the maximum capacity is 14.13 TB.

- ▶ With 72.8 GB homogeneous DDMs, the maximum capacity is 28.26 TB.
- ▶ With 145.6 GB homogeneous DDMs, the Model 800 supports a maximum capacity of 55.9 TB.

7.4.3 Supported servers environment

The ESS can be connected to a broad range of server environments, including IBM z/OS, IBM OS/400, IBM AIX, Linux, UNIX, Microsoft Windows, HP-UX, SUN Solaris and others. There are over 90 supported platforms.

For supported servers for Model 800, go to:

<http://www.ibm.com/servers/storage/disk/ess/ess800/interop-matrix.html>

For supported servers for Model 750, go to:

<http://www.ibm.com/servers/storage/disk/ess/ess750/interop-matrix.html>

7.5 Advanced Copy Services for DS8000/DS6000 and ESS

Table 7-7 compares the available copy services for the DS8000/DS6000 and ESS.

Table 7-7 Advanced Copy Services feature table

	DS6000 (1)	ESS Model 800/750	DS8000 (1)
FlashCopy	yes	yes	yes
Metro Mirror	yes	yes	yes
Global Copy	yes	yes	yes
Global Mirror	yes	yes	yes
Metro/Global Copy	yes	yes	yes
Metro/Global Mirror	no (4)	yes (2)	yes
z/OS Global Mirror	yes (3)	yes	yes

Note 1: Remote Mirror and Copy functions can also be established between DS6800 and ESS 800/750 systems, but not directly between a DS6800 and older ESS models like the F20, because these models do not support remote mirror or copy across Fibre Channel (they only support ESCON) and the DS6800 does not support ESCON.

The DS Series does not support Metro Mirror over ESCON channels.

Note 2: An RPQ is available for ESS to participate in a Metro/Global Mirror configuration. ESS in a Metro/Global Mirror configuration is not able to support the Metro/Global Mirror Incremental Resync function, which is only available on DS8000 series systems.

Note 3: The DS6000 can only be used as a target system in z/OS Global Mirror operations.

Note 4: The DS6000 is not currently supported in a Metro/Global Mirror configuration.

IBM System Storage FlashCopy

FlashCopy can help reduce or eliminate planned outages for critical applications. FlashCopy provides a point-in-time copy capability for logical volumes. FlashCopy supports many advanced capabilities, including these:

- ▶ **Data Set FlashCopy:**
Data Set FlashCopy allows a FlashCopy of a data set in a System z environment.
- ▶ **Multiple Relationship FlashCopy:**
Multiple Relationship FlashCopy allows a source volume to have multiple targets simultaneously.
- ▶ **Incremental FlashCopy:**
Incremental FlashCopy provides the capability to update a FlashCopy target without having to recopy the entire volume.
- ▶ **FlashCopy to a Remote Mirror primary:**
FlashCopy to a Remote Mirror primary gives you the possibility to use a FlashCopy target volume also as a remote mirror primary volume. This process allows you to create a point-in-time copy and then make a copy of that data at a remote site.
- ▶ **Consistency Group commands:**
Consistency Group commands allow the storage subsystems to hold off I/O activity to a LUN or group of LUNs until the FlashCopy Consistency Group command is issued. Consistency Groups can be used to help create a consistent point-in-time copy across multiple LUNs, and even across multiple storage subsystems.
- ▶ **Inband Commands over Remote Mirror link:**
In a remote mirror environment, commands to manage FlashCopy at the remote site can be issued from the local or intermediate site and transmitted over the remote mirror Fibre Channel links. This eliminates the requirement for a network connection to the remote site solely for the management of FlashCopy.

IBM System Storage Metro Mirror (Synchronous PPRC)

Metro Mirror is a remote data-mirroring technique for all supported servers, including z/OS and open systems. Its function is to constantly maintain an up-to-date copy of the local application data at a remote site which is within the metropolitan area — typically up to 300 km away using Dense Wavelength Division Multiplexing (DWDM). Greater distances are supported on special request. With synchronous mirroring techniques, data currency is maintained between sites, though the distance can have some impact on performance. Metro Mirror is used primarily as part of a business continuance solution for protecting data against disk storage system loss or complete site failure.

IBM System Storage Global Copy (PPRC Extended Distance, PPRC-XD)

Global Copy is an asynchronous remote copy function for z/OS and open systems for longer distances than are possible with Metro Mirror. With Global Copy, write operations complete on the primary storage system before they are received by the secondary storage system. This capability is designed to prevent the primary system's performance from being affected by wait-time from writes on the secondary system. Therefore, the primary and secondary copies can be separated by any distance. This function is appropriate for remote data migration, off-site backups, and transmission of inactive database logs at virtually unlimited distances.

IBM System Storage Global Mirror (Asynchronous PPRC)

Global Mirror copying provides a two-site extended distance remote mirroring function for z/OS and open systems servers. With Global Mirror, the data that the host writes to the storage unit at the local site is asynchronously shadowed to the storage unit at the remote site. A consistent copy of the data is then automatically maintained on the storage unit at the remote site. This two site data mirroring function is designed to provide a high-performance, cost-effective global distance data replication and disaster recovery solution.

IBM System Storage Metro/Global Copy

Metro/Global Copy is a cascaded three-site disk mirroring solution. Metro Mirror is used between production site A and intermediate site B. Global Copy is used between the intermediate site B and the remote site C. Metro/Global Copy is often used for migration purposes in a two site mirroring environment. Global Copy keeps the cascaded copy (which can be located at either the remote or local site) nearly current with the running two-site disk mirroring configuration.

IBM System Storage Metro/Global Mirror

Metro/Global Mirror is a cascaded three-site disk mirroring solution. Metro Mirror is used between production site A and intermediate site B. Global Mirror is used between the intermediate site B and the remote site C. In the event of a loss of access to intermediate site B, the license for DS8000 Metro/Global Mirror provides new functionality to incrementally resynchronize and establish Global Mirror from production site A to remote site C, without application impact to site A — thus maintaining out of region disaster recovery. When intermediate site B access returns, that site can be re-inserted into the three site cascading topology without impact to production site A applications. In all cases, only incremental changes have to be sent for resynchronization.

IBM System Storage z/OS Global Mirror (Extended Remote Copy XRC)

z/OS Global Mirror offers a very specific set of very high scalability and high performance asynchronous mirroring capabilities designed to match very demanding, very large System z resiliency requirements.

7.6 Introduction to Copy Services

Copy Services is a collection of functions that provide disaster recovery, data migration, and data duplication functions. With the Copy Services functions, for example, can create backup data with little or no application disruption, and back up application data to a remote site for the disaster recovery.

Copy Services run on the DS8000/DS6000 and support open systems and System z environments. These functions are supported also on the previous ESS systems.

The licensed features included in Copy Services are:

- ▶ FlashCopy, which is a Point-in-Time Copy function
- ▶ Remote Mirror and Copy functions, previously known as Peer-to-Peer Remote Copy or PPRC, which include:
 - IBM Metro Mirror, previously known as Synchronous PPRC
 - IBM Global Copy, previously known as PPRC Extended Distance
 - IBM Global Mirror, previously known as Asynchronous PPRC
- ▶ z/OS Global Mirror, previously known as Extended Remote Copy (XRC)
- ▶ Metro/Global Copy

- ▶ Metro/Global Mirror
- ▶ z/OS Metro/Global Mirror

We explain these functions in detail in the next section.

Management of the Copy Services functions is through a command-line interface (DS CLI), a Web-based interface (DS Storage Manager), or TotalStorage Productivity Center for Replication. You also can manage the Copy Services functions through the open application programming interface (DS Open API). Copy Services functions in System z environments can be invoked by TSO commands, ICKDSF, the DFSMSdss utility, and so on.

We explain these interfaces in 7.8, “Interfaces for Copy Services” on page 291.

7.7 Copy Services functions

We describe each function and the architecture of the Copy Services in this section. There are two primary types of Copy Services functions: *Point-in-Time Copy* and *Remote Mirror and Copy*. Generally, the Point-in-Time Copy function is used for data duplication and the Remote Mirror and Copy function is used for data migration and disaster recovery.

7.7.1 Point-In-Time Copy (FlashCopy)

The Point-in-Time Copy feature, which includes FlashCopy, creates full volume copies of data in a storage unit. A FlashCopy establishes a relationship between the source and target volumes, and creates a bitmap of the source volume. Once this relationship and bitmap are created, the target volume can be accessed as though all the data had been physically copied. While a relationship between the source and target volume exists, optionally, a background process copies the tracks from the source to the target volume.

Note: In this chapter, *track* means a piece of data in the DS8000/DS6000; the DS8000/DS6000 uses the logical tracks to manage the Copy Services functions.

See Figure 7-5 for an illustration of FlashCopy concepts.

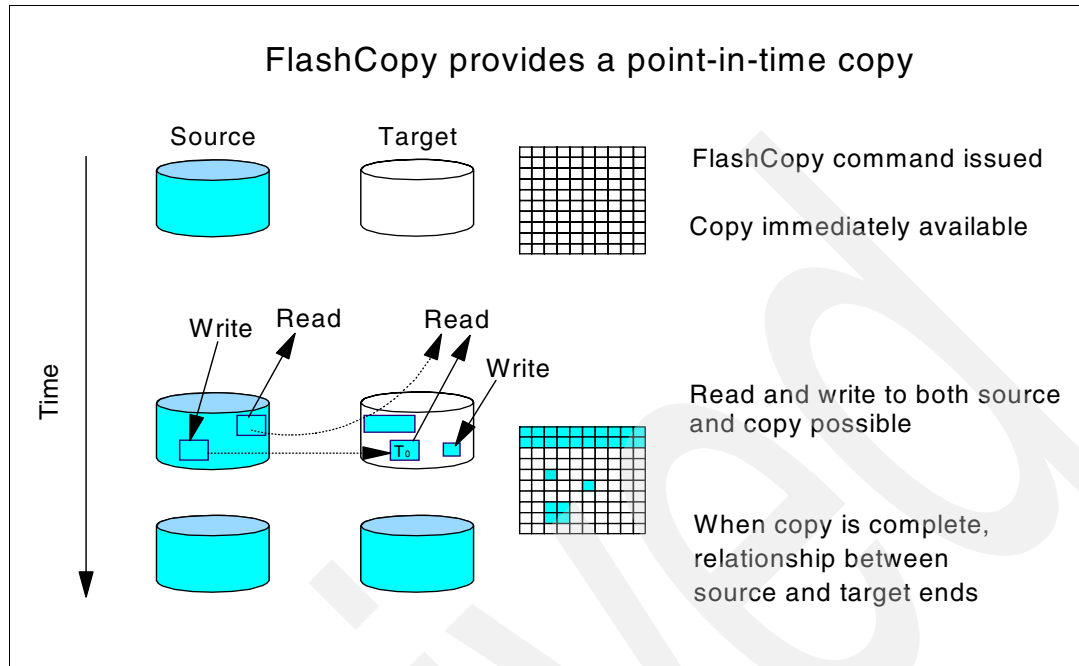


Figure 7-5 FlashCopy concepts

When a FlashCopy operation is invoked, it takes only from within a second to a few seconds to complete the process of establishing the FlashCopy pair and creating the necessary control bitmaps. Thereafter, a point-in-time copy of the source volume is available. As soon as the pair has been established, both the source and target volumes are available for read/write access.

After creating the bitmap, a background process begins to copy the real-data from the source to the target volumes. If you access the source or the target volumes during the background copy, FlashCopy manages these I/O requests as follows:

► **Read from the source volume:**

When you read some data from the source volume, it is simply read from the source volume.

► **Read from the target volume:**

When you read some data from the target volume, FlashCopy checks the bitmap and:

- If the backup data is already copied to the target volume, it is read from the target volume.
- If the backup data is not copied yet, it is read from the source volume.

► **Write to the source volume:**

When you write some data to the source volume, at first the updated data is written to the data cache and persistent memory (write cache). And when the updated data is destaged to the source volume, FlashCopy checks the bitmap and:

- If the backup data is already copied, it is simply updated on the source volume.
- If the backup data is not copied yet, first the backup data is copied to the target volume, and after that it is updated on the source volume.

► **Write to the target volume:**

When you write some data to the target volume, it is written to the data cache and persistent memory, and FlashCopy manages the bitmaps to not overwrite the latest data. FlashCopy does not overwrite the latest data by the physical copy.

The background copy might have a slight impact to your application because the physical copy requires some storage resources, but the impact is minimal because the host I/O is prior to the background copy. And if you want, you can issue FlashCopy with the *no background copy* option.

No background copy option

If you invoke FlashCopy with the no background copy option, the FlashCopy relationship is established without initiating a background copy. Therefore, you can minimize the impact of the background copy. When the DS8000/DS6000 receives an update to a source track in a FlashCopy relationship, a copy of the point-in-time data is copied to the target volume so that it is available when the data from the target volume is accessed. This option is useful if you do not have to issue FlashCopy in the opposite direction.

Benefits of FlashCopy

FlashCopy is typically used where a copy of the production data is required with little or no application downtime (depending on the application). It can be used for online backup, testing of new applications, or for creating a database for data-mining purposes. The copy looks exactly like the original source volume and is an instantly available, binary copy.

Point-in-Time Copy function authorization

FlashCopy is an optional function. To use it, you must purchase the Point-in-Time Copy 2244 function authorization model, which is 2244 Model PTC.

7.7.2 FlashCopy options

FlashCopy has many options and expanded functions to help provide data duplication. We explain these options and functions in this section.

Refresh target volume (also known as Incremental FlashCopy)

Refresh target volume provides the ability to *refresh* a LUN or volume involved in a FlashCopy relationship. When a subsequent FlashCopy operation is initiated, only the tracks changed on both the source and target have to be copied from the source to the target. The direction of the *refresh* can also be reversed.

Typically, at most, 10 to 20 percent of data is changed in a day. Therefore, if you use this function for daily backup, you can save the time for the physical copy of FlashCopy.

Figure 7-6 explains the architecture for Incremental FlashCopy.

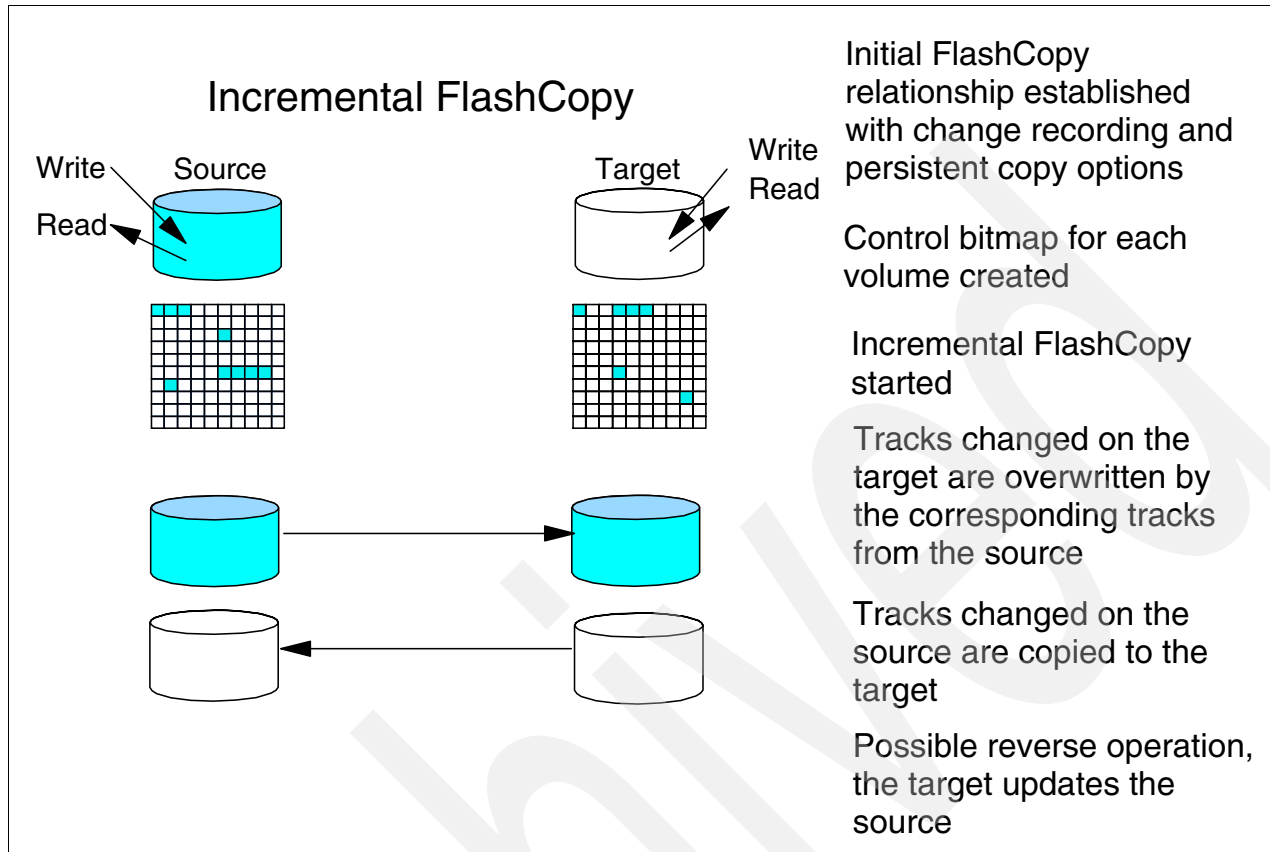


Figure 7-6 Incremental FlashCopy

In Incremental FlashCopy operations:

1. You issue full FlashCopy with the *change recording* option. This option is for creating change recording bitmaps in the storage unit. The change recording bitmaps are used for recording the tracks which are changed on the source and target volumes after the last FlashCopy.
2. After creating the change recording bitmaps, Copy Services records the information for the updated tracks to the bitmaps. The FlashCopy relationship persists even if all of the tracks have been copied from the source to the target.
3. The next time you issue Incremental FlashCopy, Copy Services checks the change recording bitmaps and copies only the changed tracks to the target volumes. If some tracks on the target volumes are updated, these tracks are overwritten by the corresponding tracks from the source volume.

You can also issue incremental FlashCopy from the target volume to the source volumes with the *reverse restore* option. The reverse restore operation is only possible after the background copy in the original direction is completed.

Data Set FlashCopy

Data Set FlashCopy allows a FlashCopy of a data set in a System z environment (Figure 7-7).

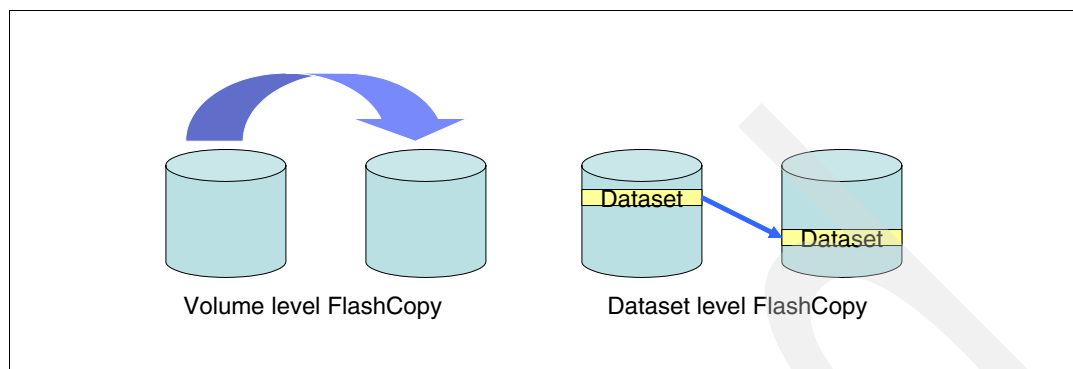


Figure 7-7 Data Set FlashCopy

Multiple Relationship FlashCopy

Multiple Relationship FlashCopy allows a source to have FlashCopy relationships with multiple targets simultaneously. A source volume or extent can be FlashCopied to up to 12 target volumes or target extents, as illustrated in Figure 7-8.

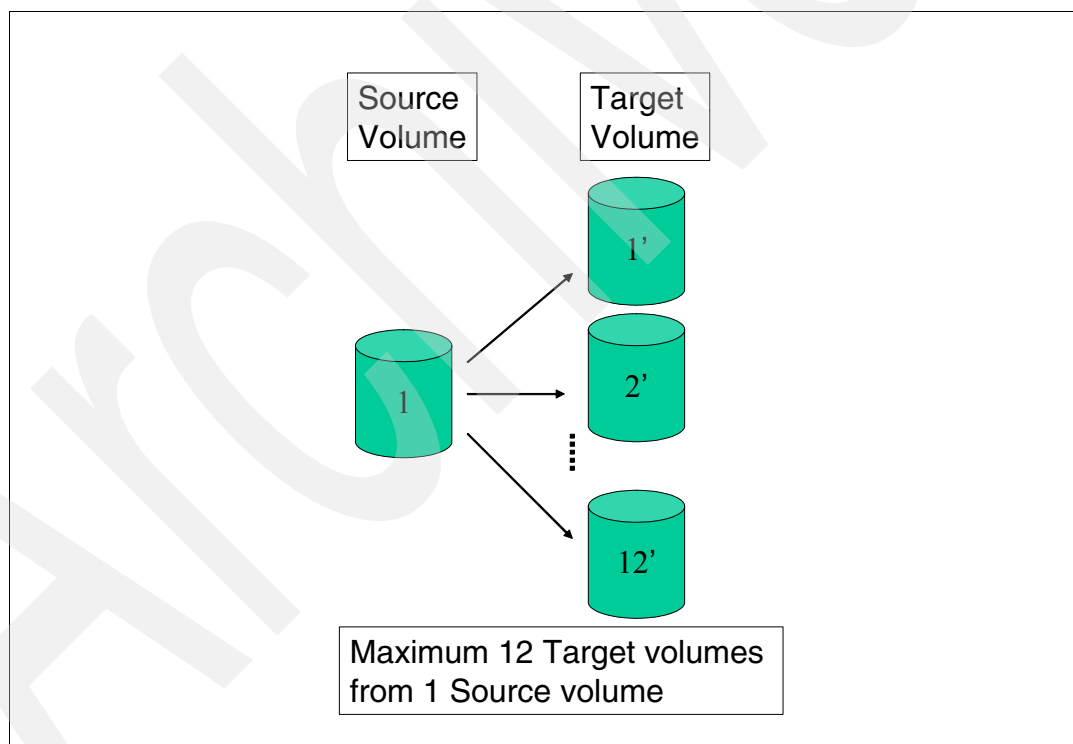


Figure 7-8 Multiple Relationship FlashCopy

Note: If a FlashCopy source volume has more than one target, that source volume can be involved only in a single incremental FlashCopy relationship.

Consistency Group FlashCopy

Consistency Group FlashCopy allows you to freeze (temporarily queue) I/O activity to a group of LUNs or volumes. Consistency Group FlashCopy helps create a consistent point-in-time copy across multiple LUNs or volumes, and even across multiple storage units.

What is Consistency Group FlashCopy?

If a consistent point-in-time copy across many logical volumes is required, and you do not wish to quiesce host I/O or database operations, then you can use Consistency Group FlashCopy to create a consistent copy across multiple logical volumes in multiple storage units.

In order to create this consistent copy, you would issue a set of **Establish FlashCopy** commands with a **freeze** option, which would suspend host I/O to the source volumes. In other words, Consistency Group FlashCopy provides the capability to temporarily queue (at the host I/O level, not the application level) subsequent write operations to the source volumes that are part of the Consistency Group. During the temporary queuing, **Establish FlashCopy** is completed. The temporary queuing continues until this condition is reset by the **Consistency Group Created** command or the time-out value expires (the default is two minutes).

Once all of the Establish FlashCopy requests have completed, a set of **Consistency Group Created** commands must be issued via the same set of DS network interface servers. The **Consistency Group Created** commands are directed to each logical subsystem (LSS) involved in the consistency group. The **Consistency Group Created** command allows the write operations to resume to the source volumes.

This operation is illustrated in Figure 7-9.

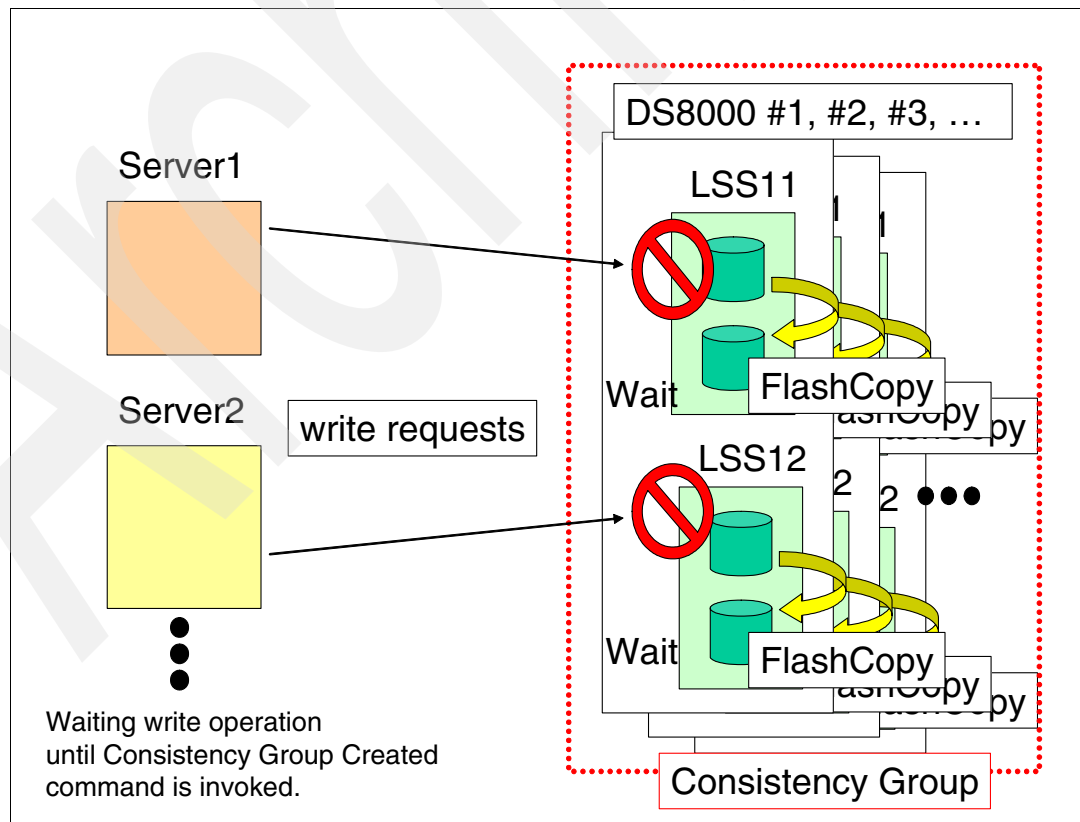


Figure 7-9 Consistency Group FlashCopy

Important: Consistency Group FlashCopy can create host-based consistent copies, they are not application-based consistent copies. The copies have *power-fail* or *crash* level consistency. This means that if you suddenly power off your server without stopping your applications and without destaging the data in the file cache, the data in the file cache could be lost and you might require recovery procedures to restart your applications. To start your system with Consistency Group FlashCopy target volumes, you might require the same operations as for crash recovery.

For example, If the Consistency Group source volumes are used with a journaled file system (like AIX JFS) and the source LUNs are not unmounted before running FlashCopy, it is likely that **fsck** must be run on the target volumes.

Establish FlashCopy on existing Remote Mirror and Copy primary

This option allows you to establish a FlashCopy relationship where the target is also a remote mirror primary volume. This enables you to create full or incremental point-in-time copies at a local site and then use remote mirroring commands to copy the data to the remote site, as shown in Figure 7-10.

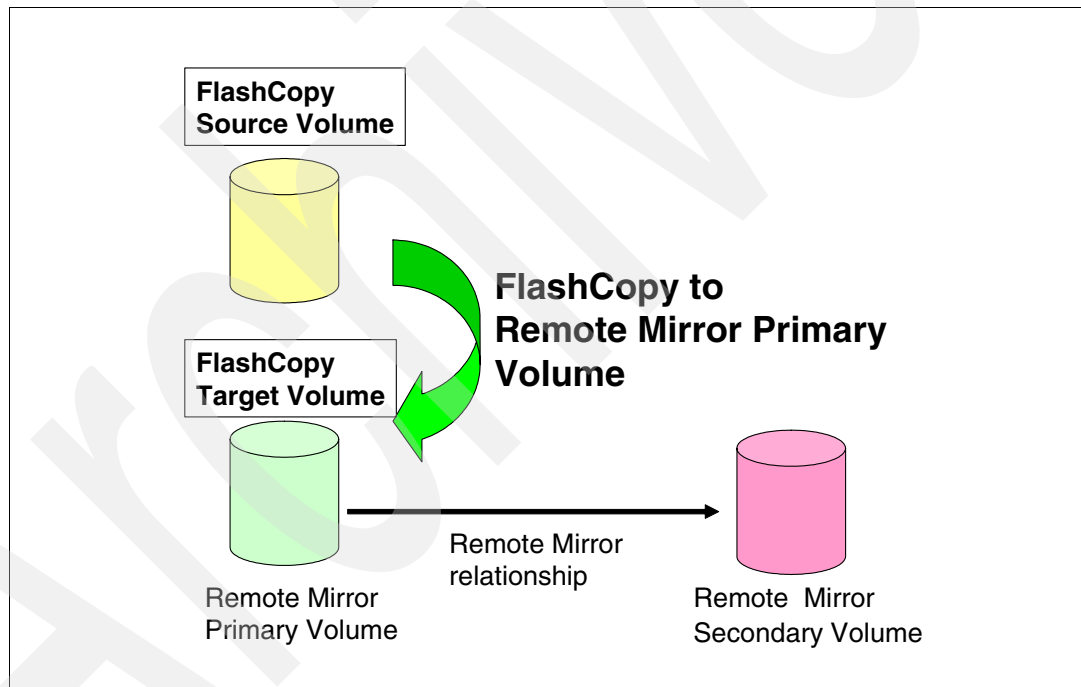


Figure 7-10 Establish FlashCopy on existing Remote Mirror and Copy Primary

Note: You cannot FlashCopy from a source to a target, where the target is also a Global Mirror primary volume.

Persistent FlashCopy

Persistent FlashCopy allows the FlashCopy relationship to remain even after the copy operation completes — the relationship must be explicitly deleted.

Inband Commands over Remote Mirror link

In a remote mirror environment, commands to manage FlashCopy at the remote site can be issued from the local or intermediate site and transmitted over the remote mirror Fibre Channel links. This eliminates the requirement for a network connection to the remote site solely for the management of FlashCopy.

7.7.3 Remote Mirror and Copy (Peer-to-Peer Remote Copy)

The Remote Mirror and Copy feature (formally called Peer-to-Peer Remote Copy, or PPRC) is a flexible data mirroring technology that allows replication between volumes on two or more disk storage systems. It can also be used for data backup and disaster recovery. Remote Mirror and Copy is an optional function - it requires purchase of the Remote Mirror and Copy 2244 function authorization model, which is 2244 Model RMC.

The DS8000 can participate in Remote Mirror and Copy solutions with the ESS Model 750, ESS Model 800, and DS6000. To establish a Remote Mirror relationship between the DS8000 and the ESS, the ESS requires licensed internal code (LIC) version 2.4.2 or later.

The Remote Mirror and Copy feature can operate in the following modes:

Metro Mirror (Synchronous PPRC)

Metro Mirror provides real-time mirroring of logical volumes between two DS8000s that can be located up to 300 km from each other. Greater distances are supported on special request. It is a synchronous copy solution where write operations are completed on both copies (local and remote site) before they are considered to be complete (Figure 7-11).

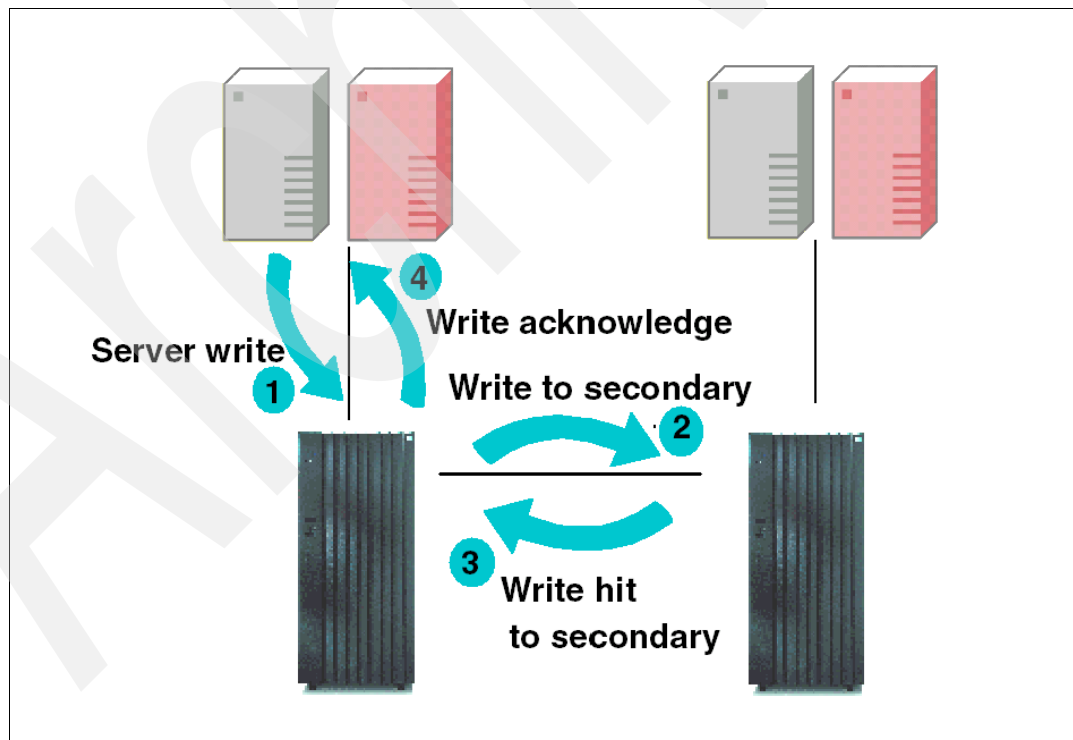


Figure 7-11 Metro Mirror

Global Copy (PPRC-XD)

Global Copy (Figure 7-12) copies data non-synchronously and over longer distances than is possible with Metro Mirror. When operating in Global Copy mode, the source volume sends a periodic, incremental copy of updated tracks to the target volume, instead of sending a constant stream of updates. This causes less impact to application writes for source volumes and less demand for bandwidth resources, while allowing a more flexible use of the available bandwidth.

Global Copy does not keep the sequence of write operations. Therefore, the copy is normally fuzzy, but you can make a consistent copy through synchronization (called a go-to-sync operation). After the synchronization, you can issue FlashCopy at the secondary site to make the backup copy with data consistency. After the establish of the FlashCopy, you can change the PPRC mode back to the non-synchronous mode.

Note: When you change PPRC mode from synchronous to non-synchronous mode, you change the PPRC mode from synchronous to suspend mode at first, and then you change PPRC mode from suspend to non-synchronous mode.

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

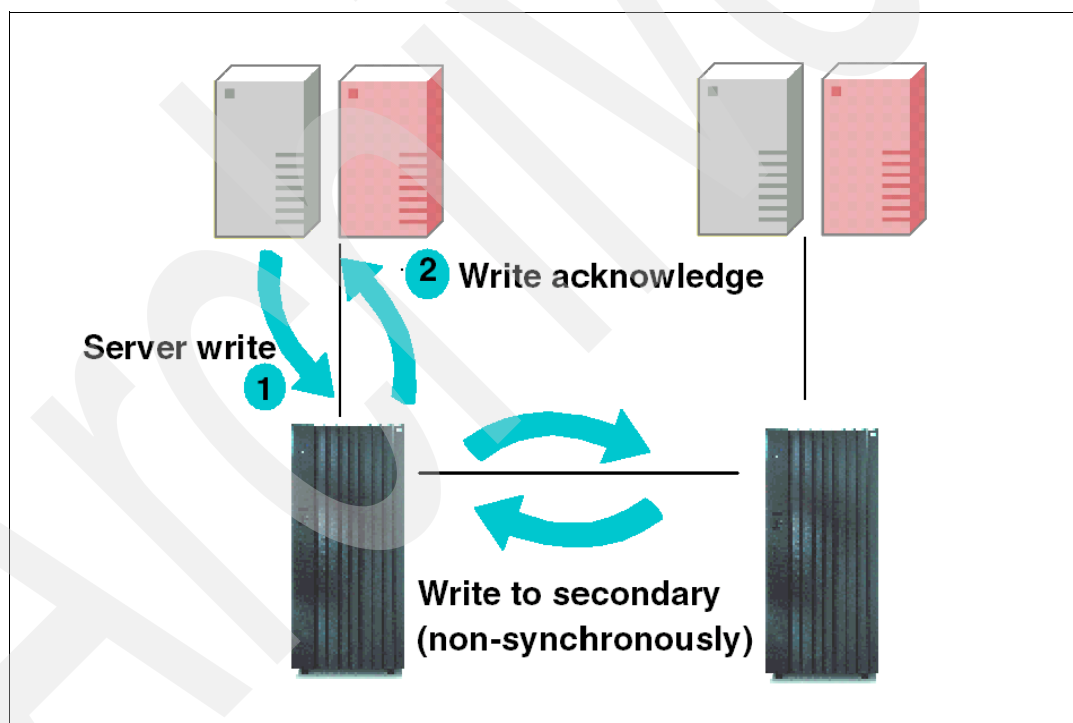


Figure 7-12 Global Copy

Global Mirror (Asynchronous PPRC)

Global Mirror provides a long-distance remote copy feature across two sites using asynchronous technology. This solution is based on the existing Global Copy and FlashCopy. With Global Mirror, the data that the host writes to the storage unit at the local site is asynchronously shadowed to the storage unit at the remote site. A consistent copy of the data is automatically maintained on the storage unit at the remote site.

Global Mirror operations provide the following benefits (see Figure 7-13):

- ▶ Support is provided for virtually unlimited distances between the local and remote sites, with the distance typically limited only by the capabilities of the network and the channel extension technology. This *unlimited* distance enables you to choose your remote site location based on business requirements and enables site separation to add protection from localized disasters.
- ▶ There is a consistent and restartable copy of the data at the remote site, created with minimal impact to applications at the local site.
- ▶ Data currency where, for many environments, the remote site lags behind the local site typically 3 to 5 seconds, minimizing the amount of data exposure in the event of an unplanned outage, is another benefit. The actual lag in data currency that you experience can depend upon a number of factors, including specific workload characteristics and bandwidth between the local and remote sites.
- ▶ Dynamic selection of the desired recovery point objective, based upon business requirements and optimization of available bandwidth, is provided.
- ▶ Session support is included, whereby data consistency at the remote site is internally managed across up to eight storage units located across the local and remote sites.
- ▶ Efficient synchronization of the local and remote sites, with support for failover and failback modes, helps reduce the time that is required to switch back to the local site after a planned or unplanned outage.

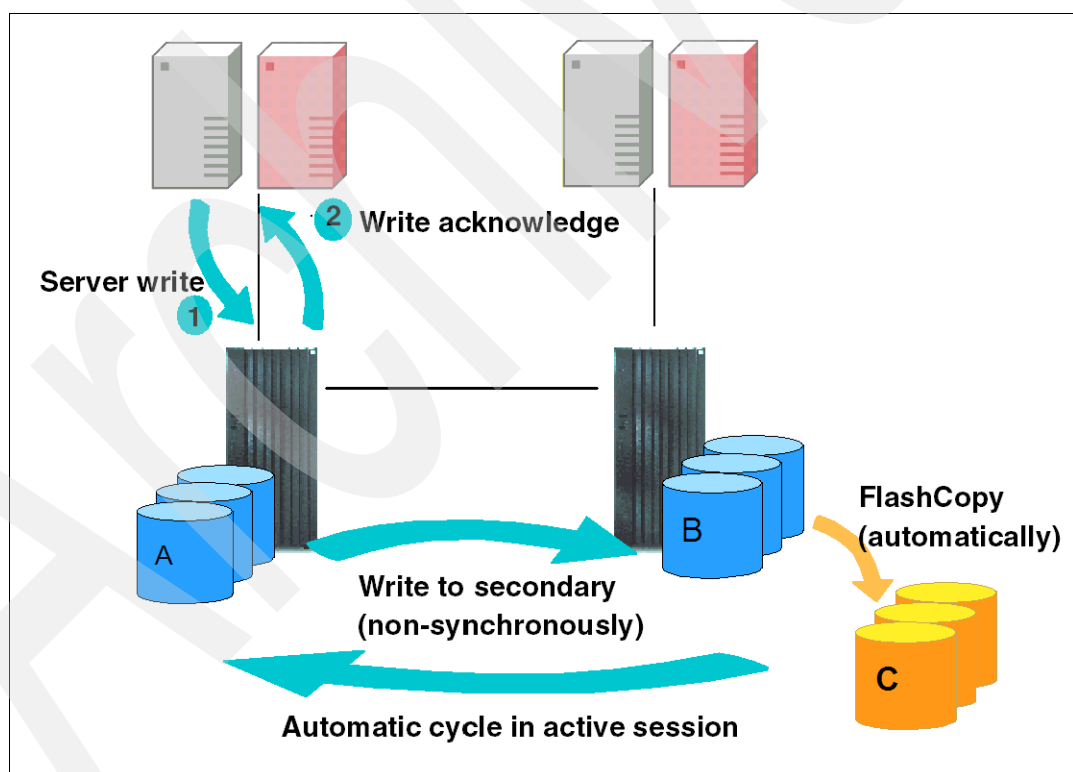


Figure 7-13 Global Mirror

How Global Mirror works

We explain how Global Mirror works in Figure 7-14.

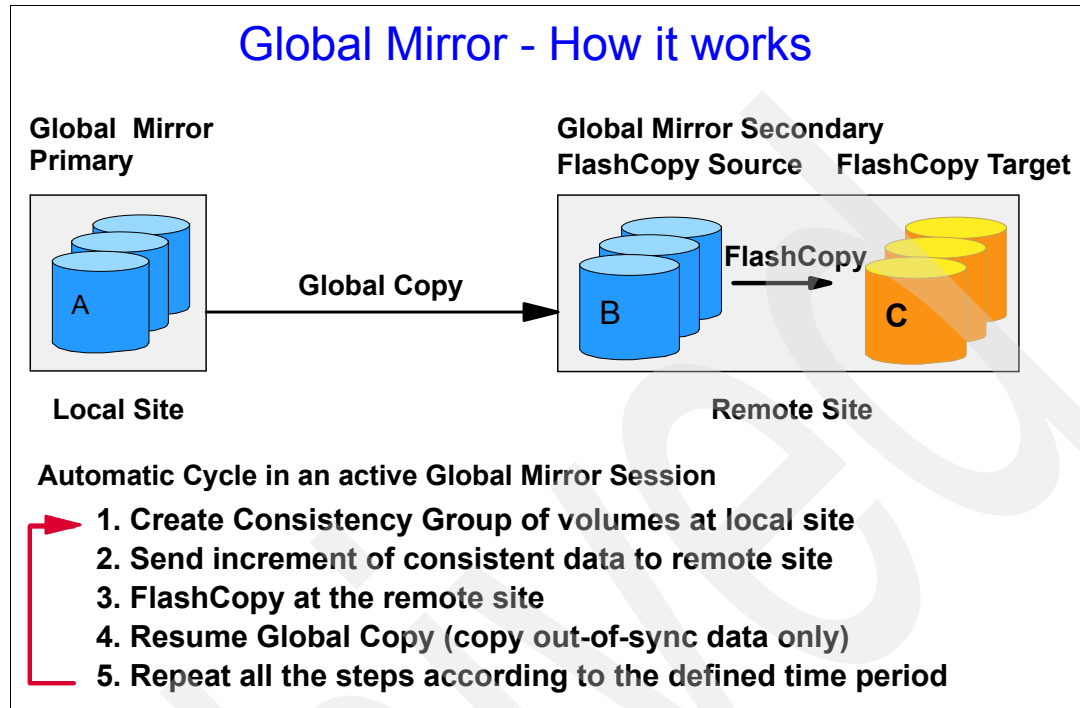


Figure 7-14 How Global Mirror works

The A volumes at the local site are the production volumes and are used as Global Copy primary volumes. The data from the A volumes is replicated to the B volumes, which are Global Copy secondary volumes. At a certain point in time, a Consistency Group is created using all of the A volumes, even if they are located in different Storage Units. This has no application impact because the creation of the Consistency Group is very quick (on the order of milliseconds).

Note: The copy created with Consistency Group is a power-fail consistent copy, not an application-based consistent copy. When you recover with this copy, you might require recovery operations, such as the `fsck` command in an AIX filesystem.

Once the Consistency Group is created, the application writes can continue updating the A volumes. The increment of the consistent data is sent to the B volumes using the existing Global Copy relationship. Once the data reaches the B volumes, it is FlashCopied to the C volumes.

The C volumes now contain the *consistent* copy of data. Because the B volumes usually contain a *fuzzy* copy of the data from the local site (not when doing the FlashCopy), the C volumes are used to hold the last point-in-time consistent data while the B volumes are being updated by the Global Copy relationship.

Note: When you implement Global Mirror, you set up the FlashCopy between the B and C volumes with *No Background copy* and *Start Change Recording* options. It means that before the latest data is updated to the B volumes, the last consistent data in the B volume is moved to the C volumes. Therefore, at some time, a part of consistent data is in the B volume, and the other part of consistent data is in the C volume.

If a disaster occurs during the FlashCopy of the data, special procedures are required to finalize the FlashCopy.

In the recovery phase, the consistent copy is created in the B volumes. You must have some operations to check and create the consistent copy.

You should check the status of the B volumes for the recovery operations. Generally, these check and recovery operations are complicated and difficult with the GUI or CLI in a disaster situation. Therefore, you might want to use some management tools, (for example, Global Mirror Utilities), or management software, (for example, TPC for Replication), for Global Mirror to automate this recovery procedure.

The data at the remote site is current within 3 to 5 seconds, but this recovery point (RPO) depends on the workload and bandwidth available to the remote site.

In contrast to the previously mentioned Global Copy solution, Global Mirror overcomes its disadvantages and automates all of the steps that have to be done manually when using Global Copy.

If you use Global Mirror, you must adhere to the following additional rules:

- ▶ You must purchase a Point-in-Time Copy function authorization (2244 Model PTC) for the secondary storage unit.
- ▶ If Global Mirror is to be used during failback on the secondary storage unit, you must also purchase a Point-in-Time Copy function authorization for the primary system.

z/OS Global Mirror (XRC)

DS8000 storage complexes support z/OS Global Mirror only on zSeries hosts. The z/OS Global Mirror function mirrors data on the storage unit to a remote location for disaster recovery. It protects data consistency across all volumes that you have defined for mirroring. The volumes can reside on several different storage units. The z/OS Global Mirror function can mirror the volumes over several thousand kilometers from the source site to the target recovery site.

With z/OS Global Mirror (Figure 7-15), you can suspend or resume service during an outage. You do not have to terminate your current data-copy session. You can suspend the session, then restart it. Only data that changed during the outage has to be re-synchronized between the copies. The z/OS Global Mirror function is an optional function. To use it, you must purchase the remote mirror for z/OS 2244 function authorization model, which is 2244 Model RMZ.

Note: DS6000 series systems can only be used as a target system in z/OS Global Mirror operations.

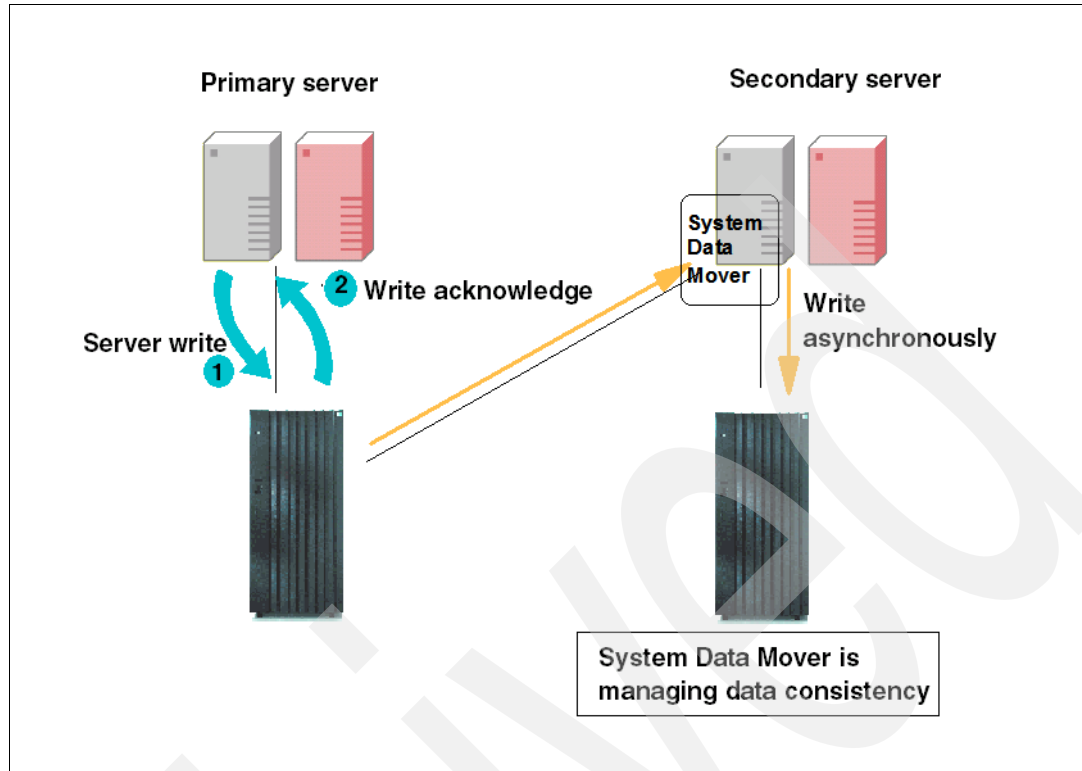


Figure 7-15 z/OS Global Mirror

Metro/Global Copy(3-site Metro Mirror and Global Copy)

Metro/Global Copy is a cascaded three-site disk mirroring solution. Metro Mirror is used between production site A and intermediate site B. Global Copy is used between the intermediate site B and the remote site C. Metro/Global Copy is often used for migration purposes in a two site mirroring environment. Global Copy keeps the cascaded copy (which can be located at either the remote or local site) nearly current with the running two-site disk mirroring configuration.

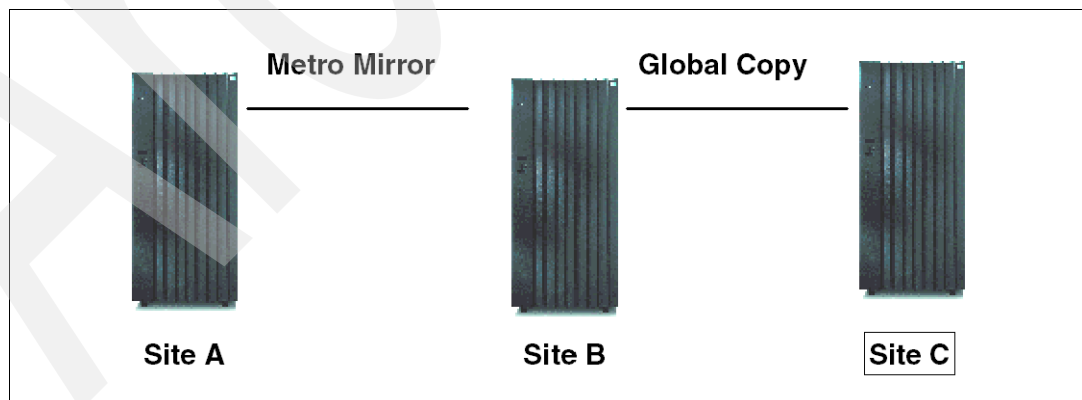


Figure 7-16 Metro/Global Copy

Global Copy does not keep the sequence of write operations. Therefore, the copy of Site C is normally fuzzy, but you can make a consistent copy through synchronization (called a go-to-sync operation). After the synchronization, you can issue FlashCopy at the secondary site to make the backup copy with data consistency. After the establish of the FlashCopy, you can change the PPRC mode back to the non-synchronous mode.

Note: When you create a consistent copy for Global Copy, you require the go-to-sync operation (synchronize the secondary volumes to the primary volumes). During the go-to-sync operation, PPRC changes from a non-synchronous copy to a synchronous copy. Therefore, the go-to-sync operation might cause performance impact to your application system. If the data is heavily updated and the network bandwidth for PPRC is limited, the time for the go-to-sync operation becomes longer.

One example of the advanced function licenses for each disk system in a Metro/Global Copy solution is shown in Figure 7-17.

Site A	Site B	Site C
<u>921/922/9A2</u> DS6800 RMC License	<u>921/922/9A2</u> DS6800 RMC License	<u>921/922/9A2</u> DS6800 RMC License PTC License
<u>922/932/9B2</u> MM License	<u>922/932/9B2</u> MM License	<u>922/932/9B2</u> MM(or GM) License PTC License

Figure 7-17 Metro/Global Copy licenses

Metro/Global Mirror (three-site Metro Mirror and Global Mirror)

Metro/Global mirror is a cascaded three-site disk mirroring solution. Metro Mirror is used between the production site A and intermediate site B. Global Mirror is used between the intermediate site B and the remote site C.

Note: FlashCopy is required when you have to keep a consistent copy on site C.

In the event of a loss of access to intermediate site B, the license for DS8000 Metro/Global Mirror provides new functionality to incrementally resynchronize and establish Global Mirror from production site A to remote site C, without application impact to site A - thus maintaining out of region disaster recovery. When intermediate site B access returns, that site can be re-inserted into the three site cascading topology without impact to production site A applications. In all cases, only incremental changes have to be sent for resynchronization. Refer to Figure 7-18.

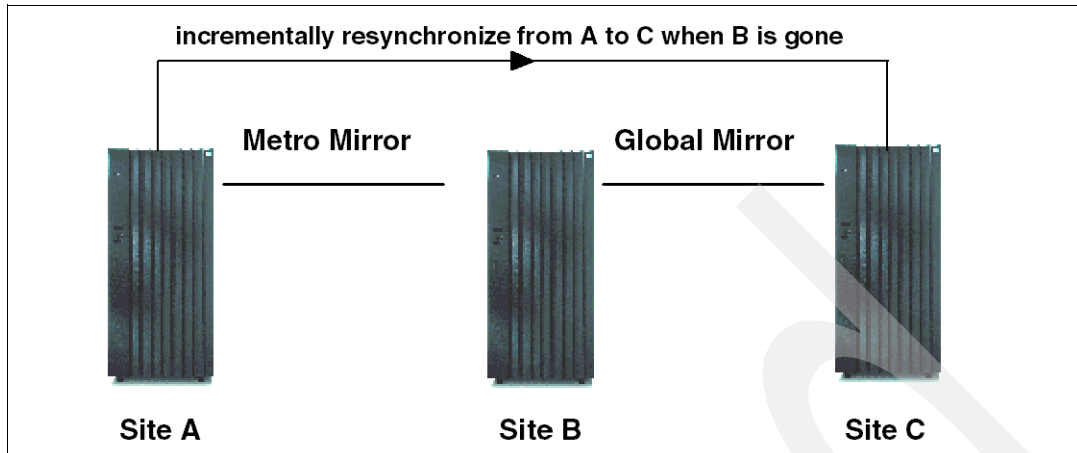


Figure 7-18 Metro/Global Mirror

One example of the advanced function licenses for each disk system in a Metro/Global Mirror solution is shown in Figure 7-19.

Site A	Site B	Site C
<u>921/922/9A2</u> MGM License RMC License	<u>921/922/9A2</u> MGM License RMC License	<u>921/922/9A2</u> MGM License RMC License PTC License
<u>922/932/9B2</u> MGM License MM License GM Add License	<u>922/932/9B2</u> MGM License MM License GM Add License	<u>922/932/9B2</u> MGM License GM License PTC License

Figure 7-19 Metro/Global Mirror licenses

z/OS Metro/Global Mirror (3-site z/OS Global Mirror and Metro Mirror)

Note: GM Add License of Site A is required for site A/C resync.

This mirroring capability uses z/OS Global Mirror to mirror primary site data to a long distance location and also uses Metro Mirror to mirror primary site data to a location within the metropolitan area. This enables a z/OS 3-site high availability and disaster recovery solution for even greater protection from unplanned outages. Refer to Figure 7-20.

The z/OS Metro/Global Mirror function is an optional function; it requires purchase of both:

- Remote Mirror for z/OS (2244 Model RMZ)
- Remote Mirror and Copy function (2244 Model RMC) for both the primary and secondary storage units

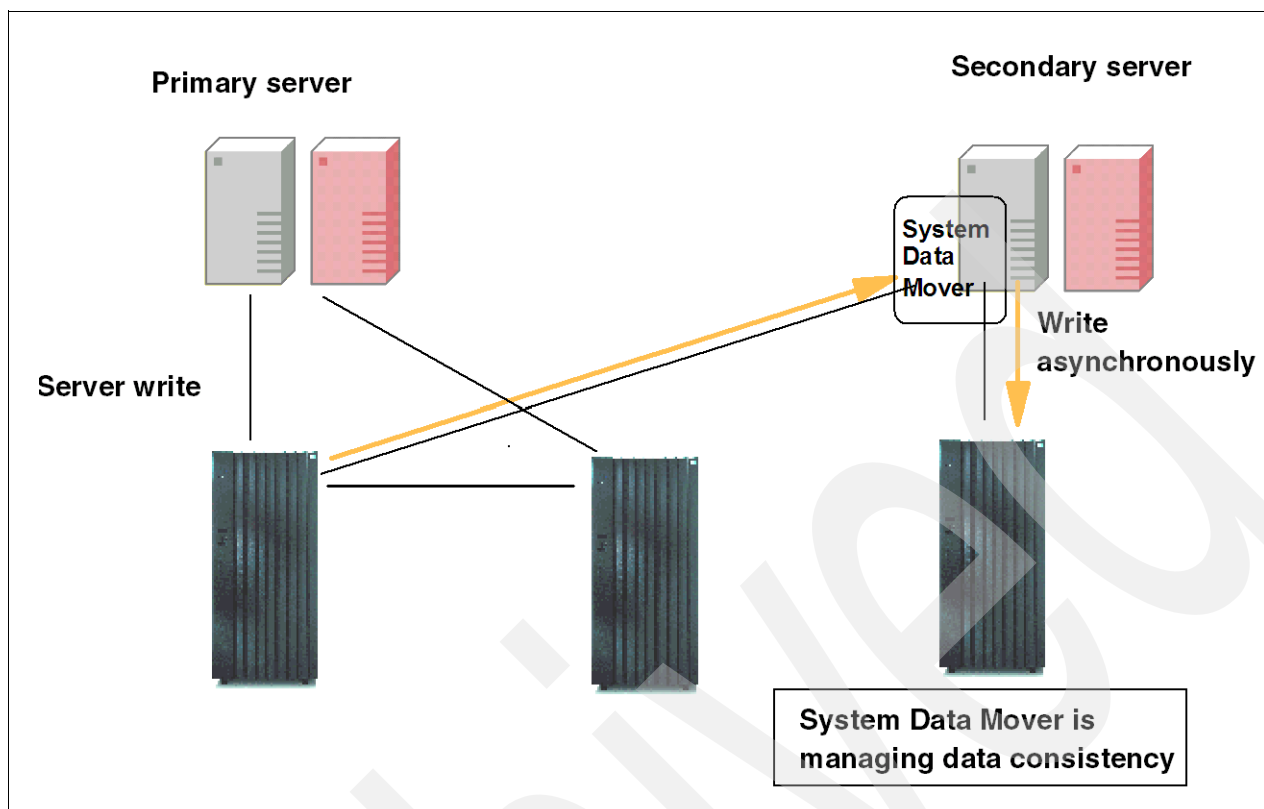


Figure 7-20 z/OS Metro/Global Mirror

An example of the advanced function Licenses for each disk system in a Metro/Global Mirror solution is shown in Figure 7-21.

Site A	Site B	Site C
<u>921/922/9A2</u> RMC License RMZ License	<u>921/922/9A2</u> RMC License RMZ License	<u>921/922/9A2</u> RMC License RMZ License PTC License
<u>922/932/9B2</u> MM License RMZ License	<u>922/932/9B2</u> MM License RMZ License	<u>922/932/9B2</u> MM License RMZ License PTC License

Figure 7-21 Z/OS Metro/Global Mirror Licenses

Note: FlashCopy is required on site C for the tertiary copy creation, which is normally being used for disaster recovery drill or testing.

7.7.4 Comparison of the Remote Mirror and Copy functions

In this section we summarize the use of and considerations for Remote Mirror and Copy functions.

Metro Mirror (Synchronous PPRC)

Here are some considerations:

► **Description:**

Metro Mirror is a function for synchronous data copy at a distance.

► **Advantages:**

There is no data loss and it allows for rapid recovery for distances up to 300 km. Greater distances are supported on special request.

► **Considerations:**

There might be a slight performance impact for write operations.

Note: If you want to use a Metro Mirror copy from the application server which has the Metro Mirror primary volume, you have to compare its function with OS mirroring.

You might experience some disruption to recover your system with Metro Mirror secondary volumes in an open system environment, because Metro Mirror secondary volumes are not online to the application servers during the Metro Mirror relationship.

You might also require some operations before assigning Metro Mirror secondary volumes. For example, in an AIX environment, AIX assigns specific IDs to each volume (PVID). Metro Mirror secondary volumes have the same PVID as Metro Mirror primary volumes. AIX cannot manage the volumes with the same PVID as different volumes. Therefore, before using the Metro Mirror secondary volumes, you have to clear the definition of the Metro Mirror primary volumes or reassign PVIDs to the Metro Mirror secondary volumes.

Some operating systems (OS) or file systems (for example, AIX LVM) have a function for disk mirroring. OS mirroring requires some server resources, but usually can keep operating with the failure of one volume of the pair and recover from the failure nondisruptively. If you use a Metro Mirror copy from the application server for recovery, you have to consider which solution (Metro Mirror or OS mirroring) is better for your system.

Global Copy (PPRC-XD)

Here are some considerations:

► **Description:**

Global Copy is a function for continuous copy without data consistency.

► **Advantages:**

It can copy data at nearly an unlimited distance, even if limited by network and channel extender capabilities. It is suitable for data migration and daily backup to the remote site.

► **Considerations:**

The copy is normally *fuzzy* but can be made consistent through synchronization.

Note: When you create a consistent copy for Global Copy, you require the go-to-sync operation (synchronize the secondary volumes to the primary volumes). During the go-to-sync operation, Metro Mirror changes from a non-synchronous copy to a synchronous copy. Therefore, the go-to-sync operation might cause performance impact to your application system. If the data is heavily updated and the network bandwidth for Metro Mirror is limited, the time for the go-to-sync operation becomes longer.

Global Mirror (Asynchronous PPRC)

Here are some considerations:

► **Description:**

Global Mirror is an asynchronous copy; you can create a consistent copy in the secondary site with an adaptable Recovery Point Objective (RPO).

Note: Recovery Point Objective (RPO) specifies how much data you can afford to re-create should the system have to be recovered.

► **Advantages:**

Global Mirror can copy with nearly an unlimited distance. It is scalable across the storage units. It can realize a low RPO with enough link bandwidth. Global Mirror causes only a slight impact to your application system.

► **Considerations:**

When the link bandwidth capability is exceeded with a heavy workload, the RPO might grow.

Note: Managing Global Mirror can be complex. Therefore, we recommend management utilities, for example, TotalStorage Productivity Center for Replication.

z/OS Global Mirror (XRC)

Here are some considerations:

► **Description:**

z/OS Global Mirror is an asynchronous copy controlled by z/OS host software, called *System Data Mover*.

► **Advantages:**

It can copy with nearly unlimited distance. It is highly scalable, and it has very low RPO.

► **Considerations:**

Additional host server hardware and software is required. The RPO might grow if bandwidth capability is exceeded, or host performance might be impacted.

7.7.5 What is a Consistency Group?

With Copy Services, you can create *Consistency Groups* for FlashCopy and Metro Mirror. Consistency Group is a function to keep *data consistency* in the backup copy. Data consistency means that the order of dependent writes is kept in the copy.

In this section we define *data consistency* and *dependent writes*, and then we explain how Consistency Group operations keep data consistency.

What is data consistency?

Many applications, such as databases, process a repository of data that has been generated over a period of time. Many of these applications require that the repository is in a consistent state in order to begin or continue processing. In general, consistency implies that the order of dependent writes is preserved in the data copy. For example, the following sequence might occur for a database operation involving a log volume and a data volume:

1. Write to log volume: Data Record #2 is being updated.
2. Update Data Record #2 on data volume.
3. Write to log volume: Data Record #2 update complete.

If the copy of the data contains any of these combinations then the data is consistent:

- ▶ Operation 1, 2, and 3
- ▶ Operation 1 and 2
- ▶ Operation 1

If the copy of data contains any of those combinations then the data is *inconsistent* (the order of dependent writes was *not* preserved):

- ▶ Operation 2 and 3
- ▶ Operation 1 and 3
- ▶ Operation 2
- ▶ Operation 3

In the Consistency Group operation, data consistency means this sequence is always kept in the backup data.

The order of non-dependent writes does not necessarily have to be preserved. For example, consider the following two sequences:

1. Deposit paycheck in checking account A.
2. Withdraw cash from checking account A.
3. Deposit paycheck in checking account B.
4. Withdraw cash from checking account B.

In order for the data to be consistent, the deposit of the paycheck must be applied *before* the withdraw of cash for each of the checking accounts. However, it does not matter whether the deposit to checking account A or checking account B occurred first, as long as the associated withdrawals are in the correct order. So for example, the data copy would be consistent if the following sequence occurred at the copy. In other words, the order of updates is not the same as it was for the source data, but the order of *dependent* writes is still preserved.

1. Deposit paycheck in checking account B.
2. Deposit paycheck in checking account A.
3. Withdraw cash from checking account B.
4. Withdraw cash from checking account A.

How does Consistency Group keep data consistency?

Consistency Group operations cause the storage units to hold I/O activity to a volume for a time period by putting the source volume into an *extended long busy* state. This operation can be done across multiple LUNs or volumes, and even across multiple storage units.

In the storage subsystem itself, each command is managed with each logical subsystem (LSS). This means that there are slight time lags until each volume in the different LSS is changed to the extended long busy state. Some people are concerned that the time lag causes you to lose data consistency, but, it is not true. We explain how to keep data consistency in the Consistency Group environments in the following section.

See Figure 7-22. In this case, three write operations (labeled 1st, 2nd, and 3rd) are dependent writes. It means that these operations must be completed sequentially.

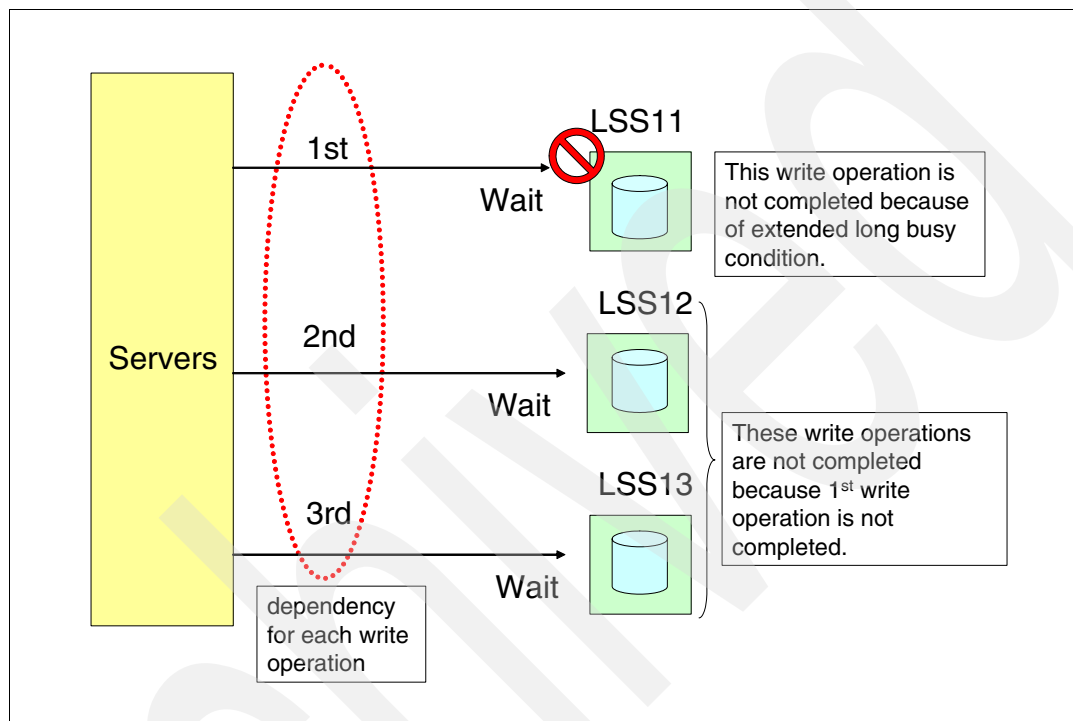


Figure 7-22 Consistency Group: Example1

Because of the time lag for Consistency Group operations, some volumes in some LSSs are in an extended long busy state and other volumes in the other LSSs are not.

In Figure 7-22, the volumes in LSS11 are in an extended long busy state, and the volumes in LSS12 and 13 are not. The 1st operation is not completed because of this extended long busy state, and the 2nd and 3rd operations are not completed, because the 1st operation has not been completed. In this case, 1st, 2nd, and 3rd updates are not included in the backup copy. Therefore, this case is consistent.

Now, refer to Figure 7-23. In this case, the volumes in LSS12 are in an extended long busy state and the other volumes in LSS11 and 13 are not. The 1st write operation is completed because the volumes in LSS11 are not in an extended long busy state. The 2nd write operation is not completed because of the extended long busy state. The 3rd write operation is not completed either, because the 2nd operation is not completed. In this case, the 1st update is included in the backup copy, and the 2nd and 3rd updates are not included. Therefore, this case is consistent.

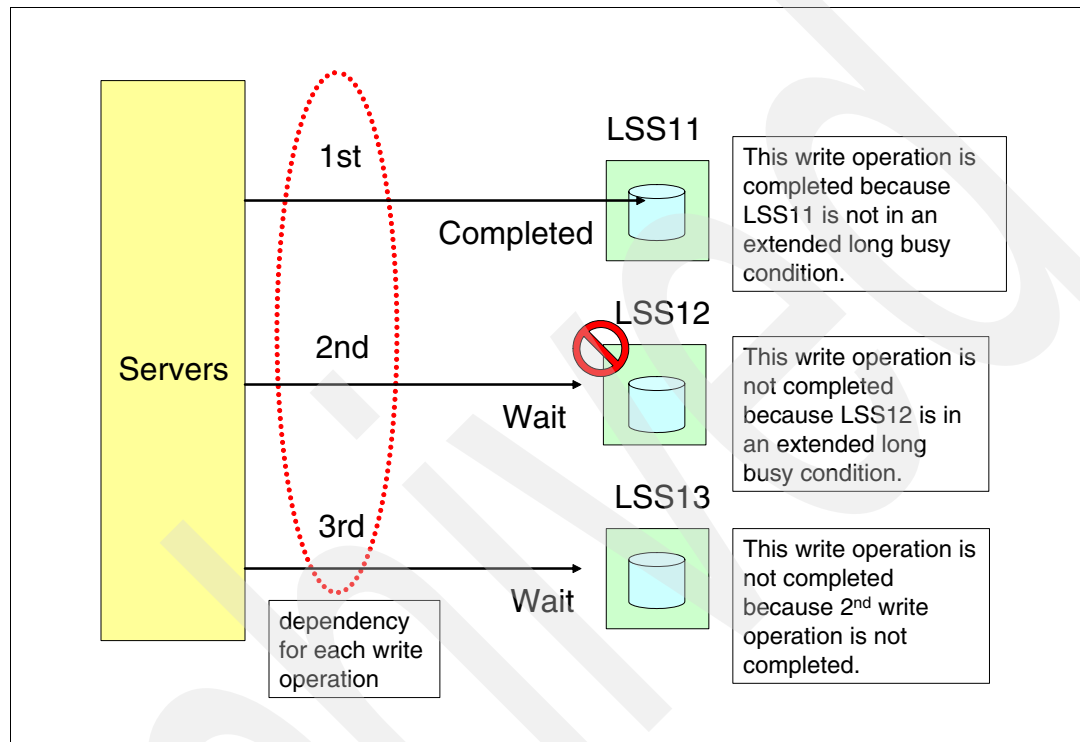


Figure 7-23 Consistency Group: Example 2

In all cases, if each write operation is dependent, the Consistency Group can keep the data consistent in the backup copy.

If each write operation is not dependent, the I/O sequence is not kept in the copy that is created by the Consistency Group operation. See Figure 7-24. In this case, the three write operations are independent. If the volumes in LSS12 are in an extended long busy state and the other volumes in LSS11 and 13 are not, the 1st and 3rd operations are completed and the 2nd operation is not completed.

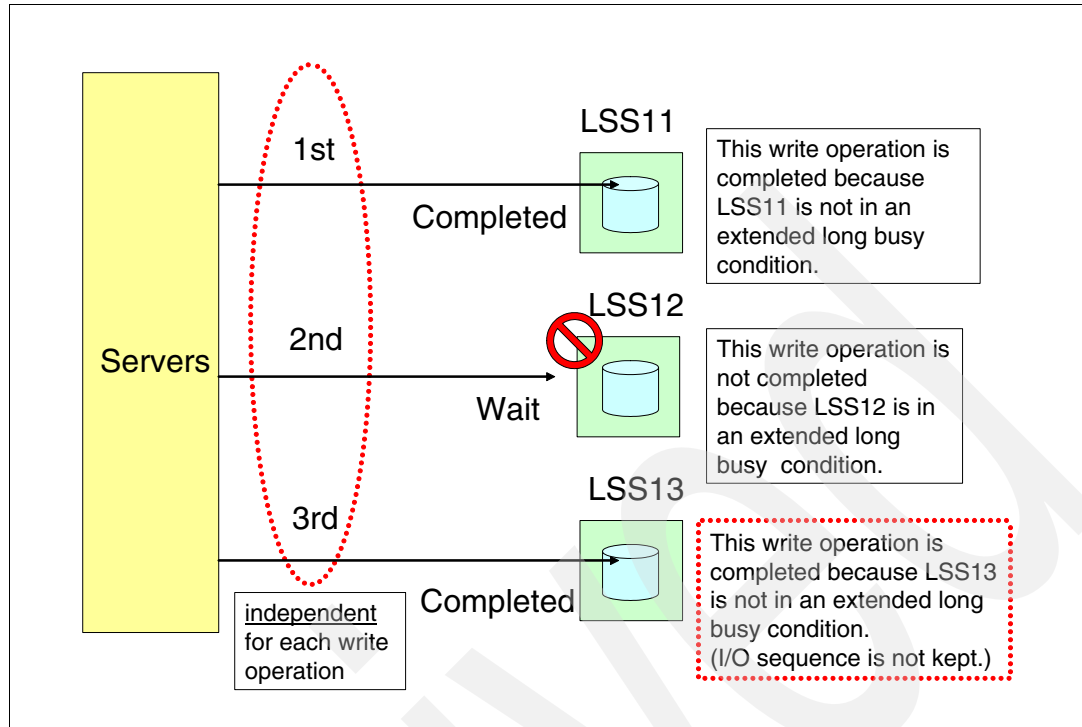


Figure 7-24 Consistency Group: Example 3

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

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

7.8 Interfaces for Copy Services

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

7.8.1 Storage Hardware Management Console (S-HMC)

Copy Services functions can be initiated over the following interfaces:

- ▶ zSeries Host I/O Interface
- ▶ DS Storage Manager web-based Interface
- ▶ DS Command-Line Interface (DS CLI)
- ▶ DS open application programming interface (DS Open API)

DS Storage Manager, DS CLI, and DS Open API commands are issued via the TCP/IP network, and these commands are invoked by the Storage Hardware Management Console (S-HMC). When the S-HMC has the command requests, including those for Copy Services, from these interfaces, S-HMC communicates with each server in the storage units via the TCP/IP network. Therefore, the S-HMC is a key component to configure and manage the DS8000.

The network components for Copy Services are illustrated in Figure 7-25.

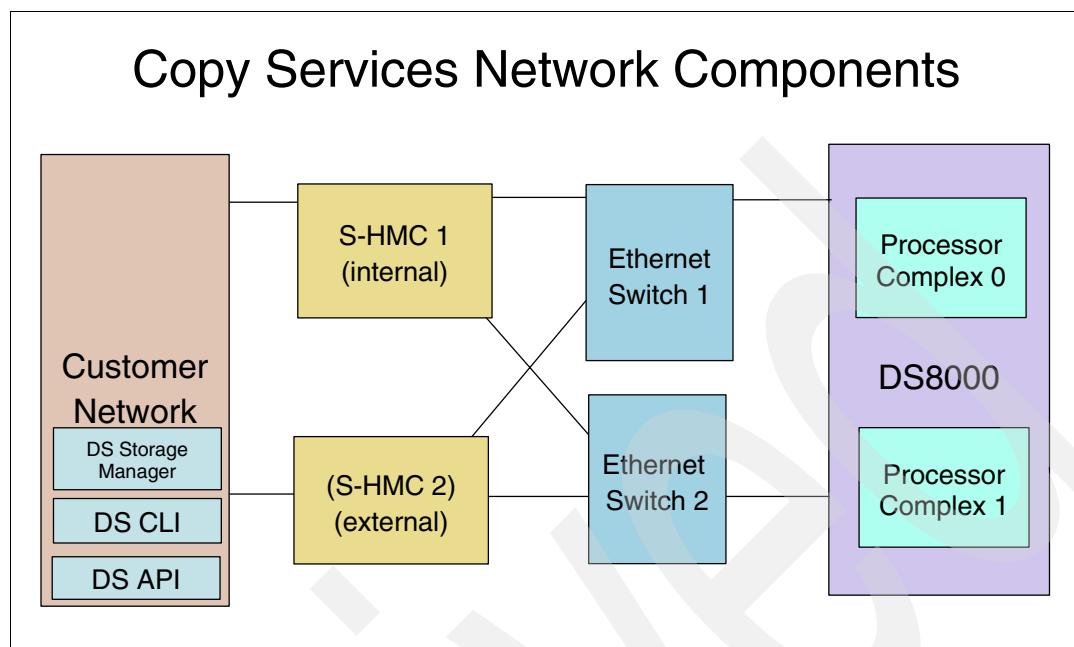


Figure 7-25 DS8000 Copy Services network components

Each DS8000 has an internal S-HMC in the base frame, and you can have an external S-HMC for redundancy.

7.8.2 DS Storage Manager Web-based interface

DS Storage Manager is a Web-based management interface used for managing the logical configurations and invoking the Copy Services functions. The DS Storage Manager has an online mode and an offline mode; only the online mode is supported for Copy Services.

DS Storage Manager is already installed in the S-HMC, and can also be installed on other systems. When managing Copy Services functions with DS Storage Manager on other systems, DS Storage Manager issues its command to the S-HMC via the TCF/IP network.

7.8.3 DS Command-Line Interface (DS CLI)

The IBM DS Command-Line Interface (CLI) helps enable open systems hosts to invoke and manage the Point-in-Time Copy and Remote Mirror and Copy functions through batch processes and scripts. The CLI provides a full-function command set for checking the storage unit configuration and performing specific application functions when necessary.

Here are a few of the specific types of functions available with the DS CLI:

- ▶ Check and verify storage unit configuration.
- ▶ Check the current Copy Services configuration that is used by the storage unit.
- ▶ Create new logical storage and Copy Services configuration settings.
- ▶ Modify or delete logical storage and Copy Services configuration settings.

7.8.4 DS Open application programming Interface (API)

The DS Open application programming interface (API) is a non-proprietary storage management client application that supports routine LUN management activities, such as LUN creation, mapping and masking, and the creation or deletion of RAID-5 and RAID-10 volume spaces. The DS Open API also enables Copy Services functions such as FlashCopy and Remote Mirror and Copy. It supports these activities through the use of the Storage Management Initiative Specification (SMIS), as defined by the Storage Networking Industry Association (SNIA)

The DS Open API helps integrate DS configuration management support into storage resource management (SRM) applications, which allow clients to benefit from existing SRM applications and infrastructures. The DS Open API also enables the automation of configuration management through client-written applications. Either way, the DS Open API presents another option for managing storage units by complementing the use of the IBM DS Storage Manager web-based interface and the DS Command-Line Interface.

You must implement the DS Open API through the IBM Common Information Model (CIM) agent, a middleware application that provides a CIM-compliant interface. The DS Open API uses the CIM technology to manage proprietary devices such as open system devices through storage management applications. The DS Open API allows these storage management applications to communicate with a storage unit.

IBM supports IBM TotalStorage Productivity Center (TPC) for the DS8000/DS6000/ESS. TPC consists of software components which enable storage administrators to monitor, configure, and manage storage devices and subsystems within a SAN environment. TPC also has a function to manage the Copy Services functions, called TPC for Replication (see 12.4, “IBM TotalStorage Productivity Center” on page 394).

7.9 Interoperability with ESS

Copy Services for DS8000 also supports the IBM Enterprise Storage Server Model 800 (ESS 800) and Model 750. To manage the ESS 800 from the Copy Services for DS8000, you have to install licensed internal code version 2.4.2 or later on the ESS 800.

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

Note: DS8000 does not support PPRC with an ESCON link. If you want to configure a PPRC relationship between a DS8000 and ESS 800, you have to use a FCP link.

Archived

The IBM System Storage DS4000

In this chapter we describe the IBM System Storage DS4000 disk storage family and how it relates to IT Business Continuity. The members of the IBM System Storage DS4000 are a true series of solutions with common functionality, centralized administration, and seamless “data intact” upgrade path. The DS4000 family provides a variety of offerings to best match individual open system storage requirements, as well as providing the flexibility to easily manage growth from entry-level to the enterprise. The IBM System Storage DS4000 family also provides an excellent platform for storage consolidation.

This chapter does not discuss the technical details of the IBM System Storage DS4000, nor does it cover all the possibilities that are available with the DS4000 Storage Manager software. To learn more about the IBM System Storage DS4000 family and the Storage Manager, we recommend the following IBM Redbooks and IBM manuals:

- ▶ *DS4000 Best Practices and Performance Tuning Guide*, SG24-6363
- ▶ *IBM System Storage DS4000 Series, Storage Manager and Copy Services*, SG24-7010
- ▶ *IBM System Storage DS4000 Storage Manager Version 9, Installation and Support Guide for AIX, HP-UX, Solaris, and Linux on POWER*, GC26-7848
- ▶ *IBM System Storage DS4000 Storage Manager Version 9.16, Copy Services User's Guide*, GC26-7850

8.1 DS4000 series

The IBM System Storage DS4000 series caters to immediately accessible, highly available, and functional storage capacity.

At the heart of the DS4000 series are dual RAID controllers that contain 4Gb Fibre Channel (FC) interfaces to connect to the host systems and the disk drive enclosures. The disk drive enclosures contain either native Fibre Channel hard disk drives (EXP710, EXP810) that are dedicated to high availability and high performance requirements or SATA hard disk drives (EXP100, EXP420) where lower cost storage is required. SATA stands for Serial Advanced Technology Attachment.

All of the DS4000 series systems have hot swappable controllers and redundant power supplies and fans. Figure 8-1 shows the positioning of the various DS4000 family members related to the available capacity, performance, and features.

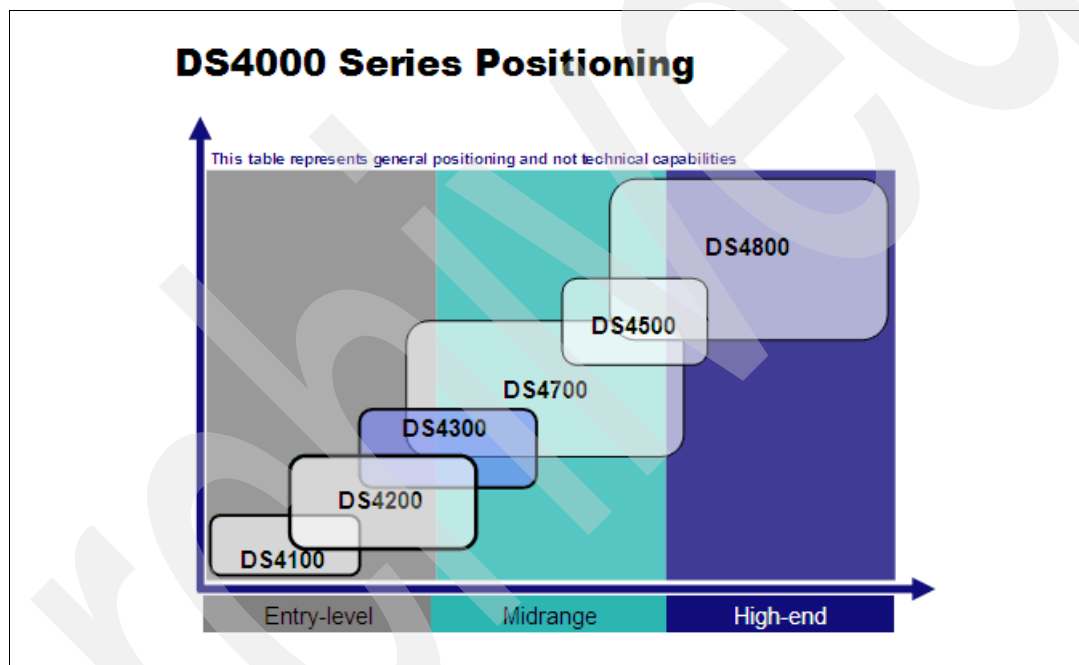


Figure 8-1 Positioning of the various DS4000 family members

8.1.1 Introducing the DS4000 series hardware overview

The various DS4000 models available, and their hardware specifications, are shown in Figure 8-2, including the connectivity, scalability, and performance capabilities of the various members in the DS4000 series.

The DS4000 series offers exceptional performance, robust functionality, and unparalleled ease of use. Building on a successful heritage, the DS4000 series offers next generation 4-Gbps FC technology while retaining compatibility with existing 1 and 2 Gbps devices.

DS4000 – Specifications Comparison

	DS4200	DS4700 Mod 72	DS4800 Mod 80	DS4800 Mod 88
CPU Processors	One 600 Mhz Xscale w/XOR	One 667 Mhz Xscale w/XOR	Intel Xeon 2.4 GHz	Intel Xeon 2.4 GHz
Cache Memory, total	2 GB	4GB	4GB	16GB
Host FC Ports, total	4 – 4Gbps Autoneg. 2, 1	8 – 4Gbps Autoneg. 2, 1	8 – 4Gbps Autoneg. 2, 1	8 – 4Gbps Autoneg. 2, 1
Disk FC Ports	4 – 4Gbps	4 – 4Gbps or 2Gbps	4 – 4Gbps or 2Gbps	8 – 4Gbps or 2Gbps
Max. Disk Drives (w/EXPs)	SATA 112	FC – 112 SATA 112	FC – 224 SATA - 224	FC – 224 SATA - 224
Max. HA Hosts	256	256	512	512
Max. Storage Partitions / LUNs	64/1024	64/1024	64/2048	64/2048
Premium Features	FlashCopy, VolumeCopy, RVM	FlashCopy, VolumeCopy, RVM, Intermix	FlashCopy, VolumeCopy, RVM, Intermix	FlashCopy, VolumeCopy, RVM, Intermix
Performance				
Cached Read IOPS	120k	120k	275k	575k
Disk Reads IOPS	11.2k	44k	58k	79.2k
Write IOPS	1.8k	9k	17k	10.9k
Cached Reads MB/s	1600	1,500	1,150	1700/1600
Disk Reads MB/s	990	990	950	1600
Disk Writes MB/s	690	850	850	1300

Figure 8-2 DS4000 comparative specifications

Figure 8-3 illustrates the capabilities of the IBM System Storage EXP810 Storage Expansion Unit. It can be attached to the DS4700 and DS4800 disk storage controllers. The EXP810 16-bay disk enclosure can accommodate Fibre Channel and SATA disk drives.

DS4000 EXP810 Drive Enclosure – Overview

- Supports up to 16 drives in 3U enclosure
 - Up to 4.8 TB physical capacity per expansion unit using sixteen 300 GB disk drives
- Full 4 Gbps FC switched ESM supported
- Requires DS4000 Storage Manager v6.16 and higher
- Requires controller firmware 06.16 and higher
- Currently supported behind DS4700 and DS4800
- Dual Environmental Service Modules (ESM) - 4Gb FC
- Fibre Channel and SATA drives supported
- Field replaceable components
- Power and Cooling are integrated into the same CRU, redundant
- For Telco industry (Optional)
 - DC (-35 to -72 vDC) power supplies
 - Air filter option in optional front bezel
 - NEBS Level 3 certification

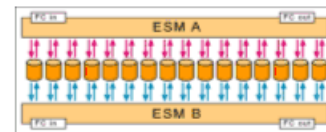


Figure 8-3 DS4000 EXP810 disk enclosures

8.1.2 DS4000 Storage Manager Software

The members of the DS4000 family are managed by the IBM DS4000 Storage Manager software. All of the members run the same robust, yet, intuitive, storage management software — designed to allow maximum utilization of the storage capacity and complete control over rapidly growing storage environments.

The IBM DS4000 Storage Manager is also designed to allow administrators to quickly configure and monitor storage from a Java™ technology-based Web browser interface as well as configure new volumes, define mappings, handle routine maintenance and dynamically add new enclosures and capacity to existing volumes. These storage management activities are done on an active DS4000, avoiding disruption to data access.

8.2 DS4000 copy functions

To understand the IBM DS4000 copy services functions, we introduce terminology to be used throughout this chapter:

- ▶ An *array* is a set of physical drives that the controller logically groups together according to the chosen RAID level protection to provide one or more *logical drives* (also called *LUNs*) to an application host or cluster.
- ▶ A *host* is a single system which can be contained in a host group.
- ▶ *Asynchronous Remote Mirroring* allows the primary disk system to acknowledge a host write request before the data has been successfully mirrored.
- ▶ *Consistency Group* is a group of disks which must be considered together in a remote copy operation. Write operations to the secondary disk system match the I/O completion order on the primary disk system for all volumes in the Consistency Group.
- ▶ *Suspend / Resume*. A resume operation synchronizes only the data written to the primary logical drive while the mirror was stopped (Resynchronization).
 - This includes a mirror that was manually halted (Suspend) or dropped due to unplanned loss of remote communication.
 - A new delta log tracks deltas to the primary volume during a communication interruption (planned or unplanned).
- ▶ Support for *dynamic mode switching* allows you to go back and forth between synchronous and asynchronous modes without breaking the remote mirror.
- ▶ The Primary Site is the location of the Primary disk system. This location with its systems provides the production data to hosts and users. Data is mirrored from here to the Secondary Site.
- ▶ The Secondary Site is the location of the Secondary disk system. This location with its systems holds the mirrored data. Data is mirrored from the Primary Site to here.
- ▶ A Mirror Repository Logical Drive is a *special logical drive* in the disk system created as a resource for the *controller owner* of the primary logical drives in a Mirror Relationship.

8.3 Introduction to FlashCopy

A FlashCopy logical drive is a point-in-time image of a logical drive. It is the logical equivalent of a complete physical copy, but it is created much more quickly than a physical copy. It also requires less disk space. On the other hand, it is not a real physical copy, because it does not copy all the data. Consequently, if the source logical drive is damaged, the FlashCopy logical drive cannot be used for recovery.

In the DS4000 Storage Manager, the logical drive from which you are basing the FlashCopy, called the *base logical drive*, can be a standard logical drive or the *secondary logical drive* in a Remote Mirror relationship. Typically, you create a FlashCopy so that an application (for example, an application to take backups) can access the FlashCopy and read the data while the base logical drive remains online and user-accessible. In this case, the FlashCopy logical drive is no longer required (it is usually disabled rather than deleted) after the backup completes.

You can also create multiple FlashCopies of a base logical drive and use the copies in write mode to perform testing and analysis. Before you upgrade your database management system, for example, you can use FlashCopy logical drives to test different configurations. Then you can use the performance data provided during the testing to help decide how to configure the live database system.

Important: For analysis, data mining, and testing without any degradation of the production logical drive performance, you should use FlashCopy in conjunction with VolumeCopy, as explained in 8.4, “Introduction to VolumeCopy”.

When you take a FlashCopy, the controller suspends I/O to the base logical drive for only a few seconds. Meanwhile, it creates a new logical drive called the *FlashCopy repository logical drive* where it stores FlashCopy metadata and copy-on-write data (Figure 8-4). It builds a metadata DB that contains only pointers. When the controller finishes creating the FlashCopy repository logical drive, I/O write requests to the base logical drive can resume.

However, before a data block on the base logical drive is modified, a copy-on-write occurs, copying the contents of blocks that are to be modified into the FlashCopy repository logical drive, for safekeeping. Subsequently, the corresponding pointer in the metadata DB changes. Since the FlashCopy repository logical drive stores copies of the original data in those data blocks, further changes to those data blocks write directly to the base logical drive without another copy-on-write. And, since the only data blocks that are physically stored in the FlashCopy repository logical drive are those that have changed since the time of the FlashCopy, the FlashCopy technology uses less disk space than a full physical copy.

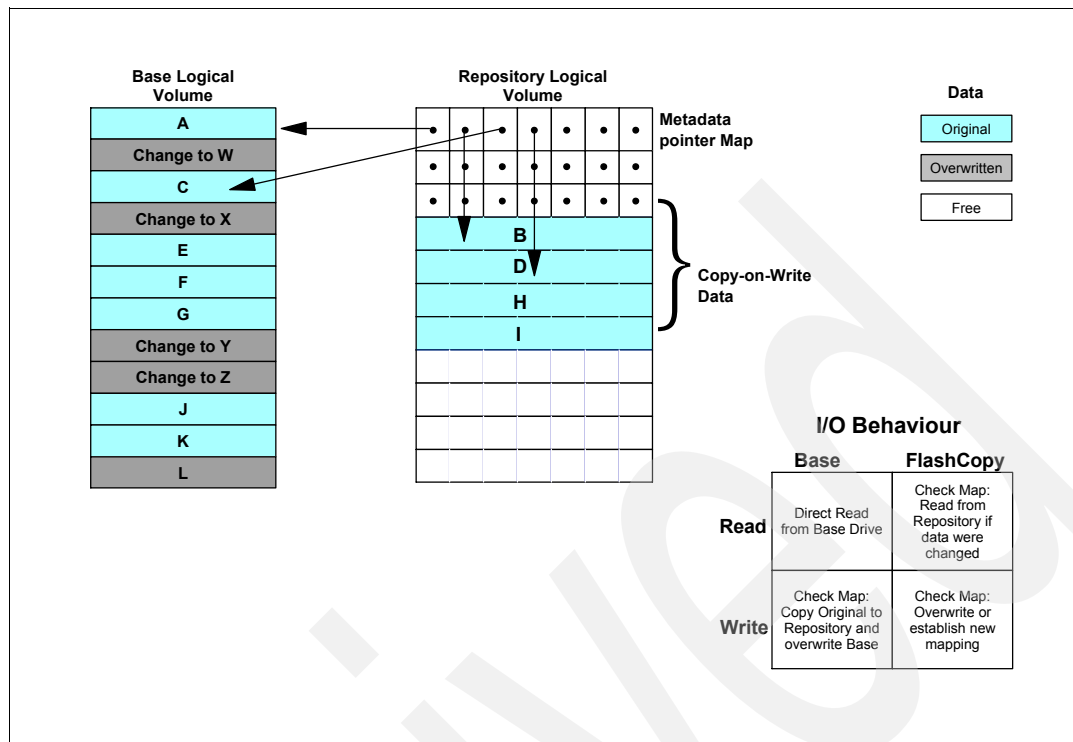


Figure 8-4 FlashCopy read and write schema

When you create a FlashCopy logical drive, you specify where to create the FlashCopy repository logical drive, its capacity, warning threshold, and other parameters. You can disable the FlashCopy when you are finished with it, for example, after a backup completes. The next time you re-create the FlashCopy, it reuses the existing FlashCopy repository logical drive. Deleting a FlashCopy logical drive also deletes the associated FlashCopy repository logical drive.

8.4 Introduction to VolumeCopy

The VolumeCopy premium feature is used to copy data from one logical drive (source) to another logical drive (target) in a single disk system (Figure 8-5). The target logical drive is an exact copy or *clone* of the source logical drive. This feature can be used to copy data from arrays that use smaller capacity drives to arrays that use larger capacity drives, to back up data, or to restore FlashCopy logical drive data to the base logical drive. The VolumeCopy premium feature includes a Create Copy Wizard, to assist in creating a logical drive copy, and a Copy Manager, to monitor logical drive copies after they have been created.

The VolumeCopy premium feature must be enabled by purchasing a feature key. FlashCopy must be installed as well. VolumeCopy is only available as a bundle that includes a FlashCopy license.

VolumeCopy is a full point-in-time replication. It allows for analysis, mining, and testing without any degradation of the production logical drive performance. It also brings improvements to backup and restore operations, making them faster and eliminating I/O contention on the primary (source) logical drive.

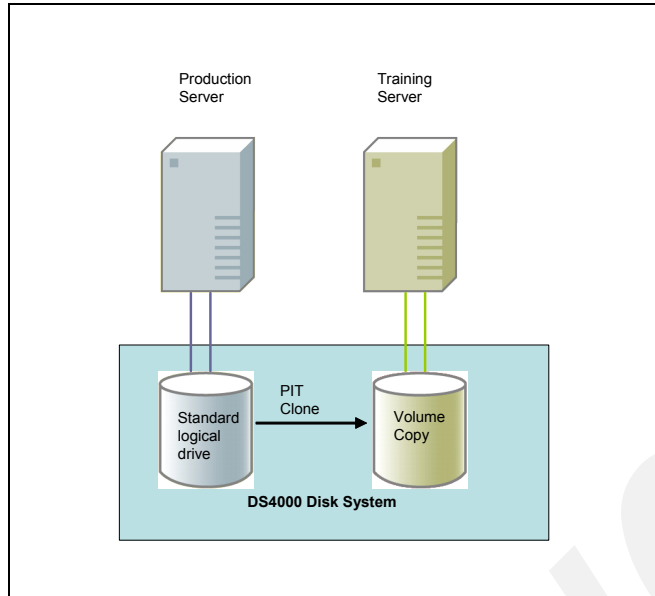


Figure 8-5 VolumeCopy

Copying data is a background operation managed by the controller firmware, which reads the source logical drive and writes the data to the target logical drive. If the Storage controller experiences a reset, the copy request is restored and the copy process resumes from the last known progress boundary.

After submitting a copy request, the source logical drive is only available for read I/O activity while a logical drive copy has a status of In Progress, Pending, or Failed. Write requests are allowed after the logical drive copy is completed. Read and write requests to the target logical drive do not take place while the logical drive copy has a status of In Progress, Pending, or Failed.

These restrictions are necessary to ensure the integrity of the point-in-time copy. If the logical drive being copied is large, this can result in an extended period of time without the ability for a production application to make updates or changes to the data.

Important: Because all write requests to the source logical drive are rejected when the VolumeCopy is taking place, it is essential to minimize the time it takes to complete the copy operation. This is why VolumeCopy must always be used in conjunction with FlashCopy; In other words, first make a FlashCopy of the source logical drive, then perform a VolumeCopy of the FlashCopy (see Figure 8-6).

As illustrated in Figure 8-6, FlashCopy, which allows a point-in-time copy to be made while maintaining read/write access, enables a complete copy to be created without interrupting the I/O activity of the production logical drive.

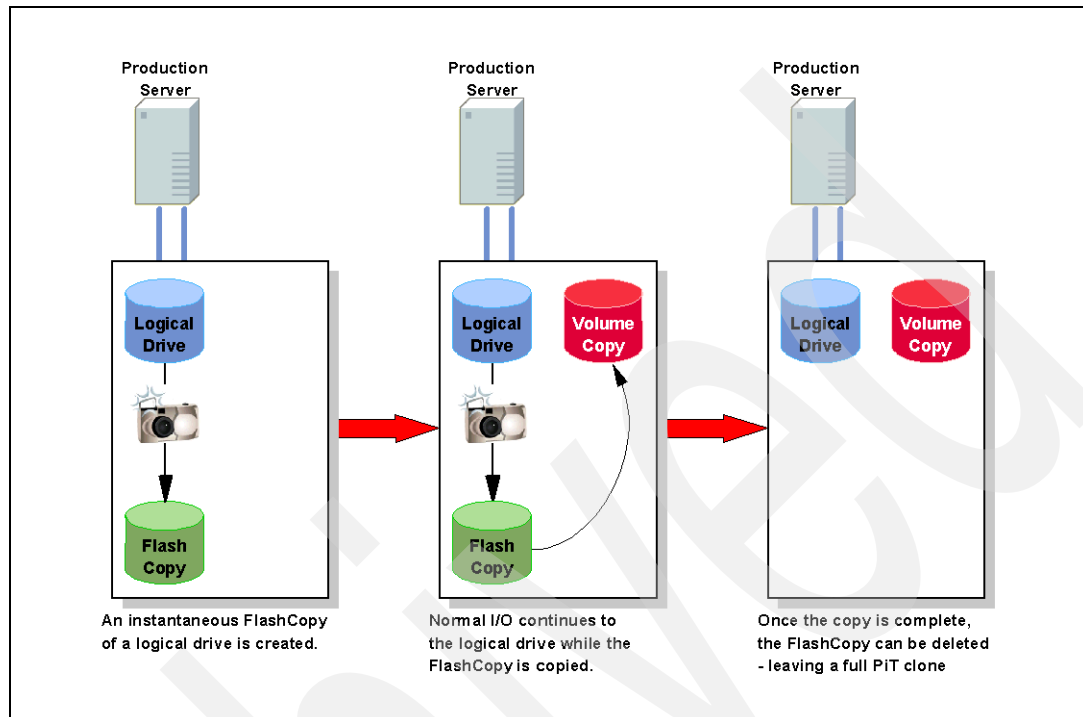


Figure 8-6 VolumeCopy integration with FlashCopy

8.5 Introduction to Enhanced Remote Mirroring

The Enhanced Remote Mirroring (ERM) option is a premium feature of the IBM DS4000 Storage Manager software. It is enabled by purchasing a premium feature key.

The Enhanced Remote Mirroring option is used for online, real-time replication of data between DS4000 systems over a remote distance (Figure 8-7). In the event of disaster or unrecoverable error at one DS4000, you can promote the second DS4000 to take over responsibility for normal I/O operations.

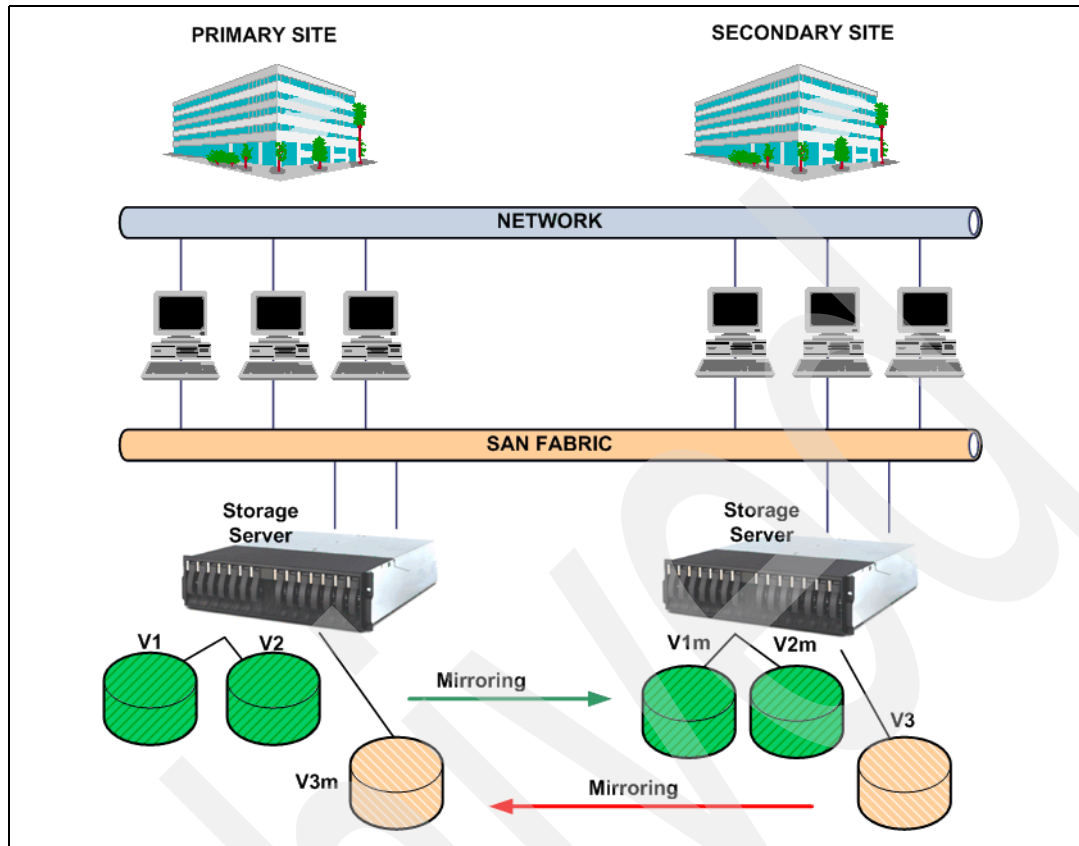


Figure 8-7 Enhanced Remote Mirroring

Data can be replicated between using different mirroring modes. Understanding how data flows between the systems is critical for setting the appropriate mirroring configuration. In the remainder of this section we explain the write operation sequences and processes involved for the different ERM mirroring modes. These are:

- ▶ Metro Mirroring
- ▶ Global Copy
- ▶ Global Mirroring

In all cases, data on the secondary logical drive of a mirror relationship can only be changed through the mirroring process. It cannot be changed by the host, or manually.

The read operations are identically treated in all three modes because, for a read request, there is no data exchange between the primary and secondary logical drives.

Tip: The mirroring mode can be changed at any time. This is called Dynamic Mode Switching.

8.5.1 Metro Mirroring (synchronous mirroring)

Metro Mirroring is a synchronous mirroring mode. This means that the controller does not send the I/O completion to the host until the data has been copied to both the primary and secondary logical drives.

When a primary controller (the controller owner of the primary logical drive) receives a write request from a host, the controller first logs information about the write request on the *mirror repository logical drive* (the information is actually placed a queue). In parallel, it writes the data to the primary logical drive. The controller then initiates a remote write operation to copy the affected data blocks to the secondary logical drive at the remote site. When the remote write operation is complete, the primary controller removes the log record from the mirror repository logical drive (deletes it from the queue). Finally, the controller sends an I/O completion indication back to the host system

Note: The owning primary controller only writes status and control information to the repository logical drive. The repository is not used to store actual host data.

When write caching is enabled on either the primary or secondary logical drive, the I/O completion is sent when data is in the cache on the site (primary or secondary) where write caching is enabled. When write caching is disabled on either the primary or secondary logical drive, then the I/O completion is not sent until the data has been stored to physical media on that site.

Figure 8-8 depicts how a write request from the host flows to both controllers to provide an instant copy.

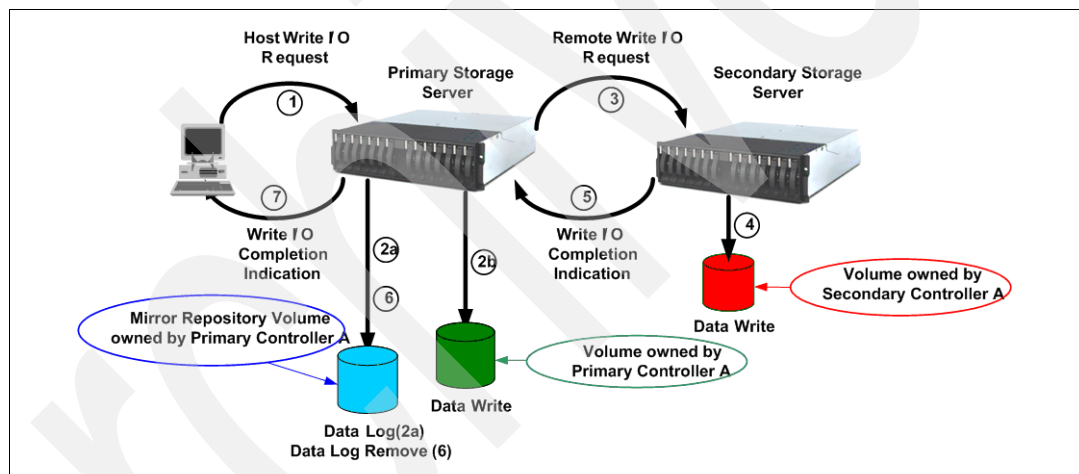


Figure 8-8 Metro Mirroring Mode (Synchronous Mirroring) data flow

When a controller receives a read request from a host system, the read request is handled on the primary disk system and no communication takes place between the primary and secondary disk systems.

8.5.2 Global Copy (asynchronous mirroring without write consistency group)

Global Copy is an asynchronous write mode. All write requests from host are written to the primary (local) logical drive and immediately reported as completed to the host system. Regardless of when data was copied to the remote disk system, the application does not wait for the I/O write request result from the remote site. However, Global Copy does not ensure that write requests at the primary site are processed in the same order at the remote site. As such, it is also referred as *asynchronous mirroring without write consistency group*.

When a primary controller (the controller owner of the primary logical drive) receives a write request from a host, the controller first logs information about the write request on the *mirror repository logical drive* (the information is actually placed in a queue). In parallel, it writes the data to the primary logical drive (or cache). After the data has been written (or cached), the host receives an I/O completion from the primary controller. The controller then initiates a background remote write operation to copy the corresponding data blocks to the secondary logical drive at the remote site. After the data has been copied to the secondary logical drive at the remote site (or cached), the primary controller removes the log record on the mirror repository logical drive (delete from the queue).

When multiple mirror relationships are defined on the disk system, the background synchronization of affected data blocks between the primary and secondary controller for the different relationships are conducted in parallel (a multi-threaded process). Thus, the write order for multiple volumes (for example write requests to a database volume and a database log volume on a database server) is not guaranteed with the Global Copy mode.

See Figure 8-9 for a logical view of the data flow.

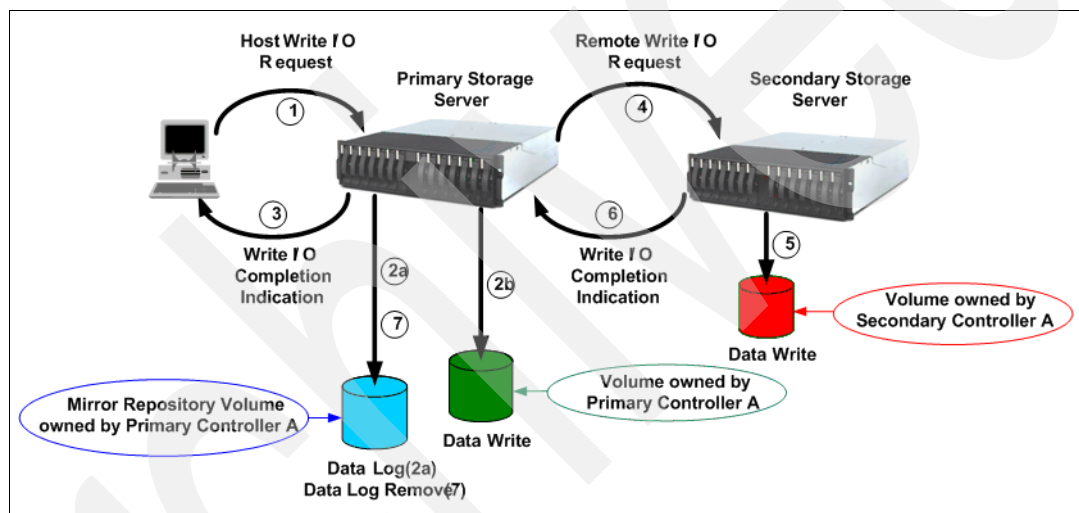


Figure 8-9 Global Copy Mode (Asynchronous Mirroring) data flow

When write caching is enabled on either the primary or secondary logical drive, the I/O completion is sent when data is in the cache on the site (primary or secondary) where write caching is enabled. When write caching is disabled on either the primary or secondary logical drive, then the I/O completion is not sent until the data has been stored to physical media on that site.

Note: The Mirror Repository logical drive can queue a number of I/O requests (up to 128 with firmware 6.1x.xx.xx). Until the maximum number has been reached, the mirrored pair state is said to be in a *Synchronized* state. If the maximum number of unsynchronized I/O requests is exceeded, the state of the mirrored pair changes to *Unsynchronized*.

The host can continue to issue write requests to the primary logical drive, but remote writes to the secondary logical drive do not take place. The requests are stored in the Remote Mirror repository on the primary site (*delta logging*).

When a controller receives a read request from a host system, the read request is handled on the primary disk system and no communication takes place between the primary and secondary disk systems.

8.5.3 Global Mirroring (asynchronous mirroring with write consistency group)

Global Mirroring is an asynchronous write mode where the order of host write requests at the primary site is preserved at the secondary site. This mode is also referred as asynchronous mirroring with write consistency group.

To preserve the write order for multiple mirrored volumes, Global Mirroring uses the *write consistency group* functionality. It tracks the order of the host write requests, queues them and sends to the remote controller in the same order.

Important: Selecting write consistency for a single mirror relationship does not change the process in which data is replicated. More than one mirror relationship must reside on the primary disk system for the replication process to change.

The volumes for which the write request order must be preserved have to be defined as members of a Write Consistency Group. The Write Consistency Group can be defined from the Storage Manager GUI.

When a primary controller (the controller owner of the primary logical drive) receives a write request from a host, the controller first logs information about the write on the *mirror repository logical drive*. It then writes the data to the primary logical drive. The controller then initiates a remote write operation to copy the affected data blocks to the secondary logical drive at the remote site. The remote write request order corresponds to the host write request order.

After the host write to the primary logical drive is completed and the data has been copied to the secondary logical drive at the remote site, the controller removes the log record from the mirror repository logical drive. Refer to Figure 8-10 for a logical view of the data flow.

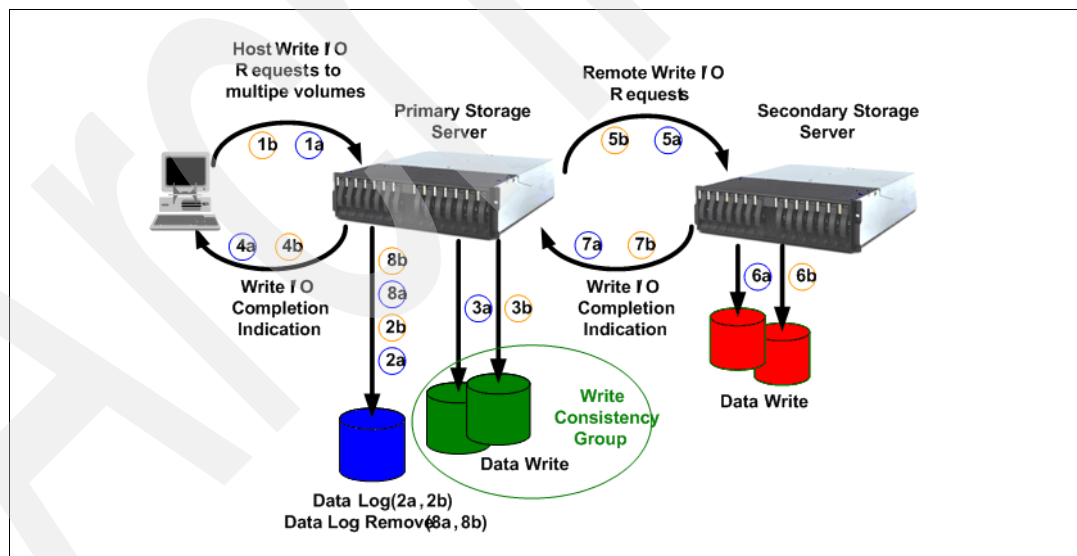


Figure 8-10 Global Mirroring logical data flow

When write caching is enabled on either the primary or secondary logical drive, the I/O completion is sent as soon as data is in the cache on the site (primary or secondary) where write caching is enabled. When write caching is disabled then the I/O completion is not sent until the data has been stored to physical media.

Note: The Mirror Repository logical drive can queue a number of I/O requests (up to 128 with firmware 6.1x.xx.xx). Until the maximum number has been reached, the mirrored pair state is said to be in a *Synchronized* state. If the maximum number of unsynchronized I/O requests is exceeded, the state of the mirrored pair changes to *Unsynchronized*.

The host can continue to issue write requests to the primary logical drive, but remote writes to the secondary logical drive do not take place. The requests are stored in the Remote Mirror repository on the primary site (*delta logging*).

Whenever the data on the primary drive and the secondary drive becomes unsynchronized, the controller owner of the primary drive initiates a changed data synchronization.

Suspend Mirror and Resume Mirror capability

This function allows you to suspend the mirroring process independently of the Mirroring Mode. While in Suspended State, the secondary disk system no longer receives any write I/Os from the primary disk system, and all data blocks that change are logged in a special volume (logical drive) called the Mirroring Repository Volume (see 8.6, “Mirror repository logical drives” on page 307).

Since the data on the secondary logical drive is not changing, you can access the “frozen” secondary logical volume and use it for test purposes or to back it up.

To resume the mirroring, you invoke the Resume Mirror function. It resynchronizes the changed data between the primary and the secondary logical drives. No full synchronization has to take place.

Change Write Mode option

You can switch among the different mirroring modes at any time, for an established mirror relationship. This is called Dynamic Mode Switching. You can switch between:

- ▶ Metro Mirroring (synchronous write mode)
- ▶ Global Copy (asynchronous write mode without Write Consistency Group)
- ▶ Global Mirroring (asynchronous write mode with Write Consistency Group)

8.6 Mirror repository logical drives

A mirror repository logical drive is a *special logical drive* in the disk system created as a resource for the controller owner of the primary logical drive in a remote logical drive mirror. Two mirror repository logical drives (one for each controller in a subsystem) are automatically created when activating the ERM Premium Feature. This is shown in Figure 8-11.

One mirror repository drive is created for each storage controller. The mirror repository logical drive stores (queues) the mirroring information, including information about the remote write request that has not yet been written to the secondary logical drive. After a confirmation of a given write request has occurred, the corresponding entry stored in the mirror repository logical drive is removed.

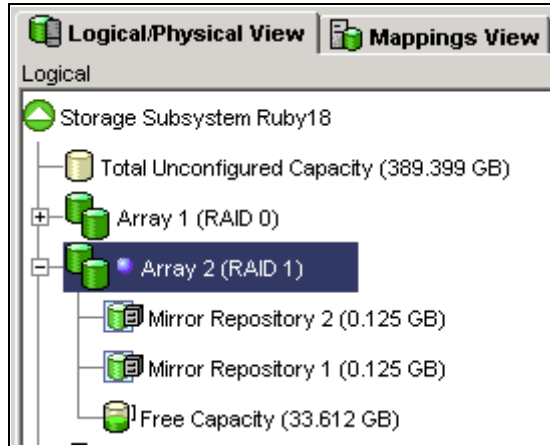


Figure 8-11 Mirror Repository Logical Drives after activating ERM feature

The mirroring process for *all* primary drives defined on a storage controller is monitored by the corresponding controller's mirror repository drive.

Notes: Two mirror repository logical drives are created, one for each controller, in every DS4000 with Enhanced Remote Mirroring activated.

No actual host data is written to the repository logical drive. It is only used for status and control data in relation to the Enhanced Remote Mirroring relationships.

The capacity is set at 128 MB for each logical drive. The segment size is set at 32 KB (or 64 blocks). The controller determines the modification priority. The drive size, segment size, and modification priority for a mirror repository logical drive cannot be changed.

8.7 Primary and secondary logical drives

To create an ERM relationship, a mirrored logical drive pair is defined consisting of a primary logical drive at the primary disk system and a secondary logical drive at a remote disk system.

A standard logical drive can only be defined in one mirrored logical drive pair.

Figure 8-12 shows that both the primary and secondary logical drives are displayed at the primary site, while only the secondary logical drive is displayed at the secondary site.

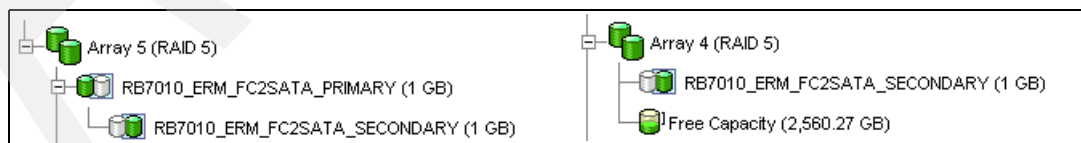


Figure 8-12 Primary and secondary volumes on the primary (right) and secondary sites (left)

8.7.1 Logical drive parameters, roles, and maximum number of mirrored pairs

The primary or secondary role is determined at the logical drive level, not the disk system level. Any given logical drive, however, can exclusively be in either primary or secondary role. In other words, a DS4000 disk system can have a combination of logical drives in a primary role and other logical drives in a secondary role.

The maximum number of mirror relationships that can be defined on a DS4000 is the number of established mirrored pairs. The number of pairs is independent of the role — primary or secondary — of the logical drives. There is no restriction on the number of logical drives defined on a DS4000. For example, you can define 32 primary and 32 secondary logical drives on a DS4800, or any other combination (1 primary and 63 secondary, 2 primary and 62 secondary, etc.), up to the allowed maximum of 64 mirrored pairs on the DS4800.

Tip: If you are upgrading the Remote Volume Mirroring Premium Feature (SM V8.4) to the Enhanced Remote Mirroring Premium Feature (SM 9.10), there are initially a restriction to a maximum of 32 mirroring pairs on the DS4400 or DS4500. This is caused by the Mirror Repository Drive size limit. You can change the size of the Mirror Repository Drive through the DS4000 Storage Manager and set it to the required size of 128 MB per storage controller and thus increase the limit of maximum mirroring pairs to 64. See 8.6, “Mirror repository logical drives” on page 307.

Characteristics such as RAID level, caching parameters, and segment size can differ between the primary and secondary logical drives of an ERM relationship. The mirrored logical drive can be configured on different Expansion Enclosures types. It is also possible to mirror a logical drive in an EXP710 to a logical drive in an EXP810.

There are, however, some restrictions:

- ▶ Only a standard logical drive can participate in a mirroring relationship.
- ▶ The secondary logical drive has to be at least as large as the primary logical drive.
- ▶ Any given standard logical drive can participate in only one mirror relationship.

Note that mirror pairs can be established between more than two DS4000's.

8.7.2 Host Accessibility of secondary logical drive

When you first create the mirror relationship, all data from the primary logical drive is copied to the remote logical drive (full synchronization).

During the full synchronization, the primary logical drive remains accessible for all normal host I/Os. The secondary logical drive can be mapped to a host, and read access is possible. The read/write behavior changes automatically if there is a role reversal of the logical drive (this applies in both directions — from secondary to primary and vice versa). Keep in mind, however, that some operating systems like Windows dynamic disks do not support mapping of identical drives to the same host.

8.7.3 Mirrored logical drive controller ownership

The logical drives belonging to Controller A in the primary Storage Server must be mirrored to the logical drives owned by Controller A in the secondary subsystem. The same rule applies for the logical drives owned by Controller B.

A primary controller only attempts to communicate with its matching controller in the secondary disk system. The controller (A or B) that owns the primary logical drive determines the controller owner of the secondary logical drive. If the primary logical drive is owned by Controller A on the primary site, the secondary logical drive is therefore owned by Controller A on the secondary side. If the primary Controller A cannot communicate with the secondary Controller A, no controller ownership changes take place, and the Remote Mirror link is broken for that logical drive pair.

If an ownership change of the logical drive on the primary site occurs (caused by an I/O path error or administrator interaction), an ownership of the logical drive on the remote controller takes place with the first write request (from the primary controller to the secondary controller) issued through the mirror connection — when the ownership transfer on the secondary side occurs and there is no “Needs Attention” status displayed.

The logical drive ownership of the secondary controller cannot be changed by either the host or through the GUI: It is entirely controlled by the primary side.

8.7.4 Enhanced Remote Mirroring and FlashCopy Premium Feature

With the latest ERM Premium Feature, you can take a FlashCopy of a primary and secondary logical drive (since read access to the secondary logical drive is now possible). Role reversals that cause a primary logical drive to reverse to a secondary logical drive do not fail any associated FlashCopy.

For data consistency, you should protect the FlashCopy source from write I/Os. By flashing a logical volume, you have to stop the application and flush the write cache entries. By flashing the secondary logical volume, you first have to stop the application and flush the write cache entries and then suspend the mirror relationship (once the secondary is in sync). The application functionality can be restored and a FlashCopy of the secondary drive can be made, after which the mirror is resumed.

8.7.5 Enhanced Remote Mirroring and VolumeCopy Premium Feature

A primary logical drive can be a source or target logical drive in a VolumeCopy. A secondary logical drive cannot be a source or target logical drive unless a role reversal was initiated after the copy had completed. If a role reversal is initiated during a copy in progress, the copy fails and cannot be restarted.

Attention: Starting a VolumeCopy for a given primary logical drive sets it to *read only* access. The host application or host OS cannot issue write requests to this resource.

8.7.6 Volume role compatibility

Table 8-1 shows the dependencies of Volume Roles on Copy Services.

Table 8-1 Volume role compatibility

	It can become:							
If volume is:	FlashCopy Volume	FlashCopy Base Volume	FlashCopy Repository	Primary Logical Drive	Secondary Logical Drive	Mirror Repository Logical Drive	Volume Copy Source	Volume Copy Target
FlashCopy Volume	NO	NO	NO	NO	NO	NO	YES	NO
FlashCopy Base Volume	NO	YES	NO	YES	NO	NO	YES	YES
FlashCopy Repository	NO	NO	NO	NO	NO	NO	NO	NO
Primary Logical Drive	NO	YES	NO	NO	NO	NO	YES	YES
Secondary Logical Drive	NO	YES	NO	NO	NO	NO	NO	NO
Mirror Repository Logical Drive	NO	NO	NO	NO	NO	NO	NO	NO
VolumeCopy Source	NO	YES	NO	YES	NO	NO	YES	YES
VolumeCopy Target	NO	NO	NO	YES	NO	NO	YES	YES

8.8 Data resynchronization process

If a link interruption or logical drive error prevents communication with the secondary disk system, the controller owner of the primary logical drive transitions the mirrored pair into an *Unsynchronized* status and sends an I/O completion to the host that sent the write request. The host can continue to issue write requests to the primary logical drive, but remote writes to the secondary logical drive do not take place. The requests are stored in the Remote Mirror repository on the primary site (delta logging).

When connectivity is restored between the controller owner of the primary logical drive and the controller owner of the secondary logical drive, a resynchronization takes place.

If the mirroring state is changed to *Suspended* state, the host can also continue to issue write requests to the primary logical drive.

There is an essential difference between the *Unsynchronized* and the *Suspended* states. The first is an error condition indication, while the second is an administrator provided status change. The behavior for data resynchronization is different for these two states.

When in *Suspended* state, the administrator must manually resume the mirroring to return to a *Synchronized* state. The *Unsynchronized* state can either be manually or automatically changed into a *Synchronized* state.

Notes: Here are some things to keep in mind:

- ▶ Data Resynchronization is required when a mirrored pair has become Unsynchronized.
- ▶ The Suspended state is a subset of the Unsynchronized state. Specification of the Unsynchronized and Suspended states let you differentiate between error (Unsynchronized) and maintenance (Suspended) conditions of a given mirrored logical drive pair.
- ▶ Only the blocks of data that have changed on the primary logical drive during Unsynchronized or Suspended state are copied to the secondary logical drive.

Normally, when resuming a mirror relationship or re-establishing the communication between the subsystems, only changed data blocks are sent to the remote site. However, there are some cases in which full synchronization of the primary and secondary logical drive is required:

- ▶ Establishing a *new* mirror relationship between two given logical drives
- ▶ Any kind of total failure of the mirror relationship members.
- ▶ Any kind of mirror repository logical drive failure
- ▶ Change of all data block track entries in the mirror repository logical drive while any kind of mirroring communication errors occurred
- ▶ Change of all data block track entries in the mirror repository logical drive in suspended state of the mirror relationship

Note: Information about changed data blocks (delta logging) is logged in the Mirror Repository Logical Drive. The resynchronization process uses this log to send only the changed data to the remote site. If during the interruption of the mirroring, all data blocks on the primary repository logical drive were changed, a full synchronization takes place. The time it takes to change all data blocks on the repository logical drive is dependent upon the number of I/O write requests, application write behavior and the capacity of the Mirror Repository logical drive.

Manual resynchronization

This is the recommended method for resynchronization of an Unsynchronized mirror pair, because it allows you to manage the resynchronization process in a way that provides the best opportunity for recovering data.

Automatic resynchronization

Automatic resynchronization starts automatically after the controller detects that communication is restored for an unsynchronized mirrored pair. When the Automatic Resynchronization option is selected and a communication failure occurs between the primary and secondary disk systems, the controller owner of the primary logical drive starts resynchronizing the primary and secondary logical drives immediately after detecting that communication has been restored.

Important: Any communication disruptions between the primary and secondary disk system while resynchronization is underway could result in a mix of new and old data on the secondary logical drive. This would render the data unusable in a disaster recovery situation.

8.9 SAN fabric and TCP/IP connectivity

There are some important requirements and rules to follow in regards to Fibre Channel connections and SAN fabric attachment for a correct implementation of ERM. SAN planning is a critical task and must include SAN ports, SAN zoning, and cabling considerations.

This section reviews these SAN considerations and also addresses the TCP/IP management network configuration.

8.9.1 SAN fabric and SAN zoning configuration

Here, we examine the SAN requirements from a general, conceptual standpoint. Detailed information and procedures for implementing the SAN fabric, configuring SAN switches, and SAN zones are beyond the scope of this book.

SAN fabric configuration

Dedicated Remote Mirroring ports (A2 and B2 host side controller ports) must be attached to a SAN fabric with support for the Directory Service and Name Service interfaces. In other words, there must be at least one SAN Switch with SAN zoning capability installed in the mirroring environment. Since ERM is typically for providing a High Availability Solution, we strongly recommend that you use at least two switches in the SAN. See Figure 8-13 for SAN fabric configuration examples.

Tip: IBM and other SAN switch vendors recommend configuring two SAN fabrics with independent SAN zones for the highest level of redundancy when implementing ERM.

SAN fabric zoning

It is also mandatory to keep ERM links and the Host links in separate SAN Zones. We recommend that you create separate zones within the fabric (assuming that only two Storage Subsystems are involved):

- ▶ The first zone, for a host connection to Controller A
- ▶ The second zone, for a host connection to Controller B
- ▶ The third zone, for Controller A Remote Mirroring links between Storage Subsystems
- ▶ The fourth zone, for Controller B Remote Mirroring links between Storage Subsystems

Note: You can have multiple zones for host connections on each controller, based upon the number of connections available. For example, the DS4800 permits up to three host connections per controller, reserving the fourth for the Remote Mirror link.

There is no special recommendation for defining the SAN zones: It can either be based on switch ports or WWPNs. However, for easier maintenance and troubleshooting, do not use mixed zoning definitions (ports and WWPN zones simultaneously); or, at least, differentiate these across the Fabrics (one Fabric with Port Zoning, one with WWPN Zoning).

Figure 8-13 shows one example of zoning a fabric for DS4500 and Enhanced Remote Mirroring.

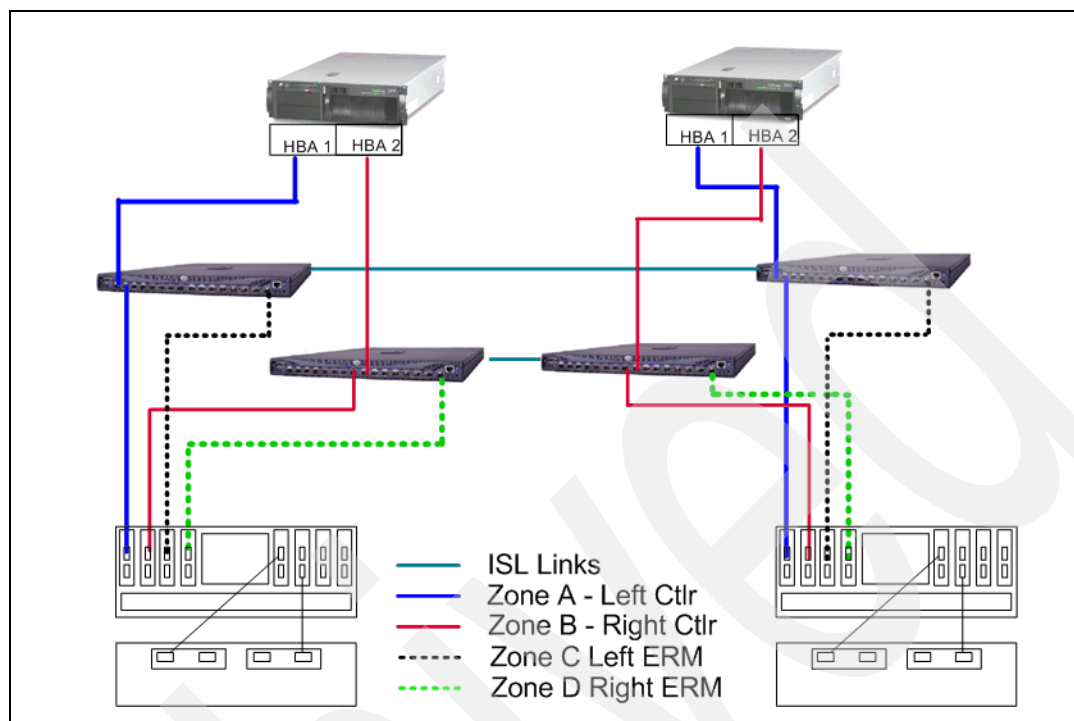


Figure 8-13 Enhanced Remote Mirroring zoning example

DS4000 Storage Server Fibre Channel configuration for ERM

The Enhanced Remote Mirroring option requires two dedicated controller host port connections for each disk system that participates in Enhanced Remote Mirroring.

When the ERM option is *activated* on a disk system, the last *host-side* controller ports become dedicated to Remote Mirror operations:

- ▶ On the DS4800 and DS4700 model 72, you use the 4th host port on controllers A and B for ERM (see Figure 8-14 and Figure 8-16 for the host port locations).
- ▶ The host port locations for the DS4700 model 70 are shown in Figure 8-15.

After ERM activation, the last host-side controller ports no longer permit host system I/O requests. The persistent reservations are also removed from these ports. The last host-side controller ports are *only* able to communicate to other disk systems that have the ERM option activated *and* are connected by the same fabric with proper SAN zoning configuration.

The level of redundancy within the fabric depends on the fabric design and Fibre Channel switch configuration. This book does not specifically address SAN design issues — you can find more information in other IBM Redbooks such as *IBM TotalStorage: SAN Product, Design, and Optimization Guide*, SG24-6384, *IBM System Storage: Implementing an Open IBM SAN*, SG24-6116, and *Introduction to Storage Area Networks*, SG24-5470.

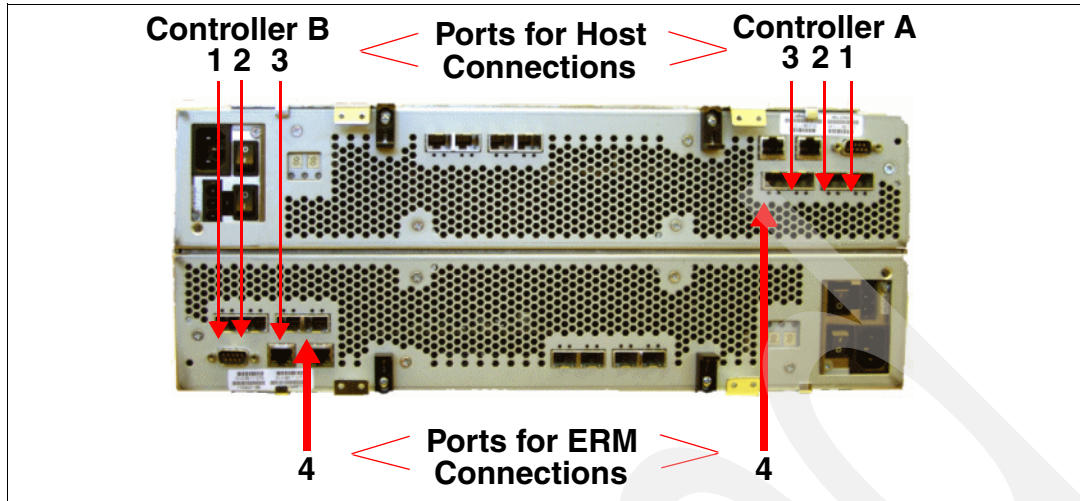


Figure 8-14 Host-side ports for host and ERM connections on the DS4800

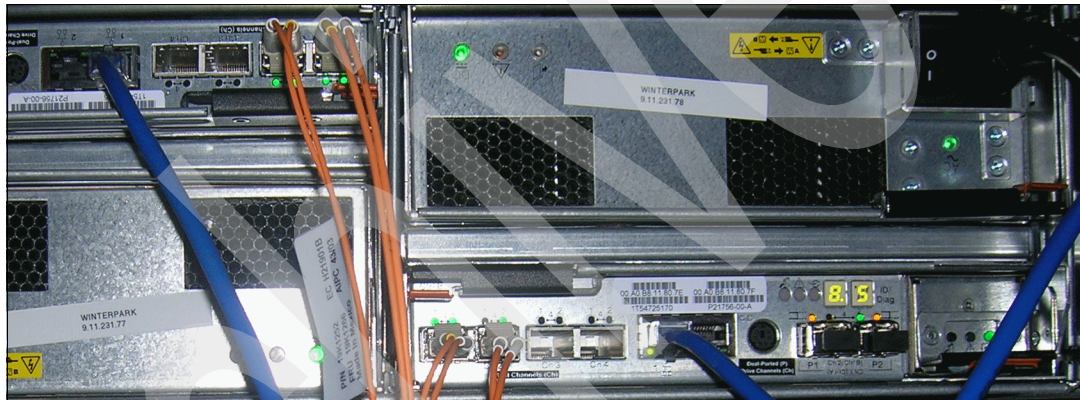


Figure 8-15 Host-side ports for host and ERM connections on DS4700 model 70

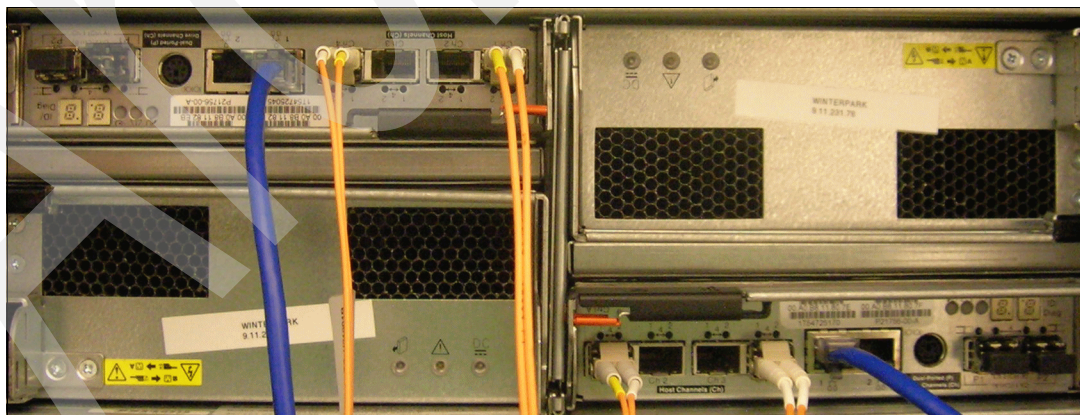


Figure 8-16 Host-side ports for host and ERM connections on DS4700 model 72

Fibre Channel distance limits

The distance between primary and remote disk systems is normally limited by the distance limitations of Fibre Channel inter-switch links (ISLs). See Table 8-2.

With the Global Copy and Global Mirror Remote Mirroring operating modes, it is now possible to establish mirroring for distances of more than 5000 km (3200 mi). For more information, refer to 8.12, “Long-distance ERM” on page 322.

Table 8-2 Fibre Channel distance limits

Fiber cable type	Laser type	Distance limit in kilometers	Distance limit in miles
Single mode 9 micron	Long wave	10 km	6.25 miles
Multi mode 50 micron	Short wave	0.5 km	0.32 mile

Important: The maximum distance for the short wave SFPs has impact on the maximum connection bandwidth. The SFP automatically changes the connection speed to 1 Gbps if the maximum distance for a 2Gbps connection is exceeded (and down to 2 Gbps if the maximum distance for a 4Gbps connection is exceeded).

For the short wave SFP, the maximum length for a 4 Gbps connection is 150 m. Refer to Table 8-3.

Table 8-3 Multi-mode fibre types and connection bandwidth

Fiber type	Speed	Maximum distance
50 micron MMF (shortwave)	1Gbps	500 m
50 micron MMF (shortwave)	2Gbps	300 m
50 micron MMF (shortwave)	4Gbps	150 m
62.5 micron MMF (shortwave)	1Gbps	175 m/300 m
62.5 micron MMF (shortwave)	2Gbps	90 m/150 m

Note that the maximum distances listed in Table 8-3 can be reduced due to low quality fiber, poor terminations, patch panels, etc.

8.10 ERM and disaster recovery

As modern business pressures increasingly require 24-hour data availability, system administrators are required to ensure that critical data is safeguarded against potential disasters. Additionally, storage administrators are searching for methods to migrate from one host to another or from one disk system to another, with as little disruption as possible.

Remote Mirroring is one method that can be implemented to assist in business continuity and disaster recovery. Once critical data has been mirrored from a primary disk system to a remote disk system, primary and secondary logical drives can have their roles reversed so that the copied data can be accessed from the remote disk system.

Next we discuss how to reverse the roles of the primary and secondary logical drives.

8.10.1 Role reversal concept

A role reversal promotes a selected secondary logical drive to become the primary logical drive of the mirrored pair. As previously explained, the roles of primary and secondary logical drives are naming conventions based on the direction of data flow. They differentiate as follows:

- ▶ The relevant administrative commands for ERM must be provided on the primary site.
- ▶ The Mirror States are determined by the primary disk system.
- ▶ The connection examination is provided by the primary disk system.
- ▶ Secondary logical drive is only read-access enabled.

A role reversal is performed using one of the following methods.

Changing a secondary logical drive to a primary logical drive

Use this option to perform a role reversal between the two paired logical drives in a mirror relationship. This option promotes a selected secondary logical drive to become the primary logical drive of the mirrored pair. If the associated primary logical drive can be contacted, the primary logical drive is automatically demoted to be the secondary logical drive. Use this option when a normally interruptible maintenance task on the primary site is required or in a case of an unrecoverable failure to the disk system that contains the primary logical drive and you want to promote the secondary logical drive so that hosts can access data and business operations can continue.

Important: When the secondary logical drive becomes a primary logical drive, any hosts that are mapped to the logical drive through a logical drive-to-LUN mapping are now able to write to the logical drive.

If a communication problem between the secondary and primary sites prevents the demotion of the primary logical drive, an error message is displayed. However, you are given the opportunity to proceed with the promotion of the secondary logical drive, even though this leads to a Dual Primary Remote Mirror status condition.

Changing a primary to a secondary logical drive

Use this option to perform a role reversal between the two paired logical drives in a Remote Mirror. This option demotes a selected primary logical drive to become the secondary logical drive of the mirrored pair. If the associated secondary logical drive can be contacted, the secondary logical drive is automatically promoted to be the primary logical drive. Use this option for role reversals during normal operating conditions. You can also use this option during a Recovery Guru procedure for a Dual Primary Remote Mirror status condition.

Important: Any hosts that are mapped to the primary logical drive through a logical drive-to-LUN mapping are no longer able to write to the logical drive.

8.10.2 Re-establishing Remote Mirroring after failure recovery

When the damaged site is back online and properly configured, mirror relationships can be resumed. Re-create a mirror relationship by completing the following steps:

1. Ensure that SAN connections and SAN zoning are properly configured.
2. From the active secondary site, define a mirror relationship using the logical drive on the recovered primary site as the secondary logical drive.
3. Ensure that storage partitioning and host mapping is properly defined on the recovered primary site so that it can take over normal operation from the secondary site.

4. Ensure that the host software is properly configured so that the host systems at the recovered primary site can take over I/O from the secondary site host systems.
5. After the full synchronization has completed, perform a manual role reversal so that the recovered primary site now possesses the active primary logical drive, and the secondary logical drive now exists on the secondary site. For more information see “Changing a primary to a secondary logical drive” on page 317.

8.10.3 Link interruptions

Loss of communication can be caused by FC link failure, but also by other hardware errors.

Fibre Channel mirror link interruptions in synchronous write mode

In synchronous mode, if the link is interrupted and the primary controller receives a write request from an attached host, the write request cannot be transferred to the secondary logical drive and the primary and secondary logical drives are no longer appropriately mirrored. The primary controller transitions the mirrored pair into *unsynchronized* state and sends an I/O completion to the primary host. The host can continue to write to the primary logical drive but remote writes do not take place.

When connectivity is restored between the controller owner of the primary logical drive and the controller owner of the secondary logical drive, depending on the configured resynchronization method, either a *automatic resynchronization* takes place or a manual resynchronization must be performed. Only the data in changed blocks is transferred during the resynchronization process. The status of the mirrored pair changes from an *Unsynchronized* state to a *Synchronization-in-Progress* state.

Note: A loss of communication between the primary and secondary do not result in the controllers attempting to change ownership of drives. The only time ownership changes is on a host path failover. This results in the secondary mirror to change ownership to match the primary on the next write I/O.

Mirror link interruptions in asynchronous write mode

In asynchronous mode, if the link is broken, the primary controller periodically attempts to reestablish the connection to the secondary controller.

The Mirror Repository Logical Drive queues a number I/O requests until the maximum number of write requests that could not be sent to the secondary subsystem has been reached. While requests are being queued, the mirrored pair remains in a Synchronized state. If the maximum number of unsynchronized I/O requests is exceeded, the state of the mirrored pair changes to Unsynchronized.

The host can continue to write to the primary logical drive but remote writes do not take place. If the link is recovered, depending on Synchronizing Settings, the resynchronization starts automatically or must be started through an administrative command (resume mirroring).

8.10.4 Secondary logical drive error

The primary controller also marks the mirrored pair as unsynchronized when a logical drive error on the secondary site prevents the remote write from completing. For example, an offline or a failed secondary logical drive can cause the Enhanced Remote Mirror to become Unsynchronized. The host can continue to write to the primary logical drive but remote writes do not take place.

After the logical drive error is corrected (the secondary logical drive is placed online or recovered to an Optimal state), then the resynchronization process can begin. Depending on Synchronizing Settings, the resynchronization starts automatically or must be started manually by issuing an administrative command (resume mirroring).

8.10.5 Primary controller failure

If a remote write is interrupted by a primary controller failure before it can be written to the secondary logical drive, the primary disk system provides controller ownership change from the preferred controller owner to the alternate controller in the disk system, the first write request to the remote site changes the ownership of the secondary logical drive. Once the transition of the ownership is completed the mirroring is proceeding as usually.

8.10.6 Primary controller reset

If a remote write is interrupted by a primary controller reset before it can be written to the secondary logical drive, there is normally no ownership change. After reboot the controller reads information stored in a log file in the mirror repository logical drive and uses the information to copy the affected data blocks from the primary logical drive to the secondary logical drive. We highly recommended suspending the mirror relationships before resetting the controller.

8.10.7 Secondary controller failure

If the secondary controller fails, the primary site can no longer communicate with the secondary logical drive. The mirror state becomes Unsynchronized. The host can continue to write to the primary logical drive but remote writes do not take place. After the secondary controller failure has been recovered, depending on the synchronization settings the primary controller changes the mirror state to Synchronizing either automatically or through a Resume command.

8.10.8 Write Consistency Group and Unsynchronized State

Mirror Relationships configured within a Write Consistency Group are considered an interdependent set with respect to the integrity of the data on the remote logical drives. A given secondary logical drive cannot be considered fully synchronized until all members of the Write Consistency Group are synchronized. Similarly, when one mirrored logical drive pair of the Write Consistency Group is in the Unsynchronized state, all members of the group halt remote write activity to protect the consistency of the remote data.

The controller changes the state of all mirrored pairs within the group to Unsynchronized, when any mirror within the set becomes Unsynchronized or Failed. All pending write requests for the member of the Write Consistency Group are moved to the Mirror Repository Logical Drive. Mirror relationships configured with the manual resynchronization setting remain in the Suspended state until a user intervenes by issuing a Resume command. Mirror relationships configured to allow automatic resynchronization automatically resynchronize.

8.11 Performance considerations

This section contains general performance considerations that apply only in the context of designing a Remote Mirror configuration. There are many factors that come into play, such as host operating system settings and different application requirements and it usually takes some experimentation and observation overtime to tune a configuration for better performance in a given environment.

8.11.1 Synchronization priority

The controller owner of a primary logical drive performs a synchronization in the background while processing local I/O write operations to the primary logical drive and associated remote write operations to the secondary logical drive.

Because the synchronization diverts controller processing resources from I/O activity, it can have a performance impact to the host application. The synchronization priority defines how much processing time is allocated for synchronization activities relative to system performance. The following priority rates are available:

- ▶ Lowest
- ▶ Low
- ▶ Medium
- ▶ High
- ▶ Highest

You can use the synchronization priority to establish how the controller owner prioritizes the resources required for synchronization process relative to host I/O activity.

The following rules are a rough estimates for the relationships between the five synchronization priorities.

Attention: Notice that logical drive size can cause these estimates to vary widely.

- ▶ A synchronization at the lowest synchronization priority rate takes approximately eight times as long as a synchronization at the highest synchronization priority rate.
- ▶ A synchronization at the low synchronization priority rate takes approximately six times as long as a synchronization at the highest synchronization priority rate.
- ▶ A synchronization at the medium synchronization priority rate takes approximately three and a half times as long as a synchronization at the highest synchronization priority rate.
- ▶ A synchronization at the high synchronization priority rate takes approximately twice as long as a synchronization at the highest synchronization priority rate.

Note: The lowest priority rate favors system performance, but the synchronization takes longer. The highest priority rate favors the synchronization, but system performance might be compromised. Logical drive size and host I/O rate loads affect the synchronization time comparisons. A synchronization at the lowest synchronization priority rate takes approximately eight times as long as a synchronization at the highest synchronization priority rate.

8.11.2 Synchronization performance and logical drive settings

In a configuration where both mirrored disk systems are holding primary and secondary logical drives (cross-mirroring) the count of host I/O requests to a local disk system and the synchronization I/O requests to the remote disk system are the two factors influencing the system's overall performance.

The write performance of a given logical drive is estimated by four key factors:

- ▶ Logical drive cache settings
- ▶ Performance of a single physical drive in the array in which the logical drive is created
- ▶ Number of physical drives in the array in which the logical drive is created
- ▶ RAID Level of the array in which the logical drive is created

A good practice is to balance resources among the mirrored logical drives. That particularly means:

- ▶ Write cache settings should be the same on both disk systems.
- ▶ Do not use Read Cache for secondary logical drives.
- ▶ Use as many drives as you can get to create an array.
- ▶ If the write cache is disabled for the secondary logical drive use more disks (than on the primary site) to create the array for the secondary logical drive.
- ▶ Use RAID Level 1 for secondary logical drives receiving data from RAID Level 3 or 5 primary logical drive to balance the performance.

With the possibility for FC/SATA Drive Intermixing, you might want to mirror data from FC drives to SATA drives (previous points still remaining).

8.11.3 Mirroring mode and performance

The mirroring mode has an impact on the disk system performance. Remember if you have intensive host I/Os that:

- ▶ Metro Mirroring is a “must synchronize” mode.
- ▶ Global Copy and Global Mirror are a “can synchronize” modes.
- ▶ Global Mirror requires more storage controller processing resources to enforce the host I/O write request order.

8.11.4 Mirroring connection distance and performance

For distances longer than 10 km (6.25 mi), you assume that asynchronous mirroring only is used.

The maximum distance that can be supported whether using short or long wave SFPs are greater in a 1Gbps fabric than in 2Gbps (refer to Table 8-2 on page 316). However, using a 1-Gbps connection, rather than 2Gbps, negatively impacts the synchronization performance.

The major performance factor for long distance solutions is the IP connection bandwidth between the Fibre Channel Extenders. If you connect to a 100 Mbps IP network, you can transfer about 10 MB of data per second (assuming the connection is stable).

- ▶ Estimate the time required for the first full synchronization.
- ▶ Check the amount of data that is changing on average on the primary logical drives over a certain period of time and compare it to the available connection bandwidth.
- ▶ Always keep some reserve for unexpected transfer rate peaks.

Important: Even if the distances are in the range of Short Wave Fibre specification, it can be a good practice to use asynchronous mirroring mode if the disk system is heavily loaded with host I/O requests. The performance of the disk system is not improved by using asynchronous mode, but the application performance is almost unaffected while using Remote Mirroring.

8.12 Long-distance ERM

With Global Copy and Global Mirror operating modes, the Enhanced Remote Mirroring provides the capability to mirror data over longer distances. Asynchronous mirroring, combined with the use of devices known as channel extenders allows replication over distances of more than 5000 km (3200 mi).

There are three main characteristics for a connection:

- ▶ **Connection bandwidth:** Defines how much information can be transferred over a connection in a given period of time. It is usually measured in bits per second (bps) or bytes per second (Bps). Fibre Channel bandwidth is 2 Gbps or 1 Gbps.
- ▶ **Connection latency:** Specifies how long it takes get a response in return to a request. Most applications are expecting values between 1–10 ms.
- ▶ **Connection protocol:** Is the set of the communication rules both communication partners have to use to understand each other.

For long distance, you could in theory use a native Fibre Channel, low latency, high bandwidth connections such as Dense Wavelength Division Multiplexing (DWDM) or Coarse Wavelength Division Multiplexing (CWDM) technologies. But this is only a speculative possibility for most situations because of the connection and equipment costs.

It is also possible to build a long distance connection based on Internet topology and TCP/IP protocols. TCP/IP implementations offer good bandwidth (10 Mbps to 1 Gbps) for storage mirroring purposes and can be secured by known and tested technologies like VPN.

The latency is the main issue using IP connections for storage mirroring. Due to the complexity of the Internet topology and many other various factors, the latency of an IP Network does not have a constant value and can vary in a large range.

The second factor is the value itself. Latency values about 50–200 ms are common values for a Internet connection. Although the storage controller I/O completion acknowledgement time-out could accommodate such latency, the host application probably would not. Also, keep in mind that you have not even considered the high I/O rates and possible connection breaks yet.

The solution to circumvent this problem is to use an asynchronous Mirroring Mode like Global Copy or Global Mirror for a connection based on IP Networks.

Remember that the Mirror Repository Logical Drive can queue a number of I/O requests until the maximum allowed difference of pending write requests between primary and secondary is attained (number of unsynchronized write requests). During this process the mirrored pair state remains Synchronized. If the maximum number of unsynchronized I/O requests is exceeded, the state of the mirrored pair changes to Unsynchronized. The time you have before the mirroring state changes from Synchronized to Unsynchronized depends mainly on two factors:

- ▶ Maximum number of unsynchronized write requests
- ▶ I/O write request per second to the primary disk system

For example, if the queue is holding a maximum of 1024 I/Os and you have an impact of 2500 I/O write requests per second, then the time difference between each I/O is:

$$1s / 2500 \text{ IOPS} = 0.4 \text{ ms.}$$

The period of time you can hold the Synchronized state with your queue can be calculated as follows:

$$1024 \text{ writes} * 0.4\text{ms/write} = 409.6\text{ms}$$

You can assume that connection can have a maximum latency of 200 ms (you must send the data and receive the confirmation). The theoretical value that can be assumed for an IP Network is about 100 km/ms. In this case you could theoretically have a maximum connection length of 20000 km. In reality, the value is slightly less because of the unknown number of routing hops, alternate traffic volume, and other factors.

Note: The queue size is 128 outstanding I/Os per logical drive (with firmware version 6.1x.xx.xx). For a subsystem with 32 mirrored pairs, the queue length is thus 128x32=4096; and for 64 pairs, it is 8192. We recommend using as many logical drives as possible for ERM long distance to keep the I/Os on one particular logical drive in the smallest possible range.

Now let us look at the communication protocol. Channel Extenders must be used to route SAN Data Frames (based on FCP) over a TCP/IP Network, and vice versa.

One of the possibilities is the IBM Proven Solution for DS4000 family, using McDATA UltraNet Edge Storage Router Fibre Channel Extenders. Using Global Copy or Global Mirror Enhanced Remote Mirroring Modes, configurations with a distance of more than 5150 km (3200 mi) can be achieved. Lab tests with simulated IP Networks have reached a distance of 10000 km (6200 mi) without synchronization problems.

Figure 8-17 shows a configuration principle for long-distance Remote Mirroring.

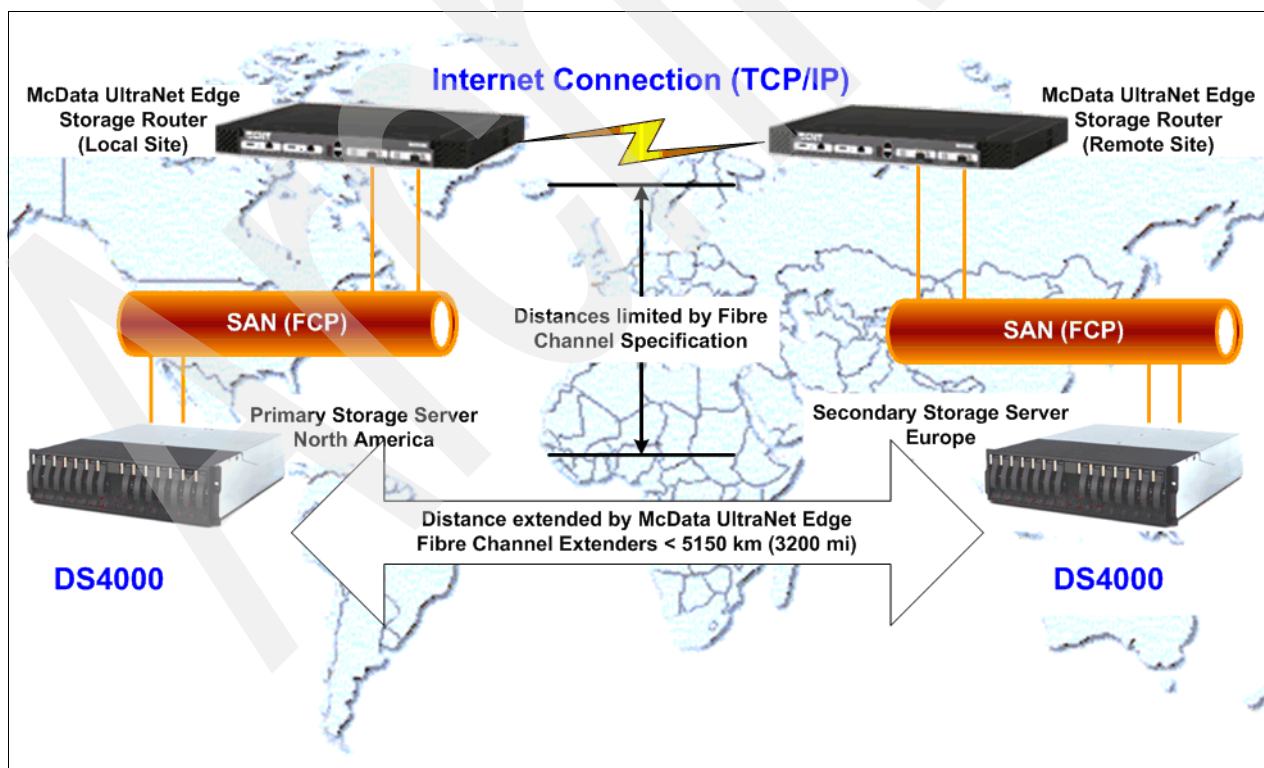


Figure 8-17 Long-distance configuration principle

8.13 Summary

The range of the IBM System Storage DS4000 solutions adequately caters to the mid-range and enterprise storage based IT Business Continuity requirements, across multiple platforms in the open systems world. This demonstrates the IBM commitment to providing workable disaster recovery solutions for all levels of storage requirements from small to very large.

For more information about DS4000, see the IBM Redbook, *IBM System Storage DS4000 Series, Storage Manager and Copy Services*, SG24-7010, and the URL:

<http://www.ibm.com/servers/storage/disk/ds4000/>

IBM System Storage N series

In this chapter we describe the features and functionalities of the IBM System Storage N series. We cover the following topics:

- ▶ N series hardware overview:
 - System Storage N3700
 - System Storage N5000 series
 - System Storage N7000 series
 - N series expansion units
- ▶ N series software:
 - Software overview
 - Detailed introduction of N-series Copy Service related functions

9.1 N series hardware overview

The IBM System Storage N series offers additional choices to organizations facing the challenges of enterprise data management. The IBM System Storage N series is designed to deliver enterprise storage and data management value with midrange affordability. Built-in serviceability and manageability features help increase reliability, simplify and unify storage infrastructure and maintenance, and deliver exceptional economy. Figure 9-1 summarizes the current products available.

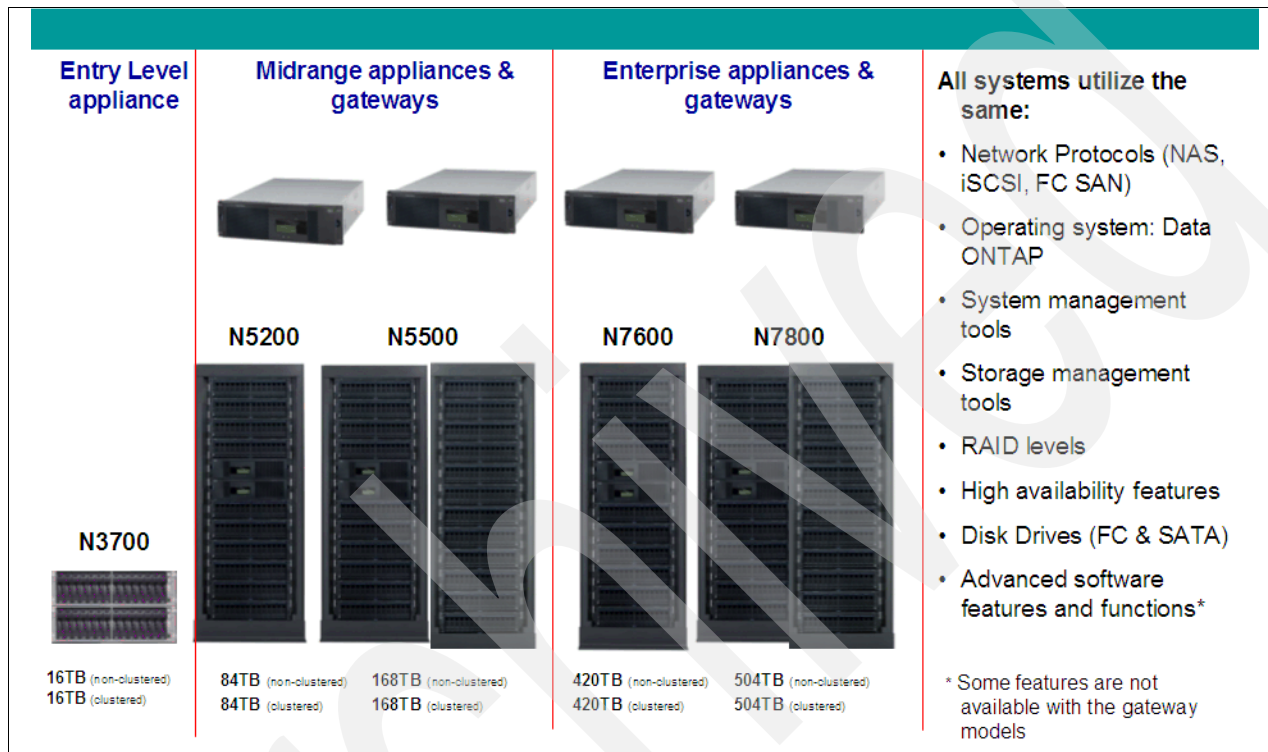


Figure 9-1 N series portfolio

The N series products provide a wide-range of network attachment capabilities to a broad range of host and client systems using multiple network access protocols, including file system NAS protocols (CIFS and NFS); and block I/O protocols, including iSCSI and FCP — all from a single hardware platform, simultaneously.

The N series products are very flexible — they can be populated with both Fibre Channel and SATA disk drives. Fibre Channel disk drives can be suitable for relatively high-performance data transaction environments. SATA disk drives provide an economical platform for disk to disk backup, disaster recovery, data archive, or data like home directories that do not require high-performance environments.

All N series systems utilize a single operating system (Data ONTAP®) across the entire platform and offer a combination of multiple advanced function software features for comprehensive system management, storage management, onboard and outboard copy services, virtualization technologies, and disaster recovery and backup solutions.

Optional WORM data protection software provides additional data security in regulatory environments where data must be stored in non-erasable and non-rewritable formats to meet the industry's newest and strict regulatory requirements for retaining company data assets.

The N series portfolio of products includes tools for managing database environments like Microsoft Exchange, Microsoft SQL. Patented RAID-DP (RAID Double Parity) helps ensure the highest availability and data loss prevention.

For the latest information about platforms supported by IBM System Storage N series, see:

<http://www.ibm.com/servers/storage/nas/interophome.html>

9.1.1 System Storage N3700

The N3700 Filer, shown in Figure 9-2, provides NAS and iSCSI functionality for entry to mid-range environments. The basic N3700 A10 is a single-node 3U model, which is upgradeable to the dual-node model A20, without requiring additional rack space. The dual-node, clustered A20 supports failover and failback functions to maximize reliability. The N3700 filer can support 14 internal hot-plug disk drives, and can attach up to three 3U EXN2000 or EXN1000 expansion units, each with a maximum of 14 drives. The N3700 also can connect to a SAN.



Figure 9-2 System Storage N3700

N3700 hardware features

There are two models available:

- ▶ N3700 A10 (2863-A10, Single Controller)
- ▶ N3700 A20 (2863-A20, Dual Controller)

The IBM System Storage N3700 has the following features:

- ▶ 3U rack-mountable integrated filer and disk storage enclosure
- ▶ Redundant hot plug power supplies and cooling
- ▶ Two integrated full duplex 10/100/1000 Ethernet ports per NAS controller
- ▶ Two integrated FC ports per NAS controller

These disk storage expansion units are supported:

- ▶ EXN2000 - FC Disk Storage Expansion Unit
- ▶ EXN1000 - SATA Disk Storage Expansion Unit

These disk drive capacities are supported:

- ▶ EXN2000 - 10K rpm FC disk drives (72 GB, 144 GB, 300 GB), 15K RPM FC disk drives (72 GB, 144 GB)
- ▶ EXN1000 - 7200 rpm SATA disk drives (250 GB, 320 GB, 500 GB)

N3700 specifications

Table 9-1 summarizes the specifications of the N3700.

Table 9-1 3700 specifications

Filer specifications	N3700 A10	N3700 A20
IBM machine type/model	2863-A10	2863-A20
Max. raw capacity	16.8TB	16.8TB
Max. number of disk drives	56	56
Max. volume size	8TB	8TB
ECC memory	1 GB	2 GB
Nonvolatile memory	128 MB	256 MB
CPU	Broadcom 650-MHz BCM1250 *2	Broadcom 650-MHz BCM1250 *4
Ethernet 10/100/1000 copper	2	4
Copper FC adapter	1	2
Optical FC adapter (host-attach SAN/Tape SAN)	1	2
Clustered failover-capable	No	Yes
Supported Expansions	EXN2000 for FC Disks EXN1000 for SATA Disks	
Supported Disks	10K rpm FC disk drives: 72/144/300 GB 15K rpm FC disk drives: 72/144 GB 7.2K rpm SATA disk drives: 250/320/500 GB	
Rack Mount (in IBM 2101 Storage Solutions Rack Model 200 or other industry-standard 19-inch rack)	Yes	

Notes:

- ▶ Storage for an N3700 NAS system is limited to a maximum of 56 disk drives in both the storage controller and all expansion units, and a maximum of 16.8 TB disk raw capacity.
- ▶ The base N3700 chassis does not support SATA disk drives.
- ▶ Only one type of disk expansion unit can be attached to any particular N3700. EXN2000 and EXN1000 cannot coexist on the same N3700.
- ▶ A maximum of two distinct drive types can be attached to any particular N3700.

9.1.2 System Storage N5000 series

The IBM System Storage N5000 series, shown in Figure 9-3, comes in both appliance and gateway models. All models come without storage in the base chassis. The appliance models can attach Fibre Channel EXN2000 and SATA EXN1000 disk expansion units. The gateway models support the attachment of external storage from IBM or other vendors.

There are no visible differences between the N5200 and N5500. The differences are in maximum storage capacity, Storage Cache and CPU processing power.



Figure 9-3 IBM System Storage N5000 series

N5000 hardware features

There are four Appliance models and four Gateway models available:

- ▶ N5200 A10 (2864-A10, Single Appliance)
- ▶ N5200 A20 (2864-A20, Clustered Appliance)
- ▶ N5500 A10 (2865-A10, Single Appliance)
- ▶ N5500 A20 (2865-A20, Clustered Appliance)
- ▶ N5200 G10 (2864-G10, Single Gateway)
- ▶ N5200 G20 (2864-G20, Clustered Gateway)
- ▶ N5500 G10 (2865-G10, Single Gateway)
- ▶ N5500 G20 (2865-G20, Clustered Gateway)

Each IBM System Storage N5000 has the following standard features:

- ▶ 19" rack-mount enclosure
- ▶ Dual redundant hot-plug integrated cooling fans and auto-ranging power supplies
- ▶ Four full-duplex 10/100/1000 Base-T Ethernet ports onboard per NAS controller
- ▶ Four 2 Gbps Fibre Channel ports onboard per NAS controller

These disk storage expansion units are supported:

- ▶ EXN2000 - FC Disk Storage Expansion Unit
- ▶ EXN1000 - SATA Disk Storage Expansion Unit

These disk drive capacities are supported:

- ▶ EXN2000 - 10K rpm FC disk drives (72 GB, 144 GB, 300 GB), 15K RPM FC disk drives (72 GB, 144 GB)
- ▶ EXN1000 - 7200 rpm SATA disk drives (250 GB, 320 GB, 500 GB)

N5000 model specifications

Table 9-2 summarizes the specifications of the N5000 models, Table 9-3 shows the characteristics of various options on the appliance models, and Table 9-4 shows the characteristics of various options on the gateway models.

Table 9-2 N5000 specifications

File specifications	N5200	N5200	N5500	N5500
IBM machine types - models	2864-A10 2864-G10	2864-A20 2864-G20	2865-A10 2865-G10	2865-A20 2865-G20
Storage configuration	Single storage controller	Dual (active/active) storage controllers	Single storage controller	Dual (active/active) storage controllers
ECC memory	2 GB	4 GB	4 GB	8 GB
Nonvolatile memory	512 MB	1 GB	512 MB	1 GB
CPU	Intel 2.8GHz Xeon® *1	Intel 2.8GHz Xeon *2	Intel 2.8GHz Xeon *2	Intel 2.8GHz Xeon *4
Onboard 10/100/1000 Ethernet ports	4	8	4	8
Onboard 2 Gbps Fibre Channel ports (configurable as storage-attached initiator or host-attached target)	4	8	4	8
PCI-X expansion slots	3	6	3	6
Rack Mount (in IBM 2101 Storage Solutions Rack Model 200 or other industry-standard 19-inch rack)	Yes			

Table 9-3 N5000 appliance specifications

Appliance disk specifications	N5200	N5200	N5500	N5500
IBM machine types - models	2864-A10	2864-A20	2865-A10	2865-A20
Max. number of Fibre Channel loops	4	8	4	8
Max. number of drives per Fibre Channel loop	84	84	84	84
Max. number of disk drives	168	168	336	336
Max. raw storage capacity	84TB	84TB	168TB	168TB
Max. storage expansion units	12	12	24	24
Max. volume size	17.6TB	17.6TB	17.6TB	17.6TB

Appliance disk specifications	N5200	N5200	N5500	N5500
Supported Expansions	EXN2000 for FC Disks EXN1000 for SATA Disks			
Supported Disks	10K rpm FC disk drives: 72/144/300 GB 15K rpm FC disk drives: 72/144 GB 7.2K rpm SATA disk drives: 250/320/500 GB			

Table 9-4 N5000 gateway specifications

Appliance disk specifications	N5200	N5200	N5500	N5500
IBM machine types - models	2864-G10	2864-G20	2865-G10	2865-G20
Max. raw storage capacity	50TB	50TB	84TB	84TB
Max. number of LUNs on back-end disk storage array	168	168	336	336
Maximum LUN size on back-end disk storage array	500 GB	500 GB	500 GB	500 GB
Max. volume size:	16TB	16TB	16TB	16TB

9.1.3 System Storage N7000 series

The IBM System Storage N7000 series, shown in Figure 9-4, comes in both appliance and gateway models. All models come without storage in the base chassis. The appliance models can attach Fibre Channel EXN2000 and SATA EXN1000 disk expansion units. The gateway models support the attachment of external storage from IBM, or other vendors.

There are no visible differences between the N7600 and N7800. The differences are in maximum storage capacity, Storage Cache and CPU processing power.

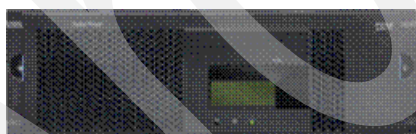


Figure 9-4 IBM System Storage N7000 series

N7000 hardware features

There are four Appliance models and four Gateway models available:

- ▶ N7600 A10 (2866-A10, Single Appliance)
- ▶ N7600 A20 (2866-A20, Clustered Appliance)
- ▶ N7800 A10 (2867-A10, Single Appliance)
- ▶ N7800 A20 (2867-A20, Clustered Appliance)
- ▶ N7600 G10 (2866-G10, Single Gateway)
- ▶ N7600 G20 (2866-G20, Clustered Gateway)
- ▶ N7800 G10 (2867-G10, Single Gateway)
- ▶ N7800 G20 (2867-G20, Clustered Gateway)

Each IBM System Storage N7000 has the following standard features:

- ▶ A 19" rack-mount enclosure
- ▶ Dual redundant hot-plug integrated cooling fans and auto-ranging power supplies
- ▶ Six full-duplex 10/100/1000 Base-T Ethernet ports onboard per NAS controller
- ▶ Eight 2 Gbps Fibre Channel ports onboard per NAS controller

These disk storage expansion units are supported:

- ▶ EXN2000 - FC Disk Storage Expansion Unit
- ▶ EXN1000 - SATA Disk Storage Expansion Unit

These disk drive capacities are supported:

- ▶ EXN2000 - 10K rpm FC disk drives (72 GB, 144 GB, 300 GB), 15K RPM FC disk drives (72 GB, 144 GB)
- ▶ EXN1000 - 7200 rpm SATA disk drives (250 GB, 500 GB)

N7000 model specifications

Table 9-5 summarizes the specifications of the N7000 models, Table 9-5 shows the characteristics of various options on the appliance models, and Table 9-7 shows the characteristics of various options on the gateway models.

Table 9-5 N7000 specifications

Filer specifications	N7600	N7600	N7800	N7800
IBM machine types - models	2866-A10 2866-G10	2866-A20 2866-G20	2867-A10 2867-G10	2867-A20 2867-G20
Storage configuration	Single storage controller	Dual (active/active) storage controllers	Single storage controller	Dual (active/active) storage controllers
ECC memory	16 GB	32 GB	32 GB	64 GB
Nonvolatile memory	512 MB	1 GB	2 GB	4 GB
CPU	AMD 2.6GHz Opteron*2	AMD 2.6GHz Opteron*4	AMD 2.6GHz Opteron*4	AMD 2.6GHz Opteron *8
Onboard Gigabit Ethernet ports	6	12	6	12
Onboard 2 Gbps Fibre Channel ports (configurable as storage-attached initiator or host-attached target)	8	16	8	16
PCI-X expansion slots	3	6	3	6
PCI-E expansion slots	5	10	5	10
Rack Mount (in IBM 2101 Storage Solutions Rack Model 200 or other industry-standard 19-inch rack)	Yes			

Table 9-6 N7000 appliance specifications

Appliance disk specifications	N7600	N7600	N7800	N7800
IBM machine types - models	2866-A10	2866-A20	2867-A10	2867-A20
Max. number of Fibre Channel loops	8	10	8	12
Max. number of drives per Fibre Channel loop	84	84	84	84
Max. number of disk drives	672	840	672	1008
Max. raw storage capacity	336TB	420TB	336TB	504TB
Max. storage expansion units	48	60	48	72
Max. volume size	16TB	16TB	16TB	16TB
Supported Expansions	EXN2000 for FC Disks EXN1000 for SATA Disks			
Supported Disks	10K rpm FC disk drives: 72/144/300 GB 15K rpm FC disk drives: 72/144 GB 7.2K rpm SATA disk drives: 250/500 GB			

Notes:

- For the initial order of the IBM System Storage N7600/7800 appliance, you cannot include EXNx000 storage expansion units containing more than two types (rotational speed and capacity) of disk drives.
- IBM System Storage N7600/7800 appliance does not support 320 GB SATA disks, while N3700 and N5000 series do.

Table 9-7 shows the specifications of the N7000 gateway.

Table 9-7 N7000 gateway specifications

Appliance disk specifications	N7600	N7600	N7800	N7800
IBM machine types - models	2866-G10	2866-G20	2867-G10	2867-G20
Max. raw storage capacity	420TB	420TB	504TB	504TB
Max. number of LUNs on back-end disk storage array	840	840	1008	1008
Maximum LUN size on back-end disk storage array	500 GB	500 GB	500 GB	500 GB
Max. volume size:	16TB	16TB	16TB	16TB

9.2 N series expansion units

An IBM System Storage N5000/N7000 series appliance system requires at least one storage expansion unit per node, either an EXN1000 or an EXN2000. These expansion units are also available as options on the N3700.

9.2.1 EXN1000 expansion unit

The EXN1000 storage expansion unit, shown Figure 9-5, provides a 3U, rack-mountable, disk enclosure containing from five to 14 SATA disk drives of either 250 GB, 320 GB, or 500 GB.



Figure 9-5 EXN1000 expansion unit

9.2.2 EXN2000 expansion unit

The EXN2000 storage expansion unit, shown in Figure 9-6, provides a 3U rack-mountable disk enclosure containing from four to 14 FC disk drives.

The EXN2000 supports the following FC disk drive speeds and capacities:

- ▶ 15,000 revolutions per minute (15K rpm) in 72 GB and 144 GB capacities
- ▶ 10,000 revolutions per minute (10K rpm) in 72 GB, 144 GB, and 300 GB capacities

9.3 N series software overview

The N series software, called Data ONTAP utilizes built-in RAID technologies to provide data reliability. Additional options are available for mirroring, replication, snapshots and backup. The software provides simple management interfaces straightforward installation, administration, and troubleshooting. It can be used in UNIX, Windows, and Web environments.

9.3.1 The IBM N series standard software features

Table 9-8 shows the standard software included with each N series.

Table 9-8 Standard software included with each N series

Data ONTAP	A highly scalable and flexible operating system which delivers flexible management and supports high availability and business continuance.
iSCSI	Support for iSCSI protocol.
AutoSupport	Provides continuous monitoring of the storage system's health using a sophisticated, event-driven logging agent.

SecureAdmin™	Enables authenticated, command-based administrative sessions between an administrative user and Data ONTAP. SecureAdmin supports authentication of both the administrative user and the filer, creating a secured, direct communication link to the filer.
FilerView®	A Web-based administration tool that allows IT administrators to fully manage N series from remote locations.
FlexVol	Creates multiple flexible volume on a large pool of disks. Dynamic, nondisruptive storage provisioning; space- and time-efficiency. These flexible volumes can span multiple physical volumes, regardless of size.
FlexShare	A control-of-service tool for simplifying storage management. Storage administrators can host and prioritize different applications on a single storage system while avoiding impact to critical applications.
SnapShot	Makes locally retained, point-in-time copies of data.
SnapMover®	Tool to migrate data among N series while avoiding impact on data availability and disruption to users.
Disk Sanitization	Physically obliterates data by overwriting disks with specified byte patterns or random data in a manner that prevents recovery of current data by any known recovery methods. Uses three successive byte overwrite patterns per cycle.
Integrated RAID manager	IBM N series and Data ONTAP provide integrated RAID management with RAID-Double Parity (default) and RAID 4.

9.3.2 N series optional software features

Table 9-9 shows the optional, chargeable software available for N series.

Table 9-9 Optional, chargeable software available for N series

CIFS	Support for Common Internet File System Protocol (CIFS)
NFS	Support for Network File System Protocol (NFS)
FCP	Support for Fibre Channel Protocol (FCP)
HTTP	Support for Hypertext Transfer Protocol (HTTP)
FTP	Support for File Transfer Protocol (FTP)
Cluster Failover	Installed on a pair of N series, Clustered Failover can help improve data availability by transferring the data service of one system to another system in the cluster. Often, the transfer of data service occurs without impacting users and applications, and the data service can be quickly resumed with no detectable interruption to business operation.
FlexClone	Generates nearly instantaneous replicas of data sets and storage volumes that require no additional storage space. Each cloned volume is a transparent virtual copy that can be used for enterprise operations.
Multistore	Quickly and easily creates separate, private logical partition. Each virtual storage partition can maintain separation from every other storage partition to prevent different enterprise departments that share the same storage resources from accessing or finding other partitions. MultiStore® helps prevent information about any virtual partition from being viewed, used or downloaded by unauthorized users.

SnapMirror	Can help deliver the disaster recovery and data distribution solutions. Replicates data at high speeds over a LAN or a WAN, to support high data availability and quick recovery for applications. SnapMirror can mirror data to one or more N series and constantly update the mirrored data to keep it current and available.
SyncMirror	Keeps data available and up-to-date by maintaining two copies of data online. When used with Clustered Failover, SyncMirror can help support even higher levels of data availability.
SnapRestore®	Helps recover data quickly in the case of a disaster. Can recover data in amounts as small as an individual file up to a multi-terabyte volume so that operations can quickly resume.
SnapVault®	Supports the frequent backup of data on N series. Provides a centralized, disk-based backup solution. Storing backup data in multiple Snapshot copies on a SnapVault secondary storage system allows for keeping multiple backups online over a period of time for faster restoration.
SnapLock	Delivers high performance and high security data function to disk-based nearline and primary N series storage. Helps manage the permanence, accuracy, integrity, and security of data by storing business records in an inalterable form and allowing for their rapid online accessibility for long periods of time
LockVault	Helps manage unstructured data used for regulatory compliance. LockVault is also integrated with backup and disaster recovery functionality to help support comprehensive, integrated protection of unstructured data.
SnapDrive®	Supports flexible and efficient utilization of storage resources, with a rich set of capabilities designed to virtualize and enhance storage management for Microsoft Windows and UNIX environments.
SnapManager	Simplifies configuration, backup, and restore operations for Exchange/SQL databases.
Single Mailbox Recovery with SnapManager for Microsoft Exchange	Fast, accurate, cost-effective backup and recovery of Microsoft Exchange data. Can take near-instantaneous online backups of Exchange databases, verify that the backups are consistent, and rapidly recover Exchange at almost any level of granularity-storage group, database, folder, single mailbox, or single message.
SnapValidator®	Provides protection for Oracle data.
DataFabric® Manager	Offers comprehensive monitoring and management for N series enterprise storage and content delivery environments. Provides centralized alerts, reports, and configuration tools.
MetroCluster	High-availability and business continuance solution - extends failover capability from within a data center to a site located many miles away. It also helps support replication of data from a primary to a remote site to maintain data currency.
Near-line feature	Optimizes N System for data protection and retention applications. It enables additional concurrent streams and SnapVault for NetBackup.

9.3.3 Details of advanced copy service functions

Data ONTAP and N series provide a number of advanced copy service features in software.

Advanced data protection with SnapShot

A SnapShot is a locally retained point-in-time image of data. SnapShot technology is a feature of the WAFL® (Write Anywhere File Layout) storage virtualization technology that is a part of Data ONTAP. A SnapShot is a “frozen,” read-only view of a WAFL volume that provides easy access to old versions of files, directory hierarchies, and LUNs.

A SnapShot (Figure 9-6) can take only a few seconds to create, regardless of the size of the volume or the level of activity on the N series. After a SnapShot copy has been created, changes to data objects are reflected in updates to the current version of the objects, as though SnapShot copies did not exist. Meanwhile, the SnapShot version of the data remains completely unchanged. A SnapShot copy incurs little performance overhead; depending on available space, users can store up to 255 SnapShot copies per WAFL volume, all of which are accessible as read-only, online versions of the data.

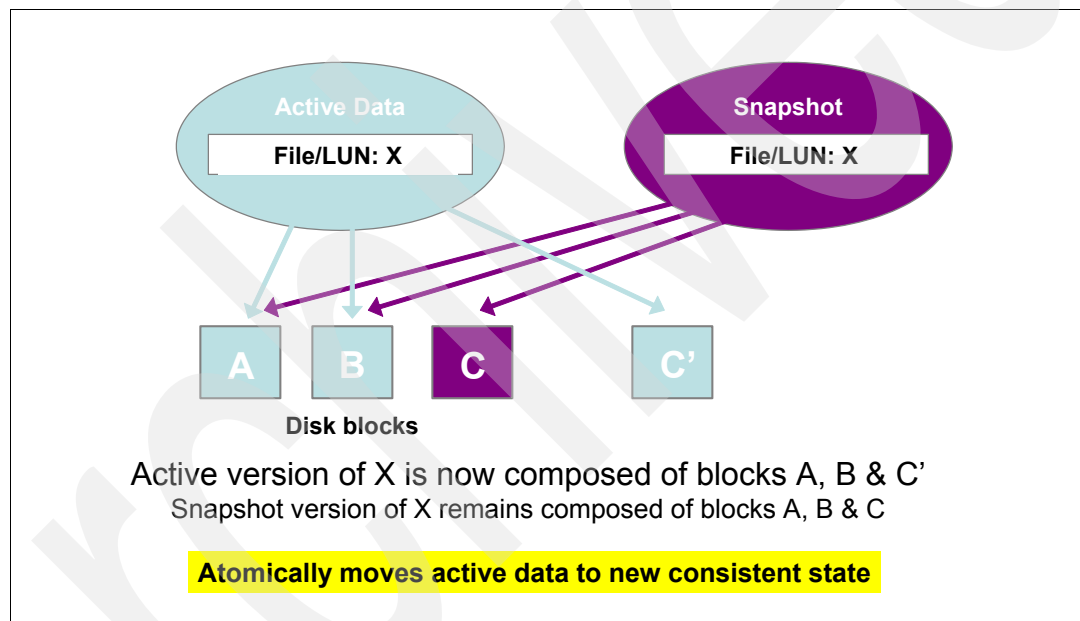


Figure 9-6 N series SnapShot - How it works

A SnapShot copy can be used to provide frequent, low-impact, user-recoverable backups of files, directory hierarchies, LUNs, and application data. A SnapShot copy can significantly improve the frequency and reliability of backups, since it is designed to avoid performance overhead and can be created on a running system.

A SnapShot supports near-instantaneous, user-managed restores. Users can directly access SnapShot copies to recover from accidental deletions, corruptions, or modifications of their data.

SnapRestore

SnapRestore capability helps recover data quickly when disaster strikes. SnapRestore technology can quickly recover large individual files or volumes through instant volume recovery. Volumes can be restored with a single command versus the file level restores that SnapShot offers.

SnapShot technology uses storage efficiently, as it stores only block-level changes between each successive SnapShot copy. Since the SnapShot process is automatic and incremental, backups are significantly faster and simpler. SnapRestore technology uses SnapShot copies to perform near-instantaneous data restoration. In contrast, non-point-in-time storage solutions can copy all of the data and require much more time and disk storage for the backup and restore operations.

SnapShot and SnapRestore working together

At time State 0 (Figure 9-7), the first SnapShot is taken. Its points to the 4K blocks are equivalent to those in the Active File System. No additional space is used at this time by SnapShot 1 because modifications to the Active File System blocks have not occurred.

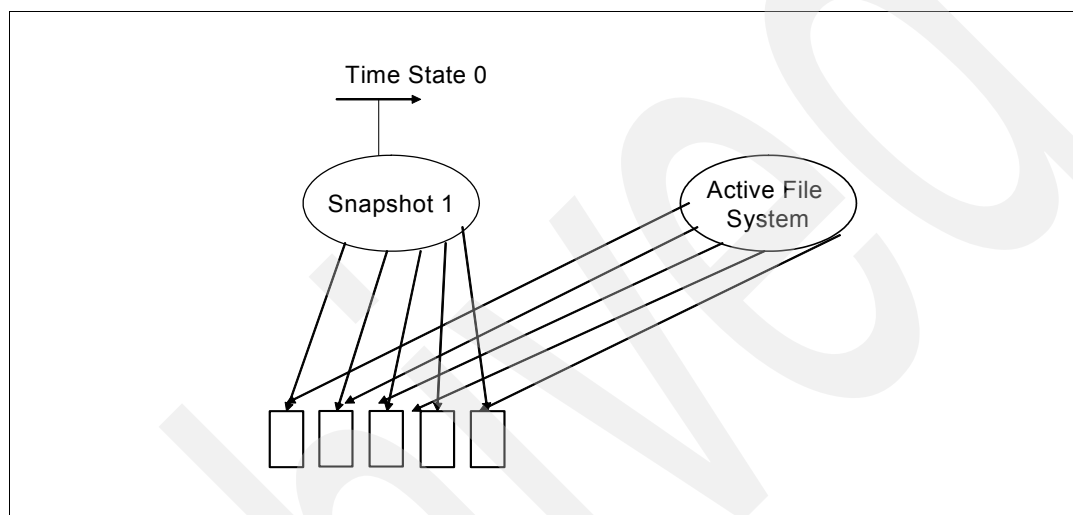


Figure 9-7 First Snap

Over time, new files are added with new blocks and modifications to files and their existing blocks are done, as shown in Figure 9-8. SnapShot 1 now points to blocks and the file system as it appeared in Time State 0; notice that one of the blocks A1 has not been modified and is still part of the Active File System. SnapShot 2 reflects a SnapShot of file modifications and adds Since Time State 0. Notice that it still points to Active File System blocks A1 and A2.

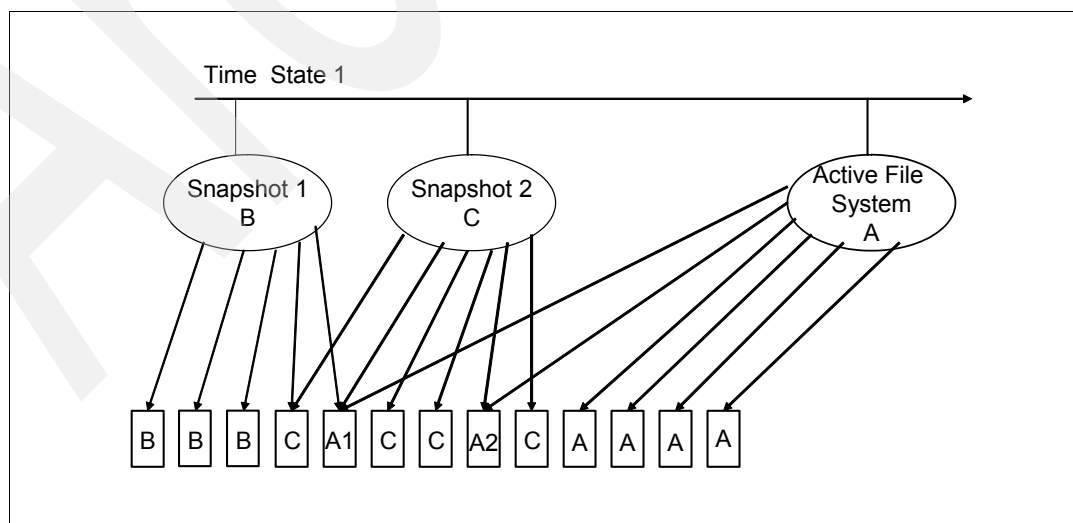


Figure 9-8 Second Snap

More files are added with new blocks and modifications to files and their existing blocks are done, as in Figure 9-9. SnapShot 1 now points to blocks and the file system as it appeared in Time State 0. SnapShot 2 reflects a SnapShot of file modifications and adds Since Time State 0. SnapShot 3 reflects modifications and adds since Time State 1 and SnapShot 2. Notice that SnapShot 1 no longer points to any Active File System Blocks.

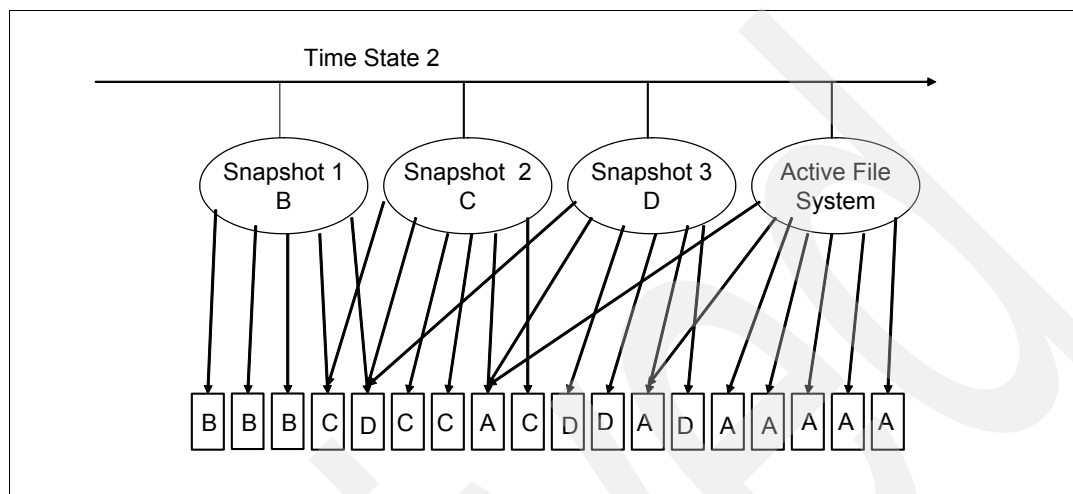


Figure 9-9 SnapShot 3

Jumping ahead to subsequent SnapShot 4 (Figure 9-10), this puts us at Time State 3, and reflects additions or modifications of 4K blocks. Notice that the first two SnapShots no longer reflect any of the Active File System Blocks.

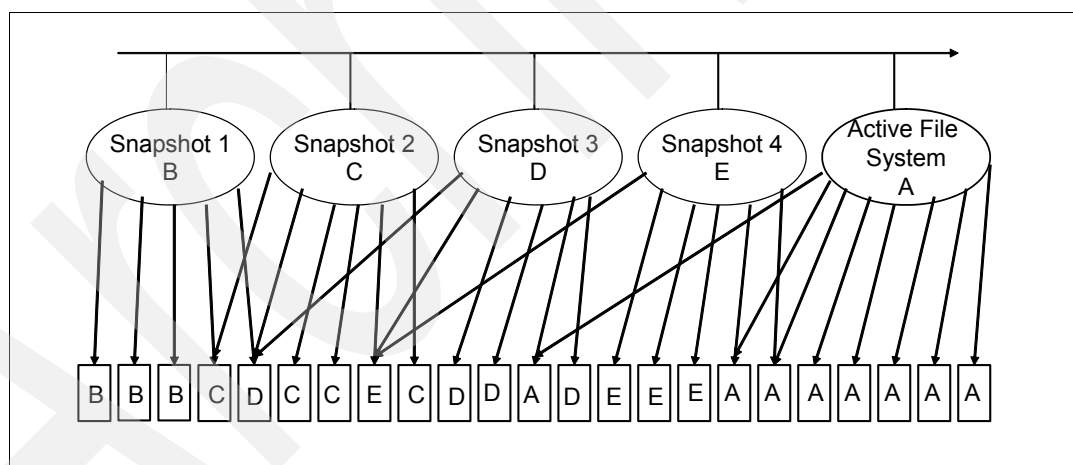


Figure 9-10 Subsequent SnapShots

In Figure 9-11, we discover some file system corruption due to a virus, and must revert to the Point in Time Snapshot as it looked in Time State 2 and SnapShot 3. The Active File system becomes SnapShot 3. Blocks that were previously pointed to solely by SnapShot 4 or the Active File System are freed up for writes again. In addition, any blocks that were pointed to only by SnapShot 4 and the previous Active File system are also freed up again.

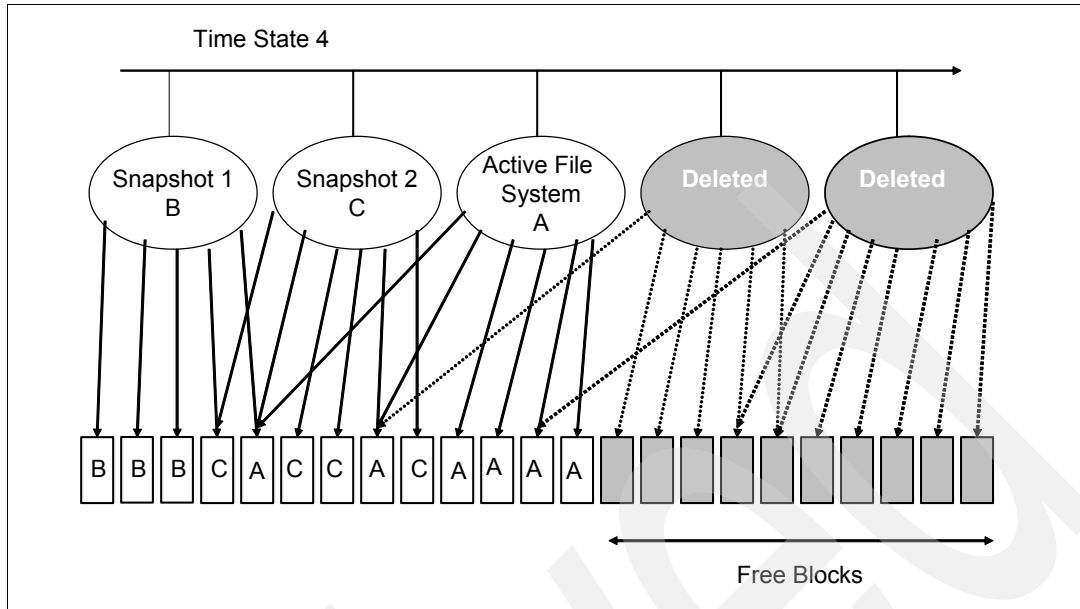


Figure 9-11 Reversion to Snap 3

In Figure 9-12 we compare Time State 4 and the Reversion to SnapShot 3 and Time State 1, which reflects the Active File System before SnapShot 3. As you can see, they are the same.

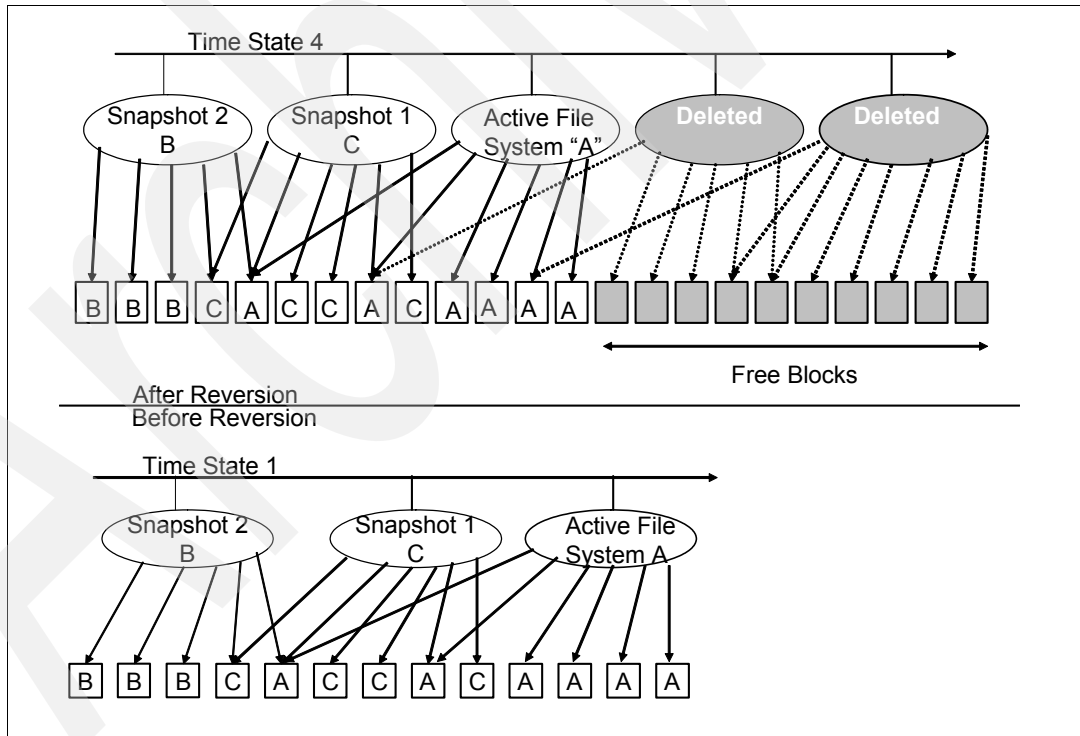


Figure 9-12 Comparison of Reversion to Snap 3 Time State 4 and Active File System at Time State 1

Automated data backup with SnapVault

SnapVault, shown in Figure 9-13, is a low overhead, disk-based online backup of heterogeneous storage systems for fast and simple restores. SnapVault is a separately licensed feature in Data ONTAP that provides disk-based data protection for filers. SnapVault

replicates selected Snapshots from multiple client filers to a common snapshot on the SnapVault server, which can store many Snapshots. These Snapshots on the server have the same function as regular tape backups. Periodically, data from the SnapVault server can be dumped to tape for extra security.

SnapVault which is a heterogeneous disk-to-disk data protection solution ideal for use with the N series. SnapVault is a low overhead, disk-based online backup of heterogeneous storage systems for fast and simple restores. A SnapVault primary system corresponds to a backup client in the traditional backup architecture. The SnapVault secondary is always an N series running Data ONTAP. SnapVault software protects data residing on a SnapVault primary.

All of this heterogeneous data is protected by maintaining online backup copies (Snapshot) on a SnapVault secondary system. The replicated data on the secondary system can be accessed via NFS or CIFS just as regular data is. The primary systems can restore entire directories or single files directly from the secondary system. There is no corresponding equivalent to the SnapVault secondary in the traditional tape-based backup architecture.

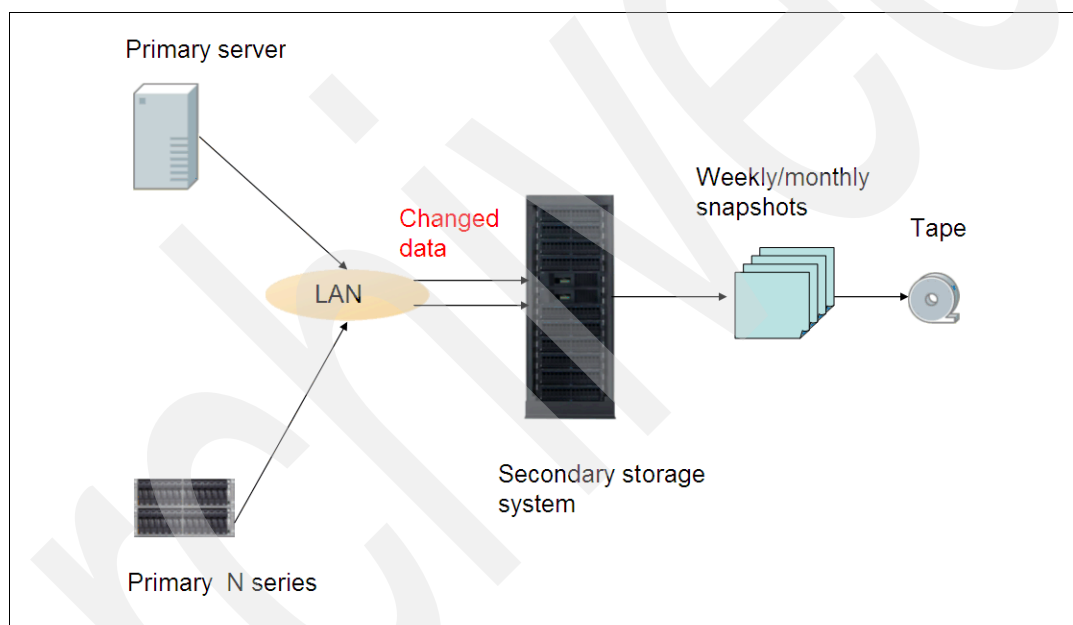


Figure 9-13 Basic SnapVault

Benefits

SnapVault offers the following benefits:

- ▶ Avoids the bandwidth limitations of tape drives so restore can be faster.
- ▶ Does not require full dumps from the primary storage, so no requirement for a backup
- ▶ Protects heterogeneous storage environments
- ▶ Performs disk-to-disk backup and recovery
- ▶ Incrementals forever model reduces backup overhead; incrementals only, changed blocks only
- ▶ Intelligent data movement reduces network traffic and impact to production systems
- ▶ Uses Snapshot technology to significantly reduce the amount of backup media
- ▶ Instant single file restore: snapshot directory displays SnapVault snapshots
- ▶ Can protect remote sites over a WAN

How does SnapVault work?

This is shown in Figure 9-14. The steps are as follows:

1. Administrators set up the backup relationship, backup schedule and retention policy.
2. The backup job starts, based on a backup schedule. One backup job back up multiple SnapVault primaries.
3. It moves data from SnapVault primary to SnapVault secondary.
4. On successful completion of a backup job, it takes SnapShot on SnapVault secondary; it only saves changed blocks on SnapVault secondary.
5. It maintains SnapShots on SnapVault secondary based on retention policy.
6. SnapMirror SnapVault secondary goes to remote location for disaster recovery (optional).
7. Traditional backup application can be leveraged to back up SnapVault secondary to tape.

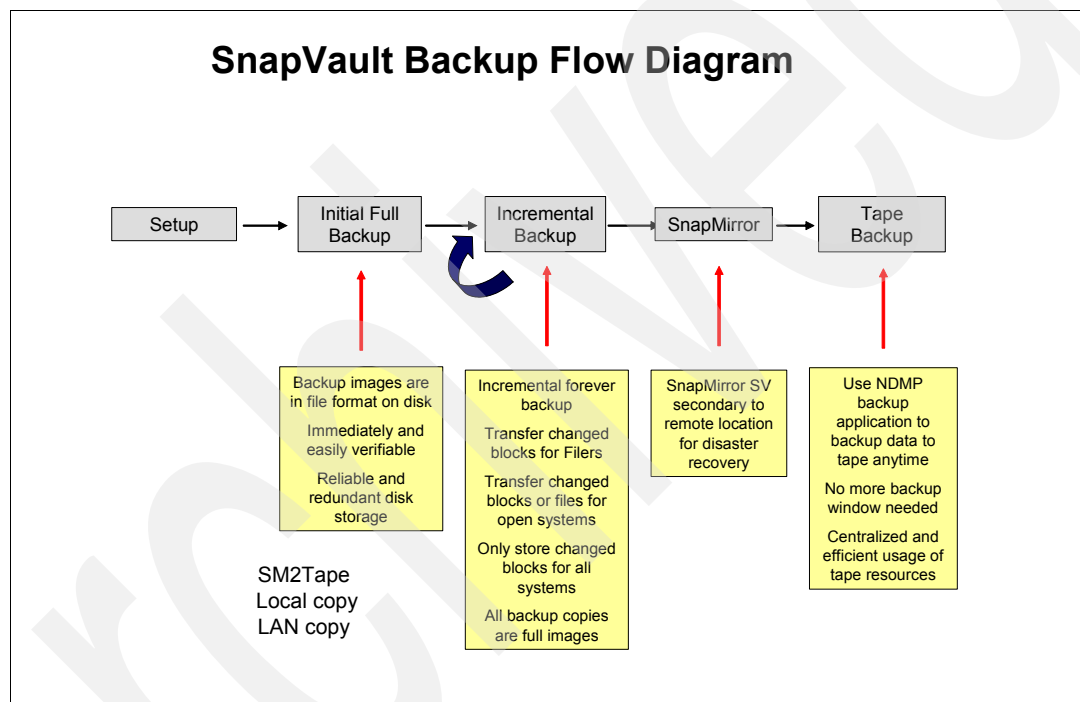


Figure 9-14 Backup flow diagram

After the simple installation of the SnapVault agent on the desired primary file and application servers, the SnapVault secondary system requests initial baseline image transfers from the primary storage system. This initial, or baseline, transfer can take some time to complete, as it is duplicating the entire source data set on the secondary much like a level-zero backup to tape. SnapVault protects data on a SnapVault primary system by maintaining a number of read-only versions of that data on a SnapVault secondary system.

The baseline transfer is typically the most time-consuming process of the SnapVault implementation as it is duplicating the entire source data set on the secondary, much like a full backup to tape. With SnapVault, one baseline transfer is required before any subsequent backups or restores can be performed, but unlike traditional tape backup environments, this initial baseline transfer is a “one time” occurrence, not a weekly event. First, a complete copy of the data set is pulled across the network to the SnapVault secondary. Each subsequent backup transfers only the data blocks that have changed since the previous backup (incremental backup or incrementals forever).

When the initial full backup is performed, the SnapVault secondary stores the data in a WAFL file system, and creates a Snapshot image of that data. Each of these snapshots can be thought of as full backups (although they are only consuming a fraction of the space). A new Snapshot is created each time a backup is performed, and a large number of Snapshots can be maintained according to a schedule configured by the backup administrator. Each Snapshot consumes an amount of disk space equal to the differences between it and the previous Snapshot.

A very common scenario is data protection of the secondary system. A SnapVault secondary system can be protected by either backup to tape or backup to another disk based system. The method to back up to a tertiary disk based system is simply volume based SnapMirror. All snapshots are transferred to the tertiary system and the SnapVault primaries can be directed to this tertiary system if necessary. In order to back up a secondary to a tape library, SnapMirror to tape can be used or simply NDMP backup to tape.

Recovering a file from a SnapVault backup is simple. Just as the original file was accessed via an NFS mount or CIFS share, the SnapVault secondary can be configured with NFS exports and CIFS shares. As long as the destination directories are accessible to the users, restoring data from the SnapVault secondary is as simple as copying from a local Snapshot image. Restores can be drag-and-drop or a simple copy command, depending on the environment. If SnapVault has been deployed in an open systems environment, the restore process can be initiated directly from the primary system that was backed up via the command line.

Recovery of an entire data set can be performed the same way if the user has appropriate access rights. SnapVault provides a simple interface to recover an entire data set from a selected Snapshot copy using the SnapVault restore command. A user can recover the complete contents of a secondary qtree/directory back to the primary with the SnapVault restore command on the primary. The primary data set is read-only until the transfer completes, at which time it becomes writable. When used alone, SnapVault creates hourly read-only Snapshots on Secondary. Hence restores are done via a copy back. Since each Snapshot refers to a complete point-in-time image of the entire file system, this means restore time is ZERO, no tape, no incremental “unwind.”

In comparison, recovery from tape can consume considerable resources. Single files can sometimes be recovered by users, but are typically recovered by an administrator. The tape that contains the backup file must be located (sometimes retrieved from an off-site storage location) and the backup application has to transfer the file from the tape location to the requesting host.

The backup administrator starts the tape restore process and retrieves the tape from the appropriate location if necessary. The tape must be loaded (from seven seconds up to two minutes), positioned to the correct location on the tape (usually several seconds, sometimes more) and finally the data is read. If a full image must be restored, data has to be recovered using the last full and subsequent incremental backups. If a restore requires recovery from one full backup and all incremental backups since that last full backup, then there is more of a chance that an error might occur with the media involved in a tape solution.

This process can be long and tedious depending on the amount of data being recovered. Hours and days can be spent during the restore process. If there is a failure during the restore, the entire process has to be reinitiated, thereby significantly adding to downtime. If the data to be restored is a large critical database, users are offline during the entirety of the restore process.

Rapid backups and restores with SnapManager

SnapManager supports rapid backup and restore of application environments.

SnapManager for Microsoft Exchange

SnapManager software can provide near-instantaneous hot backups and rapid restores for Exchange environments. It can schedule and automate Exchange database backups, use policy-based backup retention management, and simplify the migration of existing databases to N series filers. SnapManager software also offers high availability, with features that allow expansion of Exchange databases online. It supports tight integration with Microsoft Cluster Server (MSCS) and Multi Path I/O (MPIO). It also integrates with the N series Clustered Failover option and SnapMirror software to help simplify disaster recovery implementation.

SnapManager utilizes the Microsoft VSS Infrastructure, and in the Windows 2000 environment, it integrates with the Exchange Backup APIs and ESEutil to provide consistent online backups.

Depending on the restore requirements, there is a wide range of restore options: available: full Exchange Server content recovery, individual Exchange storage group recovery, individual Exchange database recovery, and virtual disk recovery. Even individual mailbox recovery can be performed with the Single Mailbox Recovery software.

Single mailbox recovery for Exchange

One of the most time-intensive (and frequent) tasks for Microsoft Exchange administrators is recovering single mailboxes or single messages. To recover single mail items from Exchange quickly, administrators must perform complex, time-consuming backups that involve backing up each mailbox separately (referred to as bricklevel backups). The alternative is a painful process of setting up a recovery server, loading the last full backup from tape, and then recovering a single mailbox.

The combination of IN series, SnapManager for Exchange, and Single Mailbox Recovery functionality supports the fast, accurate backup and recovery of Microsoft Exchange data. By directly reading the contents of SnapShot copies without the assistance of the Exchange server, N series storage with Single Mailbox Recovery functionality can restore individual mail items from a recent (hourly, daily, weekly) SnapShot. It can restore individual mailboxes, folders, messages, attachments, calendar notes, contacts, and task items directly to the production Exchange server or to a new or existing offline Outlook® PST file.

SnapManager for Microsoft SQL Server

SnapManager supports rapid SQL Server backup times, from hours to as little as seconds, and makes each backup a complete and consistent copy of the original. Backups are based on SnapShot copies, which require minimal disk space for each additional full backup. It allows back up or restore of several databases simultaneously, as well as volume expansion.

SnapManager software can integrate with the SQL Server application to automate the process of validating the consistency of data backup and checking that the data is available for restore.

The integration with SnapMirror technology supports performing remote replication of SQL Server data, thus helping to speed data recovery in the event of a disaster.

Local and remote mirroring solutions

This section discusses mirroring capabilities of N series.

SyncMirror

The SyncMirror functionality, shown in Figure 9-15, provides synchronous local mirroring from one volume to another volume attached to the same filer. It maintains a strict physical separation between the two copies of the mirrored data. In case of an error in one copy, the data is still accessible without any manual intervention.

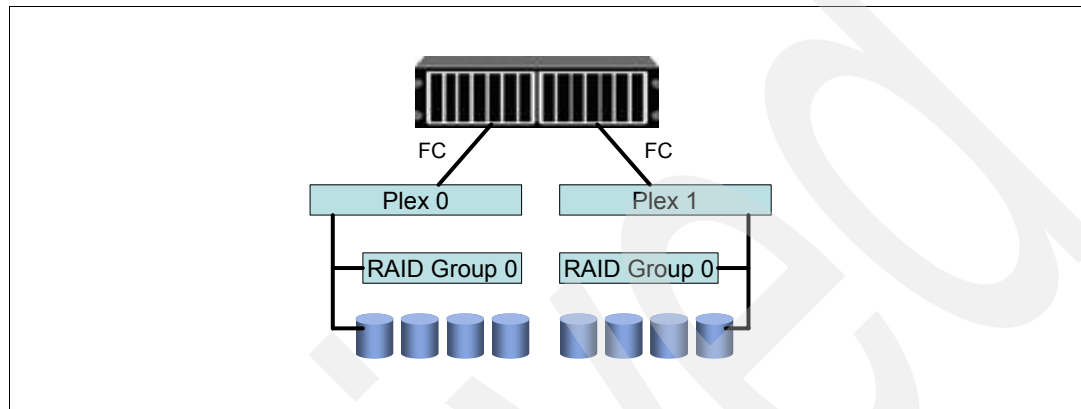


Figure 9-15 N series SyncMirror

With SyncMirror, tolerate multiple simultaneous disk failures can be tolerated across the RAID groups within the WAFL file system. This redundancy goes beyond typical mirrored (RAID-1) implementations. Because each SyncMirror RAID group is also RAID-4 or RAID-DP protected, a complete mirror could be lost and an additional single drive loss within each RAID group could occur without data loss.

SnapMirror

SnapMirror, shown Figure 9-16, is a software product that allows a data set to be replicated between N series systems over a network for backup or disaster recovery purposes. After an initial baseline transfer of the entire data set, subsequent updates only transfer new and changed data blocks from the source to the destination, which makes SnapMirror highly efficient in terms of network bandwidth utilization. The destination file system is available for read-only access, or the mirror can be “broken” to enable writes to occur on the destination.

After breaking the mirror, it can be reestablished by synchronizing the changes made to the destination back onto the source file system. In the traditional asynchronous mode of operation, updates of new and changed data from the source to the destination occur on a schedule defined by the storage administrator. These updates could be as frequent as once per minute or as infrequent as once per week, depending on user requirements.

Synchronous mode is also available, which sends updates from the source to the destination as they occur, rather than on a schedule. If configured correctly, this can guarantee that data written on the source system is protected on the destination even if the entire source system fails due to natural or human-caused disaster. A semi-synchronous mode is also provided, which can minimize loss of data in a disaster while also minimizing the performance impact of replication on the source system. In order to maintain consistency and ease of use, the asynchronous and synchronous interfaces are identical with the exception of a few additional parameters in the configuration file.

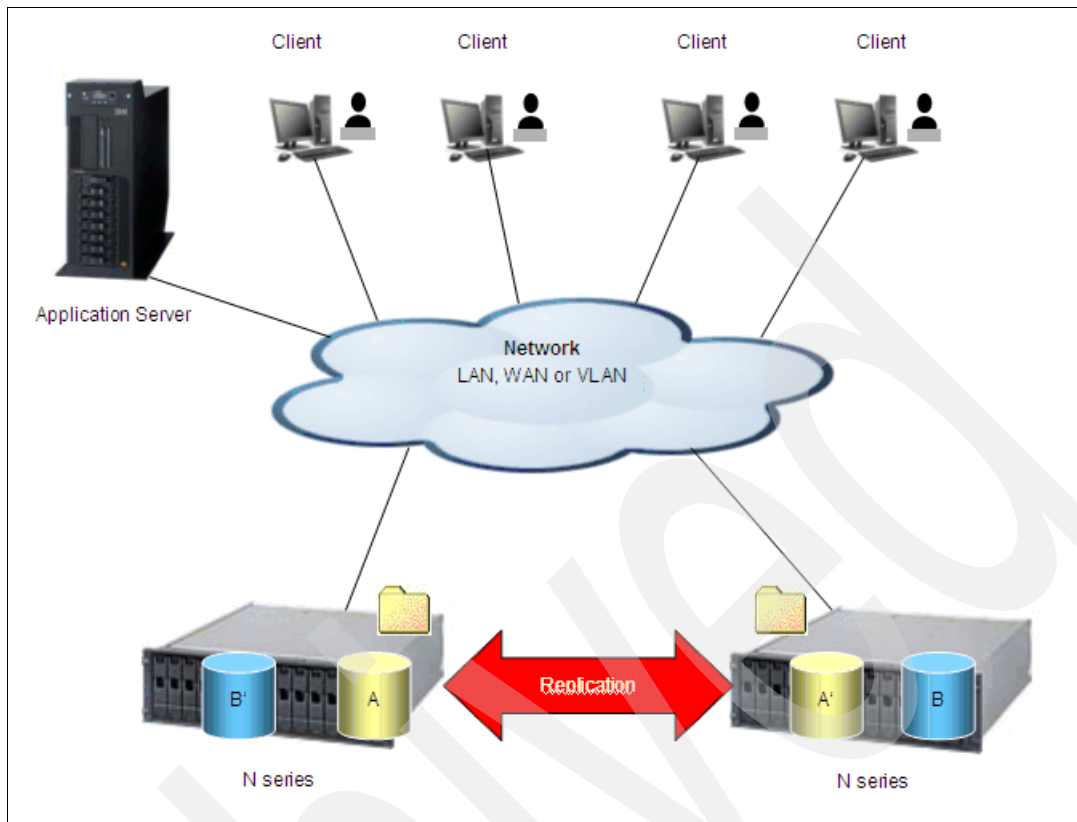


Figure 9-16 IBM N series SnapMirror

Three modes of SnapMirror

SnapMirror can be used in three different modes:

- ▶ Asynchronous
- ▶ Synchronous
- ▶ Semi-synchronous

SnapMirror asynchronous mode

In asynchronous mode, SnapMirror performs incremental, block-based replication as frequently as once per minute. Performance impact on the source N series is minimal as long as the system is configured with sufficient CPU and disk I/O resources.

The first and most important step in asynchronous mode, is the creation of a one-time, baseline transfer of the entire dataset. This is required before incremental updates can be performed. This operation proceeds as follows:

1. The primary storage system takes a Snapshot copy (a read-only, point-in-time image of the file system).
2. This Snapshot copy is called the baseline Snapshot copy (Figure 9-17).
3. All data blocks referenced by this Snapshot copy, and any previous Snapshot copies, are transferred and written to the secondary file system.
4. After initialization is complete, the primary and secondary file systems have at least one Snapshot copy in common.

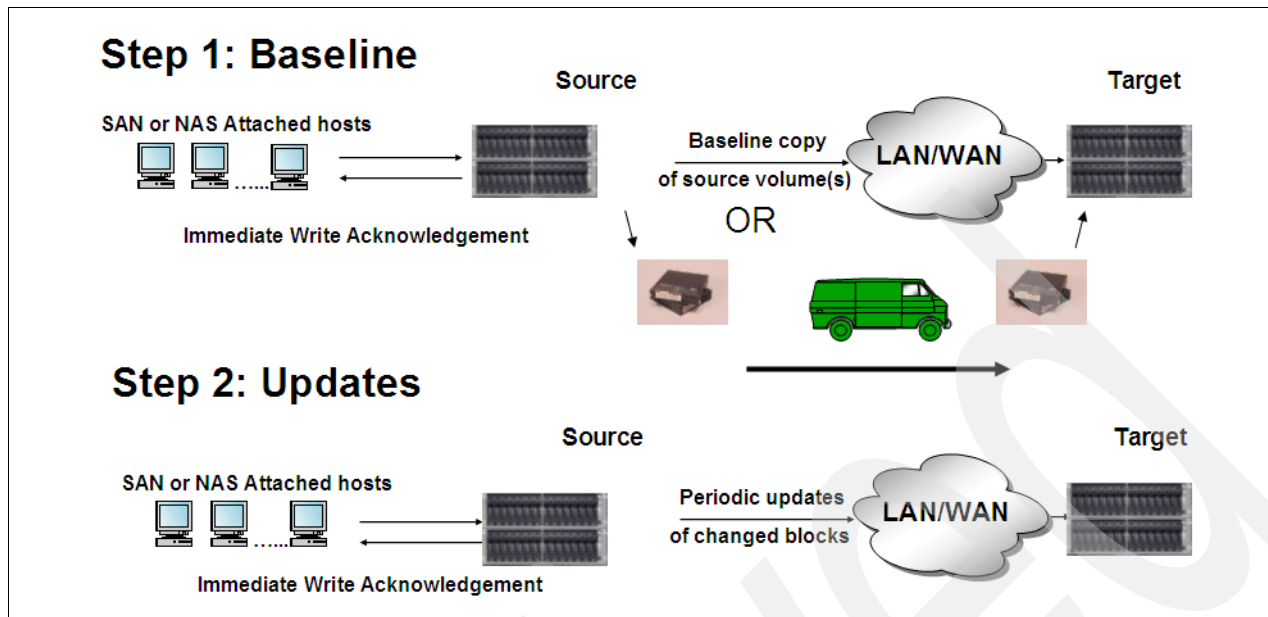


Figure 9-17 Baseline creation

After initialization, scheduled or manually triggered, updates can occur. Each update transfers only the new and changed blocks from the primary to the secondary file system. This operation proceeds as follows:

1. The primary storage system takes a Snapshot copy.
2. The new Snapshot copy is compared to the baseline Snapshot copy to determine which blocks have changed.
3. The changed blocks are sent to the secondary and written to the file system.
4. After the update is complete, both file systems have the new Snapshot copy, which becomes the baseline Snapshot copy for the next update.

Because asynchronous replication is periodic, SnapMirror can consolidate writes and conserve network bandwidth.

SnapMirror synchronous mode

Synchronous SnapMirror replicates data from a source volume to a partner destination volume at or near the same time that it is written to the source volume, rather than according to a predetermined schedule. This guarantees that data written on the source system is protected on the destination even if the entire source system fails and guarantees zero data loss in the event of a failure, but it might result in a significant impact on performance. It is not necessary or appropriate for all applications.

With traditional asynchronous SnapMirror, data is replicated from the primary storage to a secondary or destination storage device on a schedule. If this schedule were configured to cause updates once per hour, for example, it is possible for a full hour of transactions to be written to the primary storage, and acknowledged by the application, only to be lost when a failure occurs before the next update. For this reason, many setups try to minimize the time between transfers — by replicating as frequently as once per minute, which significantly reduces the amount of data that could be lost in a disaster.

This level of flexibility is good enough for the vast majority of applications and users. In most real-world environments, loss of one to five minutes of data is of trivial concern compared to the downtime incurred during such an event; any disaster that completely destroys the data on the N series would most likely also destroy the relevant application servers, critical network infrastructure, etc.

However, there are some users and applications that have a zero data loss requirement, even in the event of a complete failure at the primary site. Synchronous mode is appropriate for these situations. It modifies the application environment described above such that replication of data to the secondary storage occurs with each transaction.

SnapMirror semi-synchronous mode

SnapMirror also provides a semi-synchronous mode, sometimes called semi-sync. Synchronous SnapMirror can be configured to lag behind the source volume by a user-defined number of write operations or milliseconds. This mode is like asynchronous mode in that the application does not have to wait for the secondary storage to acknowledge the write before continuing with the transaction. Of course, for this very reason it is possible to lose acknowledged data. This mode is like synchronous mode in that updates from the primary storage to the secondary storage occur right away, rather than waiting for scheduled transfers. This makes the potential amount of data lost in a disaster very small. Semi-Synchronous minimizes data loss in a disaster while also minimizing the extent to which replication impacts the performance of the source system.

Semi-synchronous mode provides a middle ground that keeps the primary and secondary file systems more closely synchronized than asynchronous mode. Configuration of semi-synchronous mode is identical to configuration of synchronous mode, with the addition of an option that specifies how many writes can be outstanding (unacknowledged by the secondary system) before the primary system delays acknowledging writes from the clients.

If the secondary storage system is slow or unavailable, it is possible that a large number of transactions could be acknowledged by the primary storage system and yet not protected on the secondary. These transactions represent a window of vulnerability to loss of acknowledged data. For a window of zero size, clients can of course use fully synchronous mode rather than semi-sync. If using semi-sync, the size of this window is customizable based on user and application requirements. It can be specified as a number of operations, milliseconds, or seconds.

If the number of outstanding operations equals or exceeds the number of operations specified by the user, further write operations are not acknowledged by the primary storage system until some have been acknowledged by the secondary. Likewise, if the oldest outstanding transaction has not been acknowledged by the secondary within the amount of time specified by the user, further write operations are not acknowledged by the primary storage system until all responses from the secondary are being received within that time frame.

Cascading Mirrors

Cascading (Figure 9-18) is a method of replicating from one destination system to another in a series. For example, one might want to perform synchronous replication from the primary site to a nearby secondary site, and asynchronous replication from the secondary site to a far-off tertiary site. Currently only one synchronous SnapMirror relationship can exist in a cascade.

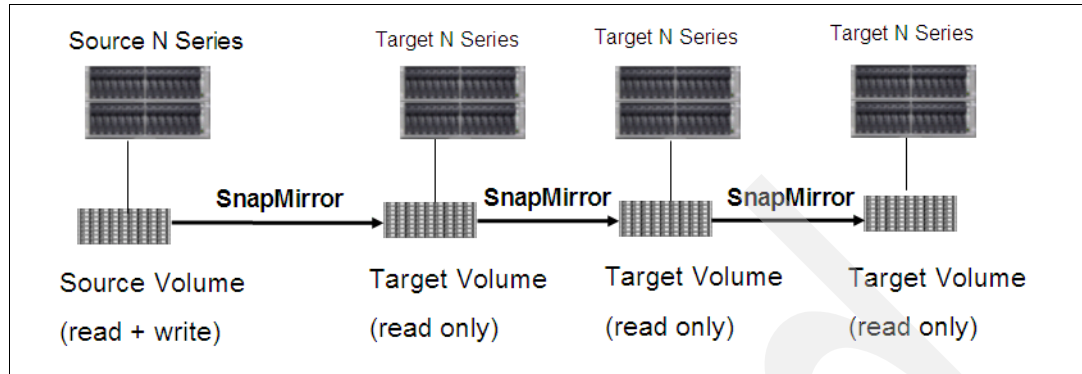


Figure 9-18 Cascading Mirrors

Metro Cluster

MetroCluster is an integrated, high availability, and business-continuation feature that allows clustering of two N5000 or N7000 storage controllers at distances up to 100 kilometers.

MetroCluster technology is an important component of enterprise data protection strategies. If a disaster occurs at a source site, businesses can continue to run and access data from a clustered node in a remote site. The primary goal of MetroCluster is to provide mission-critical applications redundant storage services in case of site-specific disasters. It is designed to tolerate site-specific disasters with minimal interruption to mission-critical applications and zero data loss by synchronously mirroring data between two sites.

A MetroCluster system is made up of the following components and requires the following licenses:

- ▶ Multiple storage controllers, HA configuration. Provides automatic failover capability between sites in case of hardware failures.
- ▶ SyncMirror. Provides an up-to-date copy of data at the remote site; data is ready for access after failover without administrator intervention.
- ▶ Cluster remote. Provides a mechanism for administrator to declare site disaster and initiate a site failover via a single command for ease of use.
- ▶ FC switches. Provide storage system connectivity between sites that are greater than 500 meters apart.

MetroCluster allows the active/active configuration to be spread across data centers up to 100 kilometers apart. In the event of an outage at one data center, the second data center can assume all affected storage operations lost with the original data center. SyncMirror is required as part of MetroCluster to ensure that an identical copy of the data exists in the second data center should the original data center be lost:

1. MetroCluster along with SyncMirror extends active/active Clustering across data centers up to 100 kilometers apart (Figure 9-19).
2. MetroCluster and SyncMirror provide the highest level of storage resiliency across a local region
3. Highest levels of regional storage resiliency ensure continuous data availability in a particular geography

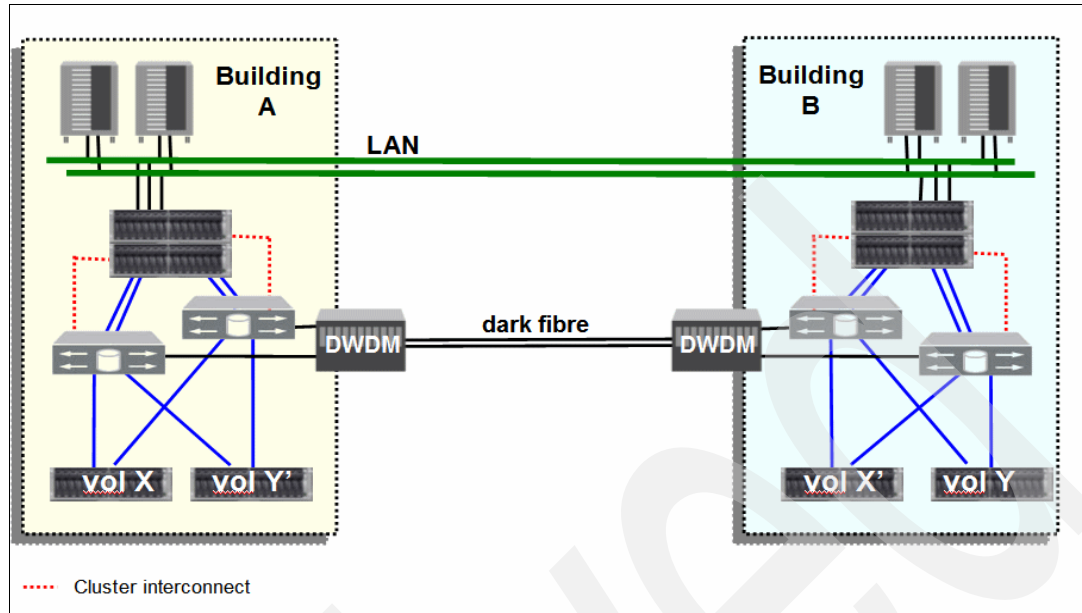


Figure 9-19 MetroCluster

Metro Cluster benefits

MetroCluster offers the following benefits:

- ▶ MetroCluster is designed to be a simple-to-administer solution that extends failover capability from within a data center to a remote site.
- ▶ It is also designed to provide replication of data from the primary site to a remote site, helping keep data at the remote site current.
- ▶ The combination of failover and data replication aids in the recovery from disaster — helping prevent loss of data — in less time than otherwise possible.
- ▶ It extends Clustered Failover capabilities from the primary site to the remote site.
- ▶ It replicates data from the primary site to the remote site to ensure that data there is completely up-to-date and available.

If Site A goes down, MetroCluster allows you to rapidly resume operations at a remote site minutes after a disaster.

9.4 More information

For more information about IBM System Storage, refer to *The IBM System Storage N Series*, SG24-7129, and see the following Web site:

<http://www.ibm.com/servers/storage/nas/>

DS300 and DS400

The IBM TotalStorage DS300 and DS400 products provide solutions for workgroup storage applications. In this chapter we discuss the following features:

- ▶ DS300 and DS400 storage systems
- ▶ DS300 and DS400 FlashCopy service

10.1 DS300 and DS400 overview

The IBM TotalStorage DS300 and DS400 products offer low-cost entry points to the IBM disk family. They are designed to work with System x and BladeCenter servers in either direct-attached or network attached configurations. These products provide solutions for workgroup storage applications, such as file, print and Web function, as well as for collaborative databases and remote booting of diskless servers. With Microsoft Windows Server 2003, they support boot from SAN capabilities without requiring the dedicated boot disk to be local to the server.

10.2 Introduction

DS300 and DS400 copy functions fall under Tier 4 for the FlashCopy function as shown in Figure 10-1. DS300 and DS400 provide scalability and performance that can help you consolidate storage. The DS300 and DS400 allow an administrator to add storage and perform upgrades easily.

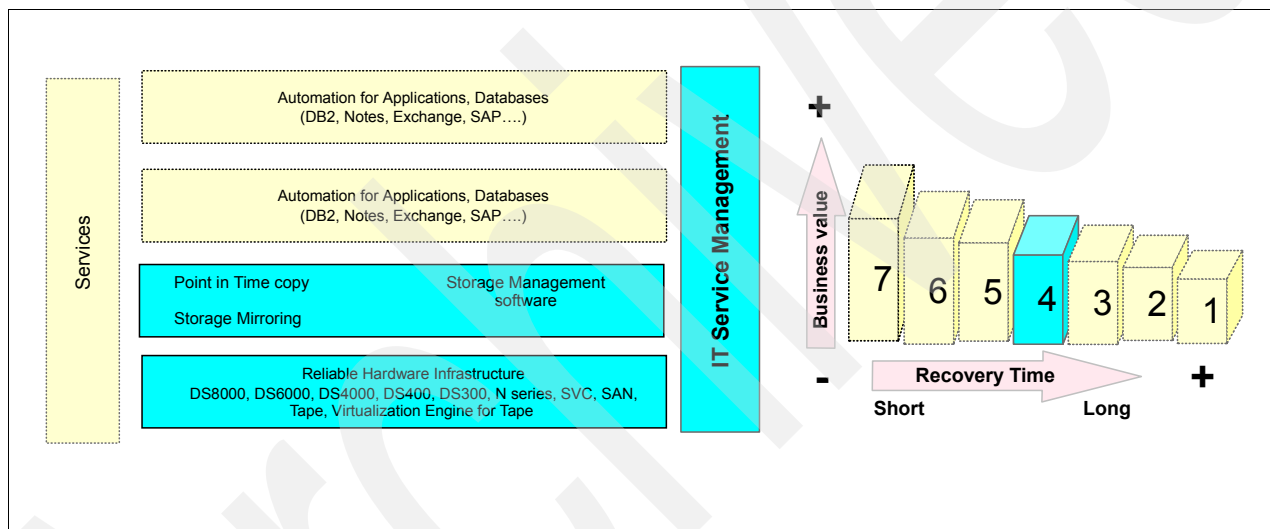


Figure 10-1 Tier level graph

Note: At the time of writing this book, there is no support for online capacity expansion and RAID migration. Online capacity expansion and RAID migration are planned to be made available in the future.

10.3 IBM TotalStorage DS400 and DS300

The IBM TotalStorage DS400 is a Fibre Channel (FC) to SCSI storage subsystem, and DS300 is an iSCSI to SCSI storage subsystem. Both DS400 and DS300 use SCSI drives, which cost less than FC disk drives, and both use a common management software, ServeRAID™ PCI RAID controllers, which simplify migration from Direct Attached Storage (DAS) to SAN configurations.

To attach to host servers, the DS300 and DS400 use an iSCSI server attachment. DS400 has features to attach via Fibre Channel (FC). The DS300 and DS400 products are equipped with redundant, hot swappable power and cooling modules, battery-backed data caches, RAID reliability, and high availability software. DS300 and DS400 have high availability functions like FlashCopy to help protect critical data while minimizing storage backup windows. DS300 and DS400 are designed for entry level storage solutions for BladeCenter and System x servers. Refer to the following Table 10-1 for a DS300 and DS400 product overview.

Table 10-1 DS300 and DS400 product specification and overview

	DS300 1RL	DS300 Single Controller	DS300 Dual Controller	DS400 Single Controller	DS400 Redundant Controller
Controller	Single controller	Single controller	Dual controller	Single controller	Dual controller
RAID level supported	0, 1, 10, 5, 50	0, 1, 10, 5, 50	0, 1, 10, 5	0, 1, 10, 5, 50	0, 1, 10, 5
Controller option upgrade	No	Yes	NA	Yes	NA
2GB FC ports per controller	1	2	2	2	2
System form factor	3U, 7 - 14 drives	3U, 14 drives	3U, 14 drives	3U, 14 drives	3U, 14 drives
Cache memory/Controller	None	256 MB upgradeable to 1 GB	256 MB upgradeable to 1 GB	256 MB upgradeable to 1 GB	256 MB upgradeable to 1 GB
Battery backup	NA	Yes	Yes	Yes	Yes
Maximum disk drives with EXP400 attach	NA	NA	NA	40	40
Maximum capacity	4.2 TB	4.2 TB	4.2 TB	12 TB	12 TB
Power supplies	Single	Dual redundant	Dual redundant	Dual redundant	Dual redundant
Supported Ultra320 SCSI HDDs	10K: ▶ 36/73/146/300 GB 15K: 36/73/146 GB	10K: ▶ 36/73/146/300 GB 15K: ▶ 36/73G/146 GB	10K: ▶ 36/73/146/300 GB 15K: ▶ 36/73/146 GB	10K: ▶ 36/73/146/300 GB 15K: ▶ 36/73/146 GB	10K: ▶ 36/73/146/300 GB 15K: ▶ 36/73/146 GB
Options	▶ Second power supply kit to support up to 14 drives ▶ FlashCopy	▶ Second controller upgrade ▶ FlashCopy	NA	▶ Second controller upgrade ▶ JBOD expansion ▶ FlashCopy	▶ JBOD expansion

Note: FlashCopy supports up to 254 volumes.

For DS400, a Fibre Channel switch is required, for example, the System Storage SAN10Q-2. for System x storage applications. It has 10 ports and each port is capable of running 1 Gbps, 2 Gbps, or 4Gbps.

DS300 operates on iSCSI that enables you to connect the storage over an IP network — iSCSI or Internet SCSI, which maps SCSI blocks into Ethernet packets. The iSCSI protocol is a method for transporting *low latency SCSI blocks across IP networks*. iSCSI allows you to build a storage area network (SAN) over IP and iSCSI removes the limitations of direct attached storage (DAS), including the inability to share storage resources across servers and to expand capacity without shutting down applications. iSCSI can be connected via multiple paths between servers and storage that enables high availability of storage to users.

iSCSI requires a software initiator to work . The recommended software initiators are:

- ▶ Microsoft iSCSI software initiator (Version 1.06a)
- ▶ Cisco iSCSI initiator for Linux and Windows (Cisco 3.43 sourceforge.net)
- ▶ Novell iSCSI initiator for NetWare

iSCSI can also be used through a hardware initiator.

Figure 10-2 shows the iSCSI architecture.

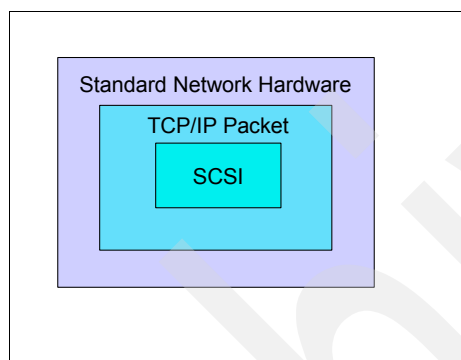


Figure 10-2 iSCSI architecture

Figure 10-3 shows the IP storage protocol.

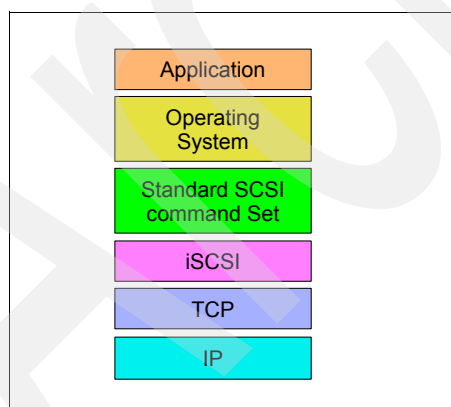


Figure 10-3 IP storage protocol summary

Refer to the following Web site for more information about the iSCSI protocol:

<http://www.snia.org>

Refer to the following Web site for more information about System x support for iSCSI:

<http://www.pc.ibm.com/us/eserver/xseries/library/cog.html>

The DS400 design architecture involves a common controller board for Fibre Channel (FC) and the DS300 design architecture involves an iSCSI subsystem controller. Both DS300 and DS400 use a common software code set for subsystems. The common management Graphic User Interface (GUI) is IBM ServerRAID Manager, which has the ability to manage both ServerRAID PCI controllers and external storage from the same management console. The ServerRAID Manager is a centralized storage management tool, which includes functions such as configuration, monitoring, and management of RAID storage resources, and is designed to provide automatic detection of events or failures. It is capable of event logging on the local management station. As a prerequisite, you must install both a management agent and the client GUI to manage the disk enclosures.

DS400 and DS300 come with optional FlashCopy Services with rollback capability.

Refer to the following Web sites for more information:

- ▶ DS400:
<http://www-1.ibm.com/servers/storage/disk/ds/ds400/index.html>
- ▶ DS300:
<http://www-1.ibm.com/servers/storage/disk/ds/ds300/index.html>

10.4 Data protection

Although DS300 and DS400 forms the entry level external storage ideal for small and medium enterprises, it offers data protection features which you can utilize to obtain a certain resiliency level on your IT infrastructure.

10.4.1 DS300, DS400 copy functions

In this section we discuss the copy functions of the DS300 and DS400.

DS300, DS400 FlashCopy

DS300 and DS400 provide FlashCopy capability to create a point-in-time copy of the source volume. You can use the FlashCopy Management Command-Line Tool to create and manage FlashCopies of application databases. A FlashCopy is a point-in-time image of an application database. It can be used as a rollback point for future application problems.

With the FlashCopy Management Command Line Tool, you can:

- ▶ List Mini Snap Agents available on the IPSAN
- ▶ List available databases
- ▶ Take a FlashCopy of the database
- ▶ Schedule a FlashCopy of the database
- ▶ Rollback to a previous FlashCopy
- ▶ Delete a FlashCopy and its associated metadata
- ▶ Print usage information

DS300/400 FlashCopy functions include:

- ▶ List instance database. Returns a list of databases available on a server instance.
- ▶ Take snapshot. Takes a FlashCopy of the database.
- ▶ Unschedule snapshot. Removes a FlashCopy job from the host's diary.
- ▶ Roll back. Rolls back the database to a specific FlashCopy.
- ▶ Delete snapshot. Deletes a FlashCopy and its associated metadata.
- ▶ Supports up to four flashes (base feature), or optionally up to 256.
- ▶ Copy on Write feature.

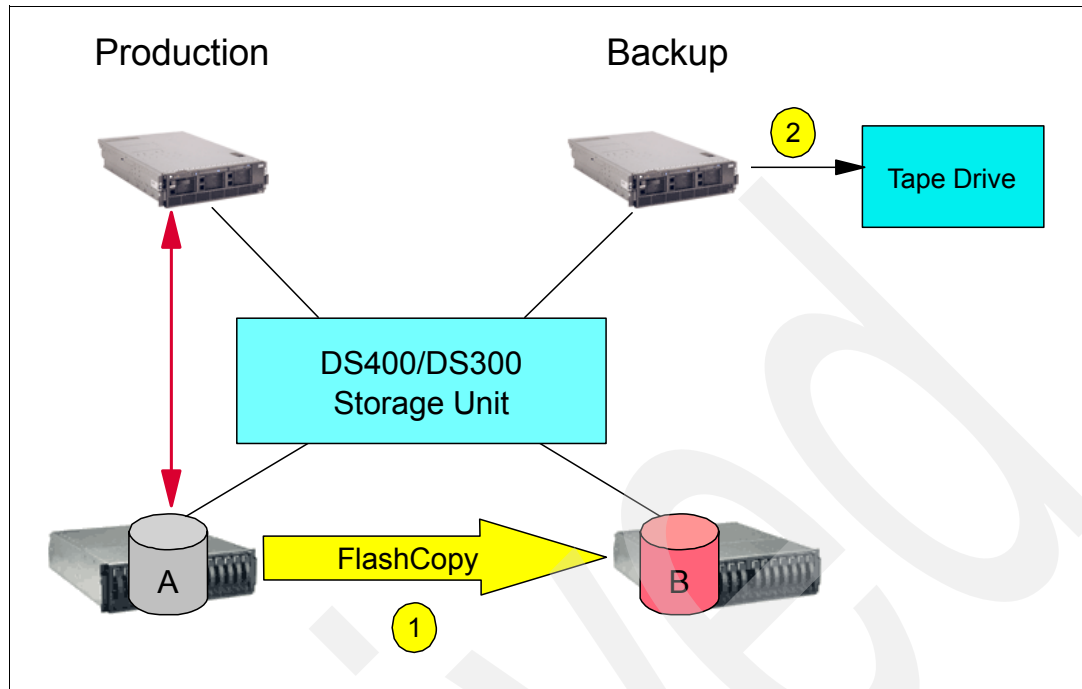


Figure 10-4 DS400 DS300 FlashCopy

There are two approaches to FlashCopy on the DS300 and DS400 for open databases or applications. The first approach is the conventional way of stopping a database before a FlashCopy can take place, and the second approach is to have an application agent that can quiesce the database before a FlashCopy is performed. Currently there is no integration for DS300 and DS400 on Microsoft Volume Shadow Copy Services (VSS) and a script has to be written in order to perform online database FlashCopy.

The *first approach* is a straightforward method: You have to stop the database to prevent any I/O writes to the disk prior to a FlashCopy. By performing a complete stop or shutdown on the database, this provides you with consistent data, ready for FlashCopy.

The *second approach* is to use an application agent. An application agent is included in individual software providers. You can check with your software provider for more information about an application agent that is capable of quiescing a database. An example of an application agent is Microsoft SQL Server Virtual Device Interface (VDI). Prior to the FlashCopy process, you have to create a script that uses the application agent to quiesce the database to prevent I/O writes to disk.

When the database is quiesced, FlashCopy can be initiated to perform a point-in-time copy of the source volume where the database resides. After the FlashCopy process completes, the application agent can resume the database I/O writes. This approach makes certain that you have a consistent FlashCopy image of your source image. The FlashCopy volume can then be mounted on a backup server for tape backup as shown in Figure 10-4. The second approach can be used if you do not want to stop the production database to perform a FlashCopy.

Boot from SAN

DS400 and DS300 allow you to boot from the SAN rather than from local disks on individual servers, which can enable you to maximize consolidation of IT resources and minimize equipment cost. Boot from SAN is a remote boot technology where the source of the boot disk is residing on the SAN. The server communicates with the SAN through host bus adapters (HBA). The HBA BIOS contains the instructions that enable the server to find the boot disk on the SAN. Booting from the SAN provides you a rapid disaster recovery, because all the boot information and production data is stored on a local or remote SAN environment.

DS300 operates on iSCSI, and it supports booting from SAN using iSCSI interconnect to the SAN, provided that iSCSI HBAs are used to enable the boot process.

Attention: Boot from SAN is not supported using a software iSCSI software initiator. No sharing of boot images is allowed, as Windows servers cannot currently share a boot image. Each server requires its own dedicated LUN to boot.

Refer to the Microsoft Web site for more information and implementation on booting from SAN:

<http://www.microsoft.com/windowsserversystem/storage/solutions/sanintegration/bootfromsaninwindows.msp>

10.5 Disaster recovery considerations

DS300 and DS400 serve as an entry level for clients who want to adopt a SAN environment. Because the FlashCopy function serves as a point-in-time copy of the source volume or production server, you can use it to off-load the backup process without having to shut down your production application server, provided that you utilize the application agent from the application software vendor. The FlashCopy volume can also be mounted for testing or data mining purposes.

Note: At the time of writing this book, DS300 does not support multipathing — it is possible to set it up in some versions of Linux; however this is not supported by IBM. You can implement a redundant network topology to ensure that no single link failure or no single device failure disrupts iSCSI attached devices.

For online FlashCopy of a database, you must make sure that your software providers provide an application agent that is capable of performing a quiesce of the database. Without this application agent, FlashCopy is not able to perform a coherent copy. If online database FlashCopy is not desired, you can stop the application or database and perform a FlashCopy. If you require an integration between databases and the FlashCopy function, you can refer to other DS storage family, for example, DS4000, DS6000, or DS8000.

The maximum configuration for the DS400 is 40 drives with two EXP400 expansion units and 14 drives for DS300. If you require more drives that exceed the maximum disk storage possible, you can refer to details on the DS4000 family in Chapter 8, “The IBM System Storage DS4000” on page 295.

Archived

Storage virtualization products

In this chapter we discuss general aspects of storage virtualization for open systems environments and describe the IBM storage virtualization products in relation to disaster recovery solutions.

The Storage Networking Industry Association (SNIA) in its SNIA Dictionary, *Network Storage Terms and Acronyms*, defines storage virtualization as:

“The act of integrating one or more (back-end) services or functions with additional (front end) functionality for the purpose of providing useful abstractions. Typically virtualization hides some of the back-end complexity, or adds or integrates new functionality with existing back-end services. Examples of virtualization are the aggregation of multiple instances of a service into one virtualized service, or to add security to an otherwise insecure service. Virtualization can be nested or applied to multiple layers of a system.”

We cover the following IBM storage virtualization products in this chapter:

- ▶ IBM System Storage SAN Volume Controller
- ▶ IBM System Storage N series Gateway
- ▶ Enterprise Removable Media Manager

For more information about SVC, refer to the IBM Web site:

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

For more information about the N series Gateway, refer to the IBM Web site:

<http://www.ibm.com/servers/storage/network/n7000/gateway/>
<http://www.ibm.com/servers/storage/network/n5000/gateway/>

11.1 Storage virtualization overview

This chapter discusses the different possibilities for storage virtualization. It describes briefly the general aspects and conclusions of the different concepts as well as the products available from IBM and how you can use them for business continuity planning.

11.1.1 Levels of storage virtualization

There are several levels of storage virtualization as illustrated in Figure 11-1.

Each level of storage virtualization has different characteristics. There are good reasons to design a product on each level. Managing the complexity of the growing storage and reducing costs are the main motivations behind the developments.

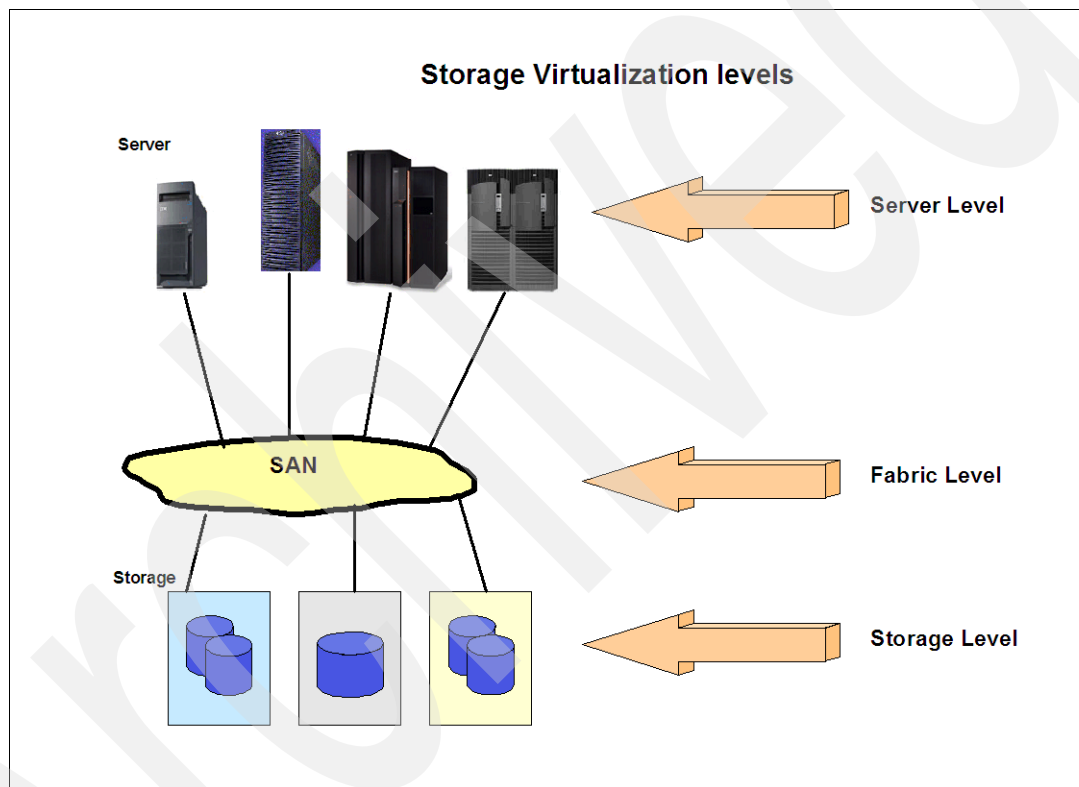


Figure 11-1 Storage virtualization levels

Storage level

The lowest level is the storage system level. Almost all disk storage systems today provide some functionality of virtualization already by assigning logical volumes to physical disks or disk groups. Here are some examples of this level:

- ▶ IBM System Storage DS8000 disk system
- ▶ IBM System Storage DS4000 disk system
- ▶ IBM System Storage TS7700 tape system

The benefit on this level is, that the storage system can optimize the mapping between the logical and physical storage, which maximizes use of the physical storage. The main restriction is, that virtualization happens within the system, and does not span multiple systems.

Fabric level

At the fabric level, virtualization can enable the independence of storage from heterogeneous servers. The SAN fabric is zoned to allow the virtualization appliances to see the storage systems and for the servers to see the virtualization appliances. Different pools of storage can be built, and advanced functions such as snapshot and remote mirroring can be introduced. Servers cannot directly see or operate on the disk systems. Here are some examples of this level:

- ▶ IBM System Storage SAN Volume Controller
- ▶ IBM System Storage N series Gateway

Server level

Abstraction at the server level is achieved using logical volume management of the server operating systems. At first glance, increasing the level of abstraction on the server seems well suited for environments without storage networks, but this can be vitally important in storage networks, too. The dependencies of the different levels of different operating systems is potentially higher than on the fabric level. A higher complexity for implementation and maintenance is the result. Virtualization can take place in the Host Bus Adapter (HBA), within the device drivers, or in the file system. Here is an example of this level:

- ▶ IBM System p5 AIX Volume Manager

In-band versus out-of-band

There are two methods for providing storage virtualization:

- ▶ In-band (also called symmetric virtualization)
- ▶ Out-of-band (also called asymmetric virtualization)

Figure 11-2 illustrates the differences between these two methods.

In an in-band solution, both data and control information flow over the same path. Levels of abstraction exist in the data path, and storage can be pooled under the control of a domain manager. In general, in-band solutions are perceived to be simpler to implement, especially on the fabric level, because they do not require special software to be installed on the servers, except the device drivers for multi-pathing and failover.

In an out-of-band implementation, the data flow is separated from the control information flow. This is achieved by separating the data and metadata (data about the data) into different places. The mapping and locking tables are maintained on a separate server that contains the files' metadata. File I/O proceeds directly over the SAN between the storage devices and servers. The metadata server grants access control requests, and handles file locking. Separating the flow of control and data in this manner allows the I/O to use the full bandwidth that a SAN provides, while control goes over a separate network or routes in the SAN that are isolated for this purpose.

Note that both in-band and out-of-band methods can be implemented on the server level or on the switch level. The IBM SAN Volume Controller uses the in-band method on the fabric level.

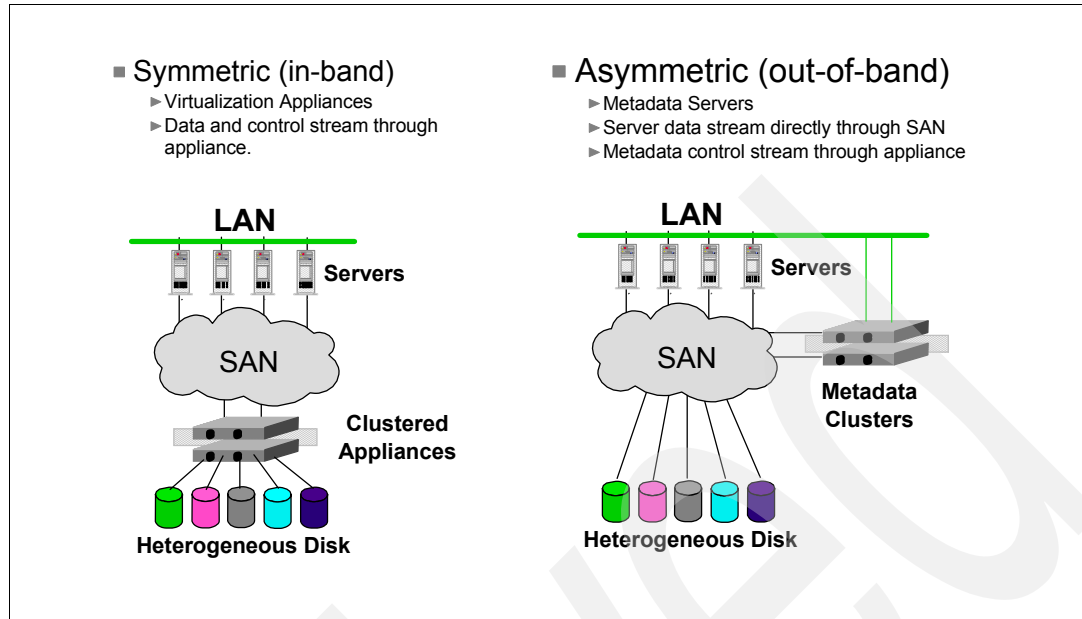


Figure 11-2 Symmetric versus asymmetric virtualization

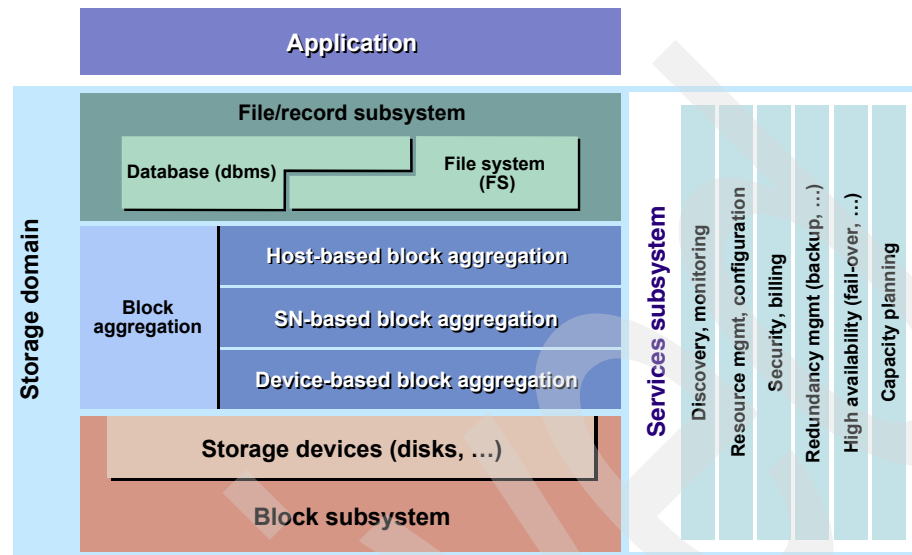
11.1.2 SNIA shared storage model

The Storage Network Industry Association (SNIA) was incorporated in December 1997 as a nonprofit trade association that is made up of over 200 companies. SNIA includes well established storage component vendors as well as emerging storage technology companies.

The SNIA mission is “to ensure that storage networks become efficient, complete, and trusted solutions across the IT community”.

IBM is an active member of SNIA and fully supports SNIA’s goals to produce the open architectures, protocols, and APIs required to make storage networking successful. IBM has adopted the SNIA Storage Model and we are basing our storage software strategy and roadmap on this industry-adopted architectural model for storage, as depicted in Figure 11-3.

The SNIA Storage Model



Copyright 2000, Storage Network Industry Association

Figure 11-3 The SNIA storage model

IBM is committed to deliver the best-of-breed products in various aspects of the SNIA storage model:

- ▶ Block aggregation
- ▶ Storage devices/block subsystems
- ▶ Services subsystems

11.1.3 The Storage Management Initiative (SMI)

In 2002 the SNIA launched the *Storage Management Initiative (SMI)*. The goal of this initiative is to develop management standards for storage technology. The SMI provides a common interface for storage vendors to incorporate in the development of new products. The SMI is based on the common protocol called *CIM (Common Information Model)*. The Web version of CIM is used, which is called *WEBEM (Web Based Enterprise Management)*. CIM defines objects and interactions.

There are many publications describing this initiative. The SNIA Web site is a central starting point:

<http://www.snia.org>

What makes it important for Business Continuity is that automation is key. Based on these standards, general tools are planned to be developed to manage these storage systems and to automate tasks useful for Business Continuity, independent of the vendor's storage systems.

The IBM disk systems ship with CIM agents, including the SAN Volume Controller, DS8000, DS6000, and DS4000.

11.1.4 Multiple level virtualization example

In a normal IT environment you might be using multiple levels of virtualization or abstraction at the same time, probably without knowing it. Consider the example shown in Figure 11-4. The example shows a stack composed of four layers of storage virtualization performed at different levels. We use the term *lvm* to represent a generic logical volume manager function. The lines represents the path of data between an application server and the storage device.

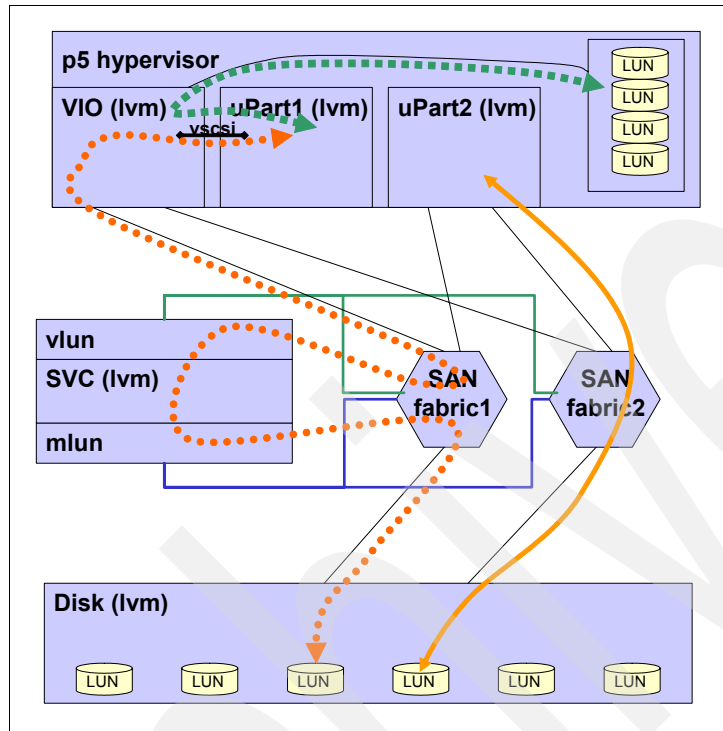


Figure 11-4 Multiple levels of virtualization

The bottom layer is the storage system layer, represented by an IBM disk system. Physical disks in the system are aggregated and partitioned into LUNs. These resulting LUNs are assigned to servers. Aggregation and mapping is performed by the lvm function.

The middle layer has a SAN Volume Controller. It uses its own lvm function, to aggregate managed disks in the storage system and then partition the resulting space into virtual disks. These virtual disks are then assigned to host servers.

The top layer is a POWER5 server. The POWER5 architecture defines the Virtual I/O Server function that is used to aggregate disks and partition the resulting space, in the example it accesses SAN Volume Controller disks. The Virtual I/O Server behaves as a logical volume manager. The Virtual I/O Server then makes these disks available to Virtual I/O clients that run in micro partitions (the example shows uPart1 and uPart2) on the POWER5 server.

The POWER5 Virtual I/O clients see SCSI LUNs published by the Virtual I/O server. The Virtual I/O client then accesses and formats these LUNs using its own lvm. This is the fourth level of virtualization in our example.

The choice of available products to perform virtualization are many and virtualization can be implemented in multiple levels at the same time.

11.2 IBM System Storage SAN Volume Controller

This section discusses very briefly the IBM System Storage SAN Volume Controller architecture, the copy functions and how the SAN Volume Controller can be integrated into a business continuity solution.

For a detailed description of the IBM SAN Volume Controller, refer to the IBM Redbook, *IBM System Storage SAN Volume Controller*, SG24-6423.

11.2.1 Glossary of commonly used terms

Here are some terms and concepts used in the SAN Volume Controller (SVC) product.

Boss node

A single node acts as the boss node for overall management of the cluster. If the boss node fails, another node in the cluster takes over the responsibilities.

Configuration node

A single node in the cluster is used to manage configuration activity. This configuration node manages a cache of the configuration information that describes the cluster configuration and provides a focal point for configuration commands.

Extent

An extent is a fixed size unit of data that is used to manage the mapping of data between mDisks and vDisks.

Grain

A grain is the unit of data represented by a single bit in a FlashCopy bitmap, 256 K in SVC.

Intracluster

FlashCopy can be performed on the intracluster environment between source and target vDisks that are part of any I/O group within the SVC cluster. Metro Mirroring is supported on intracluster provided that the vDisks are within the same I/O group. Metro Mirror across different I/O groups in the same SVC cluster is *not supported*.

Intercluster

Intercluster Metro Mirror operations requires a pair of SVC clusters that are connected by a number of moderately high bandwidth communication links. The two SVC clusters must be defined in an SVC partnership, which must be defined on both SVC clusters to establish a fully functional Metro Mirror partnership.

I/O group

An input/output (I/O) group contains two SVC nodes defined by the configuration process. Each SVC node is associated with exactly one I/O group. The nodes in the I/O group provide access to the vDisks in the I/O group.

Managed disk

Managed disk (mDisk) is a SCSI disk presented by a RAID controller and managed by the SVC. The mDisks are not visible to host systems on the SAN.

Managed disk group

The managed disk group (MDG) is a collection of mDisks that jointly contain all the data for a specified set of vDisks.

Master console

The master console is the platform on which the software used to manage the SVC runs.

Node

Node is the name given to each individual server in a cluster on which the SVC software runs.

Quorum disks

A quorum disk is used to resolve tie-break situations, when the voting set of nodes disagrees on the current cluster state. The voting set is an overview of the SVC cluster configuration running at a given point-in-time, and is the set of nodes and quorum disk which are responsible for the integrity of the SVC cluster. In a two site solution we recommend that you use two SVC clusters, one at each site, and mirror the data by using either host mirror software/functions or by SVC Metro Mirror.

SAN Volume Controller

The SAN Volume Controller is a SAN appliance designed for attachment to a variety of host computer systems, which carries out block level virtualization of disk storage.

Virtual disk

Virtual disk (vDisk) is a SAN Volume Controller device that appears to host systems attached to the SAN as a SCSI disk. Each vDisk is associated with exactly one I/O group.

11.2.2 Overview

The IBM System Storage SAN Volume Controller provides block aggregation and volume management for disk storage within the SAN. The SVC manages a number of heterogeneous back-end storage controllers and maps the physical storage within those controllers to logical disk images that can be seen by application servers in the SAN. The SAN is zoned in such a way that the application servers cannot see the back-end storage, preventing any possible conflict between SVC and the application servers both trying to manage the back-end storage.

For I/O purposes, the SVC nodes within a cluster are grouped into pairs (called IO groups), with a single pair being responsible for serving I/Os on a given virtual disk. One node within the IO group represents the preferred path for I/O to a given virtual disk, the other node represents the non-preferred path. The SVC cluster (at V4.1, current at the time of writing), can contain up to eight nodes or four I/O groups.

There are three significant benefits:

► **Simplified storage management:**

- The SVC delivers a single view of the storage attached to the SAN. Administrators can manage, add and migrate physical disks nondisruptively even between different storage subsystems.
- The SVC is the central point of administration for the logical volumes assigned to the servers.
- Performance, availability and other service level agreements can be mapped by using different storage pools and advanced functions.
- Dependencies, which exist in a SAN environment with heterogeneous storage and server systems are reduced.

► **Common platform for advanced functions:**

- FlashCopy works even between different storage systems.
- Metro Mirror/Global Mirror is done on the SAN level.
- Data migration between storage subsystems can be performed without application interruption.
- Data migration from existing native storage to SVC storage can be performed with minimal interruption.

► **Improved capacity utilization:**

- Spare capacity on underlying physical disks can be reallocated nondisruptively from an application server point of view irrespective of the server operating system or platform type. Virtual disks can be created from any of the physical disks being managed by the virtualization device.

IBM SVC supports a wide variety of disk storage and host operating system platforms. For the latest information, refer to the IBM Web site:

<http://www.ibm.com/servers/storage/software/virtualization/svc/interop.html>

Figure 11-5 shows a picture of the SAN Volume Controller mode 2145-8F4:



Figure 11-5 SVC- physical view

The SVC 4.1 hardware consists of:

- System x 366 server
- 8 GB cache per node
- 4 Gbps FC Adapter

Figure 11-6 shows a basic schematic — there are two SVC *clusters*, cluster 1 and cluster 2. Cluster 1 is composed of four *nodes*, nodes 1 to 4, in two *I/O groups* I/O group 1 and 2. Each cluster accesses *managed disks* (*mDisk*) through a *SAN storage zone*; The managed disks usually reside in *storage arrays*. These managed disks are grouped into *managed disk groups*, not shown in Figure 11-6. *virtual disks* (*vDisk*) are built on the managed disk groups. The host server *nodes* access *vDisks* through the *SAN host zone*. To use functions such as Metro Mirror, the nodes in both clusters must be connected with a *Metro Mirror SAN zone*.

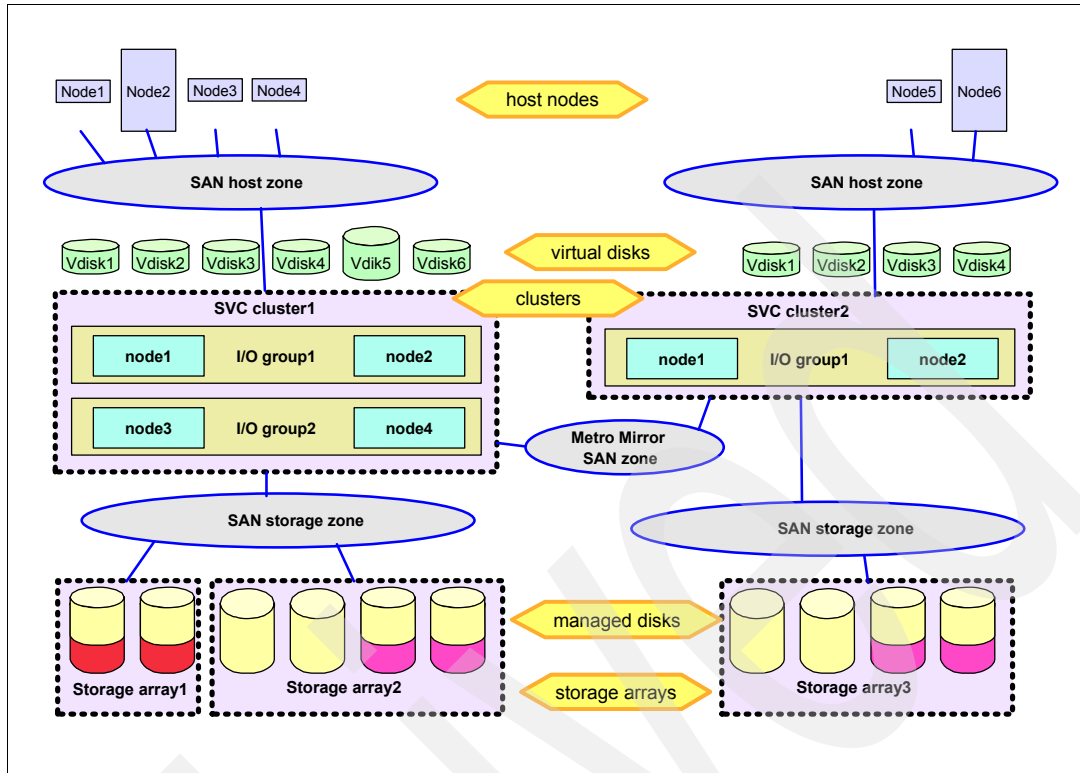


Figure 11-6 SVC - logical view

Figure 11-7 shows SVC connectivity to host systems and storage devices.

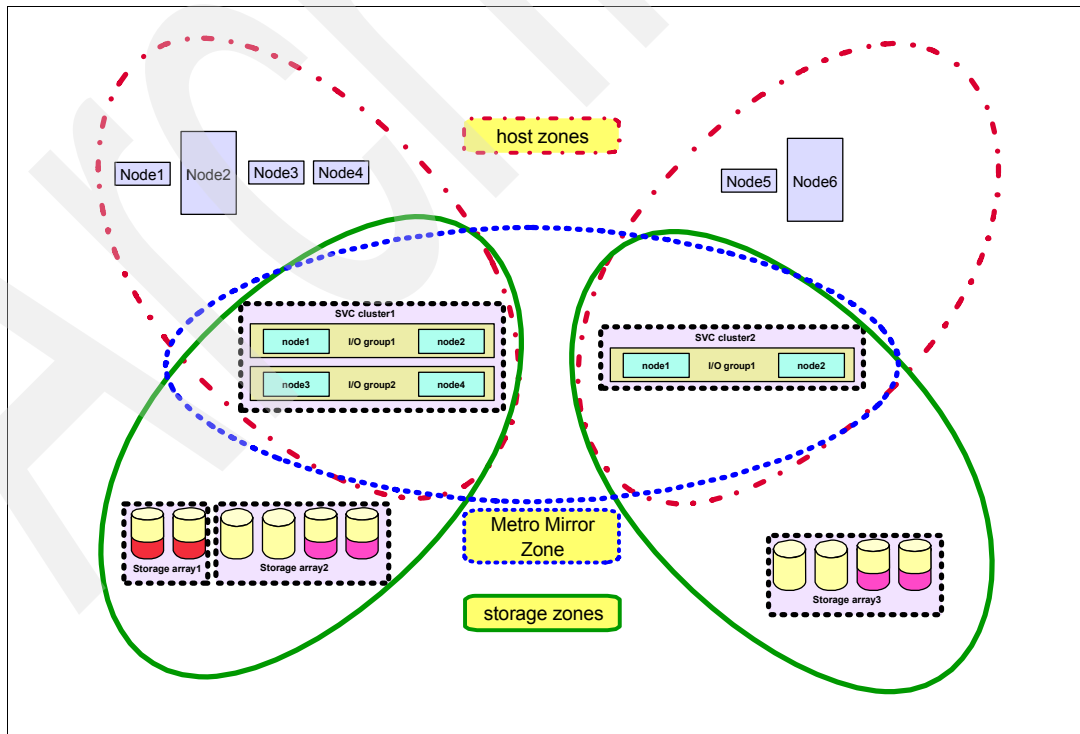


Figure 11-7 Zoning an SVC

The example shows two distinct clusters in two different locations. Host server nodes are zoned to the local SVC with host zones that allow each host node to see all local SVC node ports. Storage devices are zoned to the local SVC with storage zones, again each storage port has to see all SVC nodes. When using Metro Mirror, all nodes in both SVC clusters must have access to each other and so they must be placed in a Metro Mirror zone that encompasses all nodes in both clusters.

11.2.3 SVC copy functions

SVC provides four different copy functions:

- ▶ FlashCopy
- ▶ Metro Mirror
- ▶ Global Mirror
- ▶ Data Migration

The basic benefits of the SVC's copy functions are:

- ▶ Backup
- ▶ Nondisruptive data replacement
- ▶ Business Continuity

The SVC supports Consistency Groups for FlashCopy, Metro Mirror and Global Mirror. One advantage of using SVC Consistency Groups is that they can span the underlying storage systems — so that if the source volumes are on multiple disk systems, a single group can span them all, ensuring data consistency for disaster recovery functions. FlashCopy source volumes that reside on one disk system can write to target volumes on another disk system. Similarly, Metro Mirror and Global Mirror source volumes can be copied to target volumes on a dissimilar storage system.

SVC FlashCopy

FlashCopy (see Figure 11-8) is a point-in-time copy of a virtual disk on an SVC.

FlashCopy works by defining a FlashCopy mapping consisting of a source VDisk and a target VDisk. Multiple FlashCopy mappings can be defined and PiT consistency can be observed across multiple FlashCopy mappings using consistency groups.

When FlashCopy is started, it makes a copy of a source VDisk to a target VDisk, and the original contents of the target VDisk are overwritten. When the FlashCopy operation is started, the target VDisk presents the contents of the source VDisk as they existed at the single point in time (PiT) the FlashCopy was started. This is often also referred to as a Time-Zero copy (T_0).

When a FlashCopy is started, the source and target VDisks are instantaneously available. This is so, because when started, bitmaps are created to govern and redirect I/O to the source or target VDisk, respectively, depending on where the requested block is present, while the blocks are copied in the background from the source to the target VDisk.

Both the source and target VDisks are available for read and write operations, although the background copy process has not yet completed copying across the data from the source to target volumes.

The FlashCopy functions can be invoked via the Web-interface, CLI, TPC or scripts.

Important for Business Continuity scenarios, SVC FlashCopy supports Consistency Groups. This means that a group of virtual disks that belong to the same application can hold related data. For data consistency, all of these virtual disks must be flashed together, at a common point-in-time.

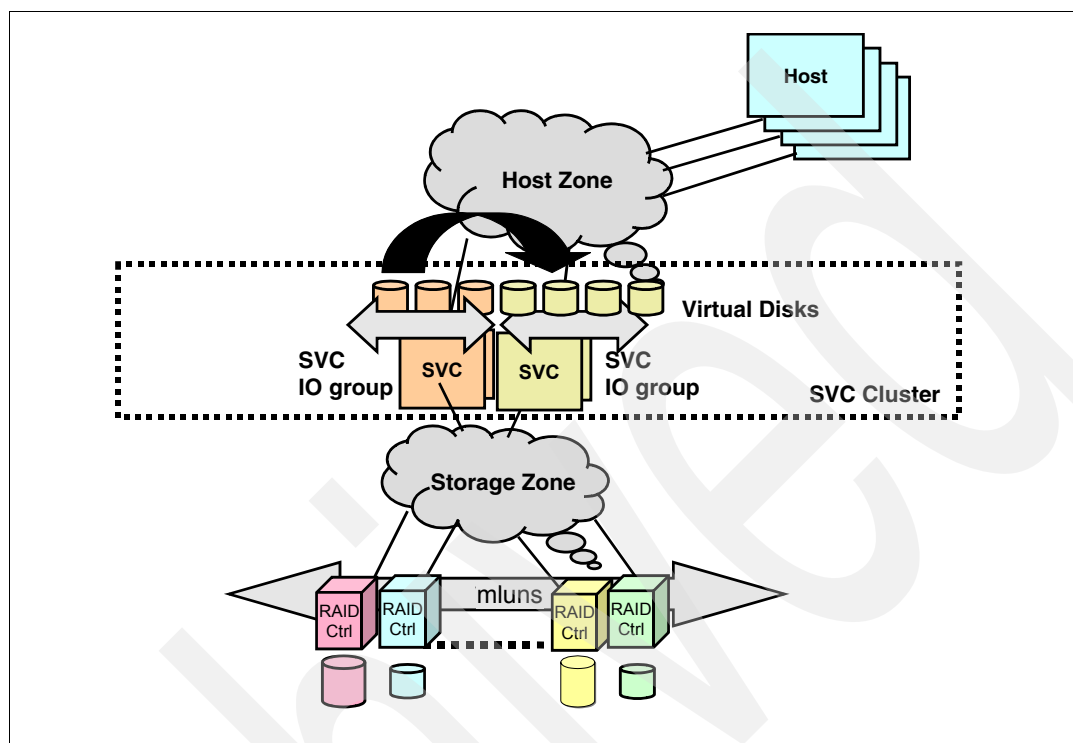


Figure 11-8 SVC FlashCopy

Important: The SVC FlashCopy function operates on the storage level and therefore independent of the application. The first step before invoking this function is to make sure that all of the application data is written to disk - in other words, that the application data is consistent. This can be achieved for example by quiescing a database and flushing all data buffers to disk.

Overview of FlashCopy features

FlashCopy supports these features:

- ▶ The target is the time-zero copy of the source (known as *FlashCopy mapping targets*).
- ▶ The source VDisk and target VDisk are available (almost) immediately.
- ▶ Consistency groups are supported to enable FlashCopy across multiple VDisks.
- ▶ The target VDisk can be updated independently of the source VDisk.
- ▶ Bitmaps governing I/O redirection (I/O indirection layer) are maintained in both nodes of the SVC I/O group to prevent a single point of failure.
- ▶ It is useful for backup, improved availability, and testing.

Be aware that at the time of writing, there are some important considerations and restrictions:

- ▶ A one-to-one mapping of source and target virtual disk is required:
 - No multiple relationships
 - No cascading

- ▶ Source and target must be within the same cluster, either within or across I/O groups.
- ▶ Source and target virtual disk must be the same size.
- ▶ There is no FlashCopy incremental support.
- ▶ There is no FlashCopy Space Efficient support.
- ▶ The size of a source and target VDisk cannot be altered (increased or decreased) after the FlashCopy mapping is created.
- ▶ The maximum quantity of source VDisks per I/O group is 16 TB.

SVC Metro Mirror

Metro Mirror works by defining a Metro Mirror relationship between VDisks of equal size. To provide management (and consistency) across a number of Metro Mirror relationships, consistency groups are supported (as with FlashCopy).

The SVC provides both intracluster and intercluster Metro Mirror as described next.

Intracluster Metro Mirror

Intracluster Metro Mirror can be applied within any single I/O group.

Metro Mirror across I/O groups in the same SVC cluster is not supported, since Intracluster Metro Mirror can only be performed between VDisks in the same I/O group.

Intercluster Metro Mirror

Intercluster Metro Mirror operations requires a pair of SVC clusters that are separated by a number of moderately high bandwidth links. The two SVC clusters must be defined in an SVC partnership, which must be performed on both SVC clusters to establish a fully functional Metro Mirror partnership.

The Metro Mirror functions can be invoked via the Web-interface, CLI, TPC, or scripts.

Metro Mirror remote copy technique

Metro Mirror is a synchronous remote copy, which we briefly explain. To illustrate the differences between synchronous and asynchronous remote copy, we also explain asynchronous remote copy.

Synchronous remote copy

Metro Mirror is a fully synchronous remote copy technique, which ensures that updates are committed at both primary and secondary VDisks before the application is given completion to an update.

Figure 11-9 illustrates how a write to the master VDisk is mirrored to the cache for the auxiliary VDisk before an acknowledge of the write is sent back to the host issuing the write. This ensures that the secondary is real-time synchronized, in case it is required in a failover situation.

However, this also means that the application is fully exposed to the latency and bandwidth limitations of the communication link to the secondary site. This might lead to unacceptable application performance, particularly when placed under peak load. This is the reason for the distance limitations when applying Metro Mirror.

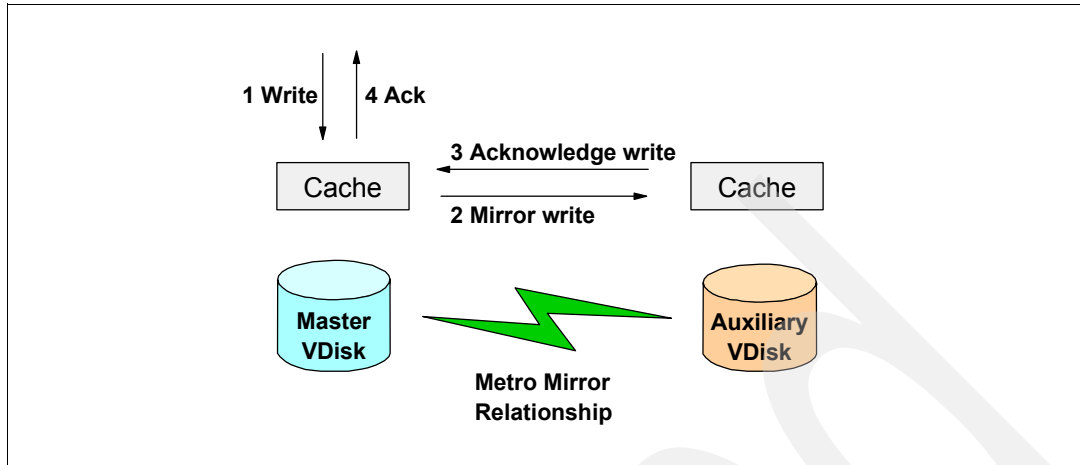


Figure 11-9 Write on VDisk in Metro Mirror relationship

Asynchronous remote copy

In an asynchronous remote copy, the application is given completion to an update when it is sent to the secondary site, but the update is not necessarily committed at the secondary site at that time. This provides the capability of performing remote copy over distances exceeding the limitations of synchronous remote copy.

Figure 11-10 illustrates that a write operation to the master VDisk is acknowledged back to the host issuing the write before it is mirrored to the cache for the auxiliary VDisk. In a failover situation, where the secondary site has to become the primary source of your data, then any applications that use this data must have their own built-in recovery mechanisms, for example, transaction log replay.

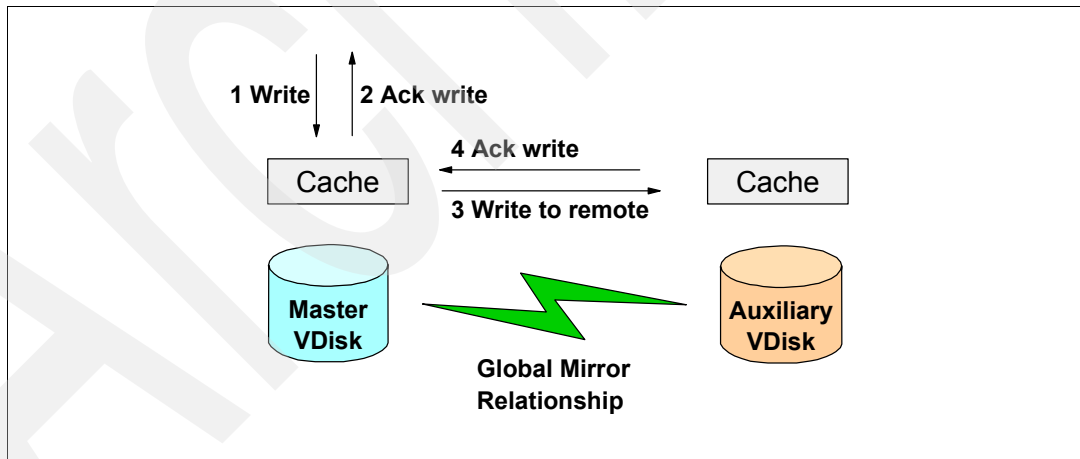


Figure 11-10 Write on VDisk in Global Mirror relationship

Overview of Metro Mirror features

The SVC Metro Mirror supports the following features:

- ▶ Synchronous remote copy of VDisks dispersed over metropolitan scale distances is supported.
- ▶ SVC implements the Metro Mirror relationship between VDisk pairs, each VDisk in the pair being managed by an SVC cluster.

- ▶ SVC supports intracluster Metro Mirror, where both VDisks belong to the same IO group within the same cluster.
- ▶ SVC supports intercluster Metro Mirror, where each VDisk belongs to their separate SVC cluster. A given SVC cluster can be configured for partnership with another cluster. A given SVC cluster can only communicate with one other cluster. All intercluster Metro Mirror takes place between the two SVC clusters in the configured partnership.
- ▶ Intercluster and intracluster Metro Mirror can be used concurrently within a cluster for different relationships.
- ▶ SVC does not require a control network or fabric to be installed to manage Metro Mirror. For intercluster Metro Mirror, SVC maintains a control link between the two clusters. This control link is used to control state and co-ordinate updates at either end. The control link is implemented on top of the same FC fabric connection as SVC uses for Metro Mirror I/O.
- ▶ SVC implements a configuration model which maintains the Metro Mirror configuration and state through major events such as failover, recovery, and resynchronization to minimize user configuration action through these events.
- ▶ SVC maintains and polices a strong concept of consistency and makes this available to guide configuration activity.
- ▶ SVC implements flexible resynchronization support enabling it to re-synchronize VDisk pairs which have suffered write I/O to both disks and to resynchronize only those regions which are known to have changed.

Be aware that at the time of writing, there are some important considerations and restrictions:

- ▶ A SAN Volume Controller cluster can only be in a relationship with one other SAN Volume Controller cluster.
- ▶ There can be a one-to-one Metro Mirror volume relationship only.
- ▶ The source and target disks must be the same size.
- ▶ Support is provided for up to 256 Consistency Groups per SVC Cluster.
- ▶ Support is provided for up to 1024 relationships per SVC Cluster.
- ▶ There can be up to 16 TB per I/O group; up to 64 TB per SAN Volume Controller cluster
- ▶ Source and target cannot be part of an existing remote copy relationship.
- ▶ Neither the source nor the target can be a target of a FlashCopy relationship.

SVC Global Mirror

Global Mirror works by defining a GM relationship between two VDisks of equal size and maintains the data consistency in an asynchronous manner. Therefore, when a host writes to a source VDisk, the data is copied from the source VDisk cache to the target VDisk cache. At the initiation of that data copy, confirmation of I/O completion is transmitted back to the host.

Note: The minimum firmware requirement for SVC GM functionality is V4.1.1. Any cluster or partner cluster not running this minimum level does *not* have GM functionality available. Even if you have a Global Mirror relationship running on a downlevel partner cluster and you only wish to use intracluster GM, the functionality is not available to you.

The SVC provides both intracluster and intercluster Global Mirror, which are described next.

Intracluster Global Mirror

Although Global Mirror is available for intracluster, it has no functional value for production use. Intracluster Global Mirror provides the same capability for less overhead. However, leaving this functionality in place simplifies testing and does allow client experimentation and testing (for example, to validate server failover on a single test cluster).

Intercluster Global Mirror

Intercluster Global Mirror operations require a pair of SVC clusters that are commonly separated by a number of moderately high bandwidth links. The two SVC clusters must each be defined in an SVC cluster partnership to establish a fully functional Global Mirror relationship.

The Global Mirror functions can be invoked via the Web-interface, CLI, TPC or scripts.

Overview of Global Mirror features

SVC Global Mirror supports the following features:

- ▶ Asynchronous remote copy of VDisks dispersed over metropolitan scale distances is supported.
- ▶ SVC implements the Global Mirror relationship between VDisk pairs, with each VDisk in the pair being managed by an SVC cluster.
- ▶ SVC supports intracluster Global Mirror, where both VDisks belong to the same IO group with the same cluster.
- ▶ SVC supports intercluster Global Mirror, where each VDisk belongs to their separate SVC cluster. A given SVC cluster can be configured for partnership with another cluster. A given SVC cluster can only communicate with one other cluster. All intercluster Global Mirror takes place between the two SVC clusters in the configured partnership.
- ▶ Intercluster and intracluster Global Mirror can be used concurrently within a cluster for different relationships.
- ▶ SVC does not require a control network or fabric to be installed to manage Global Mirror. For intercluster Global Mirror the SVC maintains a control link between the two clusters. This control link is used to control state and co-ordinate updates at either end. The control link is implemented on top of the same FC fabric connection as the SVC uses for Global Mirror I/O.
- ▶ SVC implements a configuration model which maintains the Global Mirror configuration and state through major events such as failover, recovery, and resynchronization to minimize user configuration action through these events.
- ▶ SVC maintains and polices a strong concept of consistency and makes this available to guide configuration activity.

SVC implements flexible resynchronization support enabling it to re-synchronize VDisk pairs which have suffered write I/O to both disks and to resynchronize only those regions which are known to have changed.

Global Mirror Architecture

Global Mirror extends Metro Mirror's I/O Algorithms:

- ▶ I/O complete is acknowledged to host at primary once the following actions occur:
 - Data is committed to cache in both nodes in I/O group.
 - Log of I/O and its sequence number are committed on both nodes of I/O group.
- ▶ Secondary I/O is transmitted concurrently with processing of primary I/O:
 - Local host I/O normally completes before secondary I/O completes.
- ▶ Extra messages are passed between nodes in cluster:
 - They are used to detect concurrent host I/O among VDisks in same consistency group.
- ▶ Concurrent I/O is assigned a shared "sequence number":
 - During high loads, many I/Os share a sequence number to minimize overhead.

- ▶ Secondary I/Os are applied at secondary in “sequence number” order:
 - Writes within a “sequence number” are non-dependent and written in any order once received at secondary.
 - Many sequence number “batches” can be outstanding on the long distance link.
 - At the secondary site, a single sequence number is applied at a time.
- ▶ Existing redundancy is used to continue operation if node or link fails:
 - Recovery begins by re-driving writes that were in flight.
 - Log of I/O established during initial host write used to support recovery.
 - Recovery uses read to primary VDisk to capture data for replay of write.

Dealing with “Colliding Writes” is done as follows:

- ▶ Sequence number scheme only supports a single active write for any given 512 byte sector/block.
- ▶ Later writes are delayed until processing for earlier writes is complete:
 - Any application that overwrites same sector with back-to-back writes sees synchronous, not asynchronous, performance
- ▶ This is not believed to be a problem for major applications:
 - Databases avoid over-writing same sectors in close succession
 - Might be a problem for custom applications
 - Similar, but not identical, limitation exists in DSx000 GM scheme
 - Avoid using GM for data that does not have to be mirrored (such as scratch disks)

Be aware that at the time of writing, there are some important considerations and restrictions:

- ▶ An SVC cluster can only be in a relationship with one other SVC cluster.
- ▶ There can be a one-to-one Global Mirror volume relationship only.
- ▶ The source and target disks must be the same size.
- ▶ Support is provided for up to 256 Consistency Groups per SVC Cluster.
- ▶ Support is provided for up to 1024 relationships per SVC Cluster.
- ▶ There can be up to 16 TB per I/O group; up to 64 TB per SVC cluster
- ▶ Source and target cannot be part of an existing remote copy relationship.
- ▶ Neither the source nor the target can be a target of a FlashCopy relationship. Disaster recovery considerations with SVC.

SAN Volume Controller Data Migration

Data migration is the process of moving data from one storage volume to another. This might be required because you want to move less valuable data to cheaper storage, or perhaps a new faster storage system has been purchased, and you want to move some critical data onto it. In non-virtualized environments, this is a disruptive, time-consuming operation, requiring application downtime. Some storage systems do not support a direct remote copy function — which forces host-based volume drain, back up to tapes or restore from tapes during the migration. In addition data is unavailable during the migration process.

The SVC allows the mapping of Virtual Disk (VDisk) extents to Managed Disk (MDisk) extents to be changed, without interrupting host access to the VDisk. This functionality is utilized when performing VDisk migrations, and can be performed for any VDisk defined on the SVC. This functionality can be used for:

- ▶ Redistribution of VDIs, and thereby the workload within an SVC cluster across back-end storage:
 - Moving workload onto newly installed storage.
 - Moving workload off old/failing storage, ahead of decommissioning it.
 - Moving workload to re-balance a changed workload.

- ▶ Migrating data from older back-end storage to SVC managed storage.
- ▶ Migrating data from one back-end controller to another using the SVC as a data block mover and afterwards removing the SVC from the SAN.
- ▶ Migrating data from Managed Mode back into Image mode prior to removing the SVC from a SAN.

Figure 11-11 shows the benefit of using SVC to do data migration.

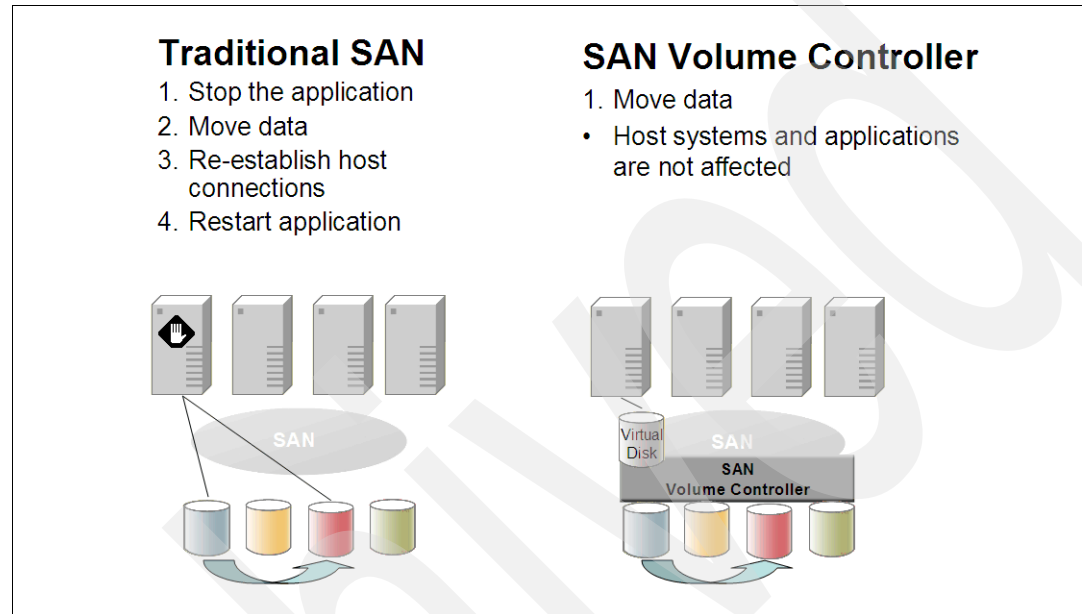


Figure 11-11 Data Migration by SVC versus by traditional method

For more information about SVC, refer to the IBM Redbook, *IBM System Storage SAN Volume Controller*, SG24-6423.

11.2.4 Business Continuity considerations with SAN Volume Controller

Basically, the advanced functions of the SVC are used as the DS6000/8000 advanced functions and the same tier levels can be achieved as with the DS6000/8000 (for example, Synchronous Metro Mirror is Tier 6 for DS6000/8000 and SVC).

Similar scenarios such as database split mirror, shadow database/log mirroring plus forward recovery, as described in the chapter, "Databases and applications: high availability options" in the *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547, can be implemented on the SVC as well as on the DS6000/8000. The benefit of the SVC over a storage system such as DS6000/8000 is that the SVC copy functions operate on a higher level. A much wider variety of combinations can therefore be considered, especially with the use of different storage subsystems.

11.3 IBM System Storage N series Gateway

The IBM System Storage N series Gateway is a network-based virtualization solution that virtualizes tiered, heterogeneous storage arrays and enables clients to leverage the dynamic virtualization capabilities of Data ONTAP software across a broad set of high-end and modular storage arrays from Hitachi, HP, IBM, Engenio, StorageTek™, and Sun.

IBM System Storage N series Gateway supports selected platforms. For current supporting information, refer to the IBM Web site:

<http://www.ibm.com/servers/storage/nas/interophome.html>

The IBM System Storage N series Gateway provides proven and innovative data management capabilities for sharing, consolidating, protecting, and recovering data for applications and users and can integrate into SAN infrastructures in some environments. These innovative data management capabilities, when deployed with disparate storage systems, simplify heterogeneous storage management.

The IBM System Storage N series Gateway presents shares, exports, or LUNs built on flexible volumes that reside on aggregates. The IBM System Storage N series Gateway is also a host on the storage array SAN. Disks are not shipped with the N series Gateway. N series Gateways take storage array LUNs (which are treated as disks) and virtualize them through Data ONTAP, presenting a unified management interface.

- ▶ An IBM System Storage N series Gateway is similar to IBM N series storage system in most dimensions
 - IBM N series storage systems use disk storage provided by IBM only, the Gateway models support heterogeneous storage.
 - Data ONTAP is enhanced to enable the Gateway series solution.
- ▶ A RAID array provides LUNs to the Gateway:
 - Each LUN is equivalent to an IBM disk.
 - LUNs are assembled into aggregates/volumes, then formatted with WAFL file system, just like the IBM N series storage systems.

The IBM System Storage N series Gateway, an evolution of the N series product line, is a network-based virtualization solution that virtualizes tiered, heterogeneous storage arrays, allowing clients to leverage the dynamic virtualization capabilities available in Data ONTAP across multiple tiers of IBM and 3rd party storage. Like all N series storage systems, the N series Gateway family is based on the industry-hardened Data ONTAP microkernel operating system, which unifies block and file storage networking paradigms under a common architecture and brings a complete suite of N series advanced data management capabilities for consolidating, protecting, and recovering data for applications and users.

Organizations that are looking for ways to leverage SAN-attached storage to create a consolidated storage environment for the various classes of applications and storage requirements throughout their enterprise. They might be looking for ways to increase utilization, simplify management, improve consolidation, enhance data protection, enable rapid recovery, increase business agility, deploy heterogeneous storage services and broaden centralized storage usage by provisioning SAN capacity for business solutions requiring NAS, SAN or IP SAN data access.

These organizations have:

- ▶ Significant investments or a desire to invest in a SAN architecture
- ▶ Excess capacity and an attractive storage cost for SAN capacity expansion
- ▶ Increasing requirements for both block (FCP, iSCSI) and file (NFS, CIFS, etc.) access
- ▶ Increasing local and remote shared file services and file access workloads.

They are seeking solutions to cost effectively increase utilization; consolidate distributed storage, Direct Access Storage and file services to SAN storage; simplify storage management; and improve storage management business practices.



Figure 11-12 Heterogeneous storage

IBM N series Gateway benefits

IBM System Storage N series Gateway provides a number of key features that enhance the value and reduce the management costs of utilizing a SAN. An N series Gateway provides the following benefits:

- ▶ Simplifies storage provisioning and management
- ▶ Lowers storage management and operating costs
- ▶ Increases storage utilization
- ▶ Provides comprehensive simple-to-use data protection solutions
- ▶ Improves business practices and operational efficiency
- ▶ Transforms conventional storage systems into a better managed storage pool — see Figure 11-13.

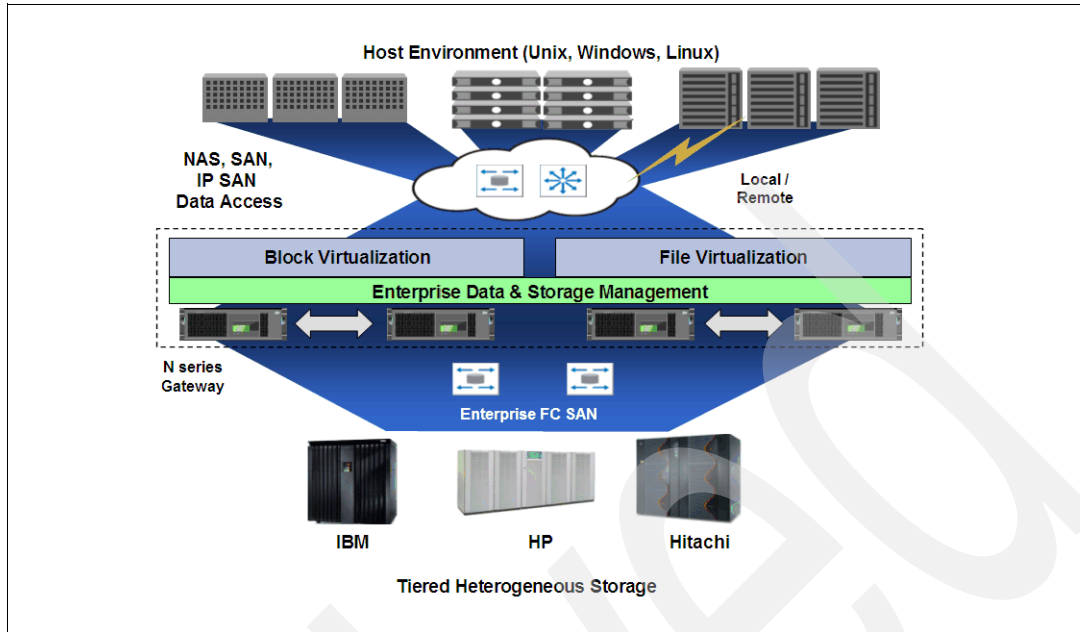


Figure 11-13 Tiered heterogeneous storage

N series Gateway hardware overview

There are 2 models of N series Gateway: N5000 Series and N7000 Series. These are shown in Table 11-1 and Table 11-2.

Table 11-1 N5000 gateway specifications

Appliance disk specifications	N5200	N5200	N5500	N5500
IBM machine types - models	2864-G10	2864-G20	2865-G10	2865-G20
Max. raw storage capacity	50 TB	50 TB	84 TB	84 TB
Max. number of LUNs on back-end disk storage array	168	168	336	336
Maximum LUN size on back-end disk storage array	500 GB	500 GB	500 GB	500 GB
Max. volume size: 1 TB=1,048,576,000,000 bytes.	16 TB	16 TB	16 TB	16 TB

Table 11-2 N7000 gateway specifications

Appliance disk specifications	N7600	N7600	N7800	N7800
IBM machine types - models	2866-G10	2866-G20	2867-G10	2867-G20
Max. raw storage capacity	420 TB	420 TB	504 TB	504 TB
Max. number of LUNs on back-end disk storage array	840	840	1008	1008f

Appliance disk specifications	N7600	N7600	N7800	N7800
Maximum LUN size on back-end disk storage array	500 GB	500 GB	500 GB	500 GB
Max. volume size: 1 TB=1,048,576,000,000 bytes.	16 TB	16 TB	16 TB	16 TB

N series software overview

The operating system for the IBM System Storage N series products is the Data ONTAP software. It is a highly optimized, scalable and flexible operating system that can handle heterogeneous environments. It integrates into UNIX, Windows, and Web environments.

Data ONTAP software includes the following standard base system features:

- ▶ Data ONTAP operating system software
- ▶ iSCSI SAN protocol support
- ▶ FlexVol
- ▶ Double parity RAID (RAID-DP)
- ▶ RAID4
- ▶ FTP file access protocol support
- ▶ SnapShot
- ▶ FilerView
- ▶ SecureAdmin
- ▶ Disk Sanitization
- ▶ iSCSI Host Attach Kit for AIX
- ▶ iSCSI Host Attach Kit for Windows
- ▶ iSCSI Host Attach Kit for Linux

Note: RAID-DP and RAID4 is not a software feature of N series Gateway, since the Gateway only uses external disk systems

The following protocols for the N series are available as extra charge features:

- ▶ CIFS - Provides File System access for Windows environments over an IP network.
- ▶ NFS - Provides File System access for UNIX and Linux environments over an IP network.
- ▶ HTTP - Allows a user to transfer displayable Web pages and related files.
- ▶ FCP - Allows transfer of data between storage and servers in block I/O formats utilizing FCP protocols across a Fibre Channel SAN.

The following software products are available as extra charge features:

- ▶ Cluster Failover
- ▶ FlexClone
- ▶ MultiStore
- ▶ SnapLock Compliance
- ▶ SnapLock Enterprise
- ▶ SnapMirror
- ▶ SnapMover
- ▶ SnapRestore
- ▶ SnapVault
- ▶ Open Systems SnapVault
- ▶ LockVault
- ▶ SnapDrive for Windows
- ▶ SnapDrive for UNIX: AIX, Solaris, HP-UX, Linux

- ▶ SnapValidator
- ▶ SyncMirror
- ▶ SnapDrive
- ▶ SnapManager for SQL
- ▶ SnapManager for Exchange
- ▶ Single Mailbox Recovery for Exchange
- ▶ Data Fabric Manager

Notes:

- ▶ LockVault Compliance is not supported by N series Gateway, while LockVault Enterprise is supported.
- ▶ SnapVault Compliance is not supported by N series Gateway, while SnapVault Enterprise is supported.
- ▶ Metro Cluster is not supported by N series Gateway.
- ▶ Nearline™ Bundle is not supported by N series Gateway.
- ▶ RAID4/RAID-DP is not supported by N series Gateway, because the gateway only uses external disk systems.

For more information, see Chapter 9, “IBM System Storage N series” on page 325.

11.4 Volume Managers

This section describes Volume Managers as another means of storage virtualization.

11.4.1 Overview

Volume managers are an example of symmetrical storage virtualization on the server level.

The Volume Manager can be additional software or can be built into the operating system (such as IBM AIX) running on the server level. It controls the I/Os on this level and can replicate them to mirror the data to two different storage subsystems. This compares to RAID-1 functionality on a file level. The file itself is stored on two independent storage subsystems. Each of the storage systems provides the LUN where this file is stored, RAID protection, and is protected by additional HA functionality, depending on the storage system. A combination of Server Clustering and Volume Manager is a good choice. As the data mirroring with a Volume Manager is synchronous, this is a valuable implementation in local or intermediate dispersed solutions. The main advantage is that the operating system level keeps control of the I/Os and the mirror. See Figure 11-14.

Examples include the AIX logical volume manager, which supports AIX only, and VERITAS Volume manager, which supports various operating systems.

The general advantages to using Volume Managers for storage virtualization include:

- ▶ Support of heterogeneous storage hardware
- ▶ Data replication functionality on the server level
- ▶ Recovery functionality on the server level

The general disadvantages include:

- ▶ Additional complexity
- ▶ A variety of dependencies (hardware components such as servers, HBAs, switches, storage and software levels of the operating system)

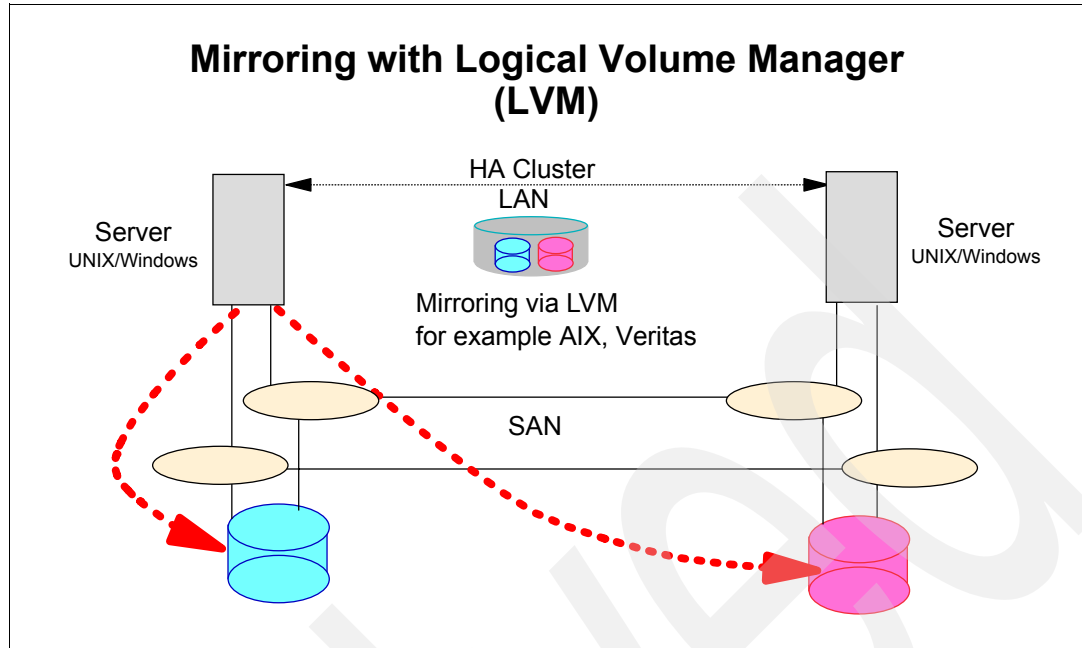


Figure 11-14 Mirroring with a Logical Volume Manager

Distance

Without server clustering the distance for mirroring is dependent on the SAN distance; the standard distance is 10km.

Using server clustering, the distance depends on the type of cluster, which might be much less than 10 km or even more. If the cluster supports more than 10km, then additional actions such as an RPQ might be necessary for the SAN part.

Tier levels

Synchronous data mirroring fits into BC Tier level 6. In combination with server clustering you can reach BC Tier level 7, because server clustering can add sophisticated automation for take-over.

11.5 Enterprise Removable Media Manager

This section covers Enterprise Removable Media Manager (eRMM), which is an IBM service offering for advanced tape management.

Enterprise Removable Media Manager provides features known from the mainframe's DFSMSrmm™ for open systems. It complements the IBM Open Storage Software Family to provide storage virtualization and advanced storage management for removable media. eRMM automatically configures drives for IBM Tivoli Storage Manager and it gathers audit trails and statistical data for the complete cartridge life cycle.

11.5.1 Introduction to Enterprise Removable Media Manager

Management for removable media is one of the biggest challenges for today's heterogeneous open systems tape environments.

The following issues are widespread with tapes and tape media:

- ▶ Tape resources are statically linked to applications.
- ▶ Resource sharing in heterogeneous and even in large homogeneous configurations is very limited.
- ▶ Adding or changing tape hardware requires changes to every application.
- ▶ Cartridge management has to be done by each application.
- ▶ There is a multitude of media changer and management interfaces (SCSI, IBM TS3500, IBM 3494, STK ACSLS).
- ▶ There is a lack of centralized management and monitoring.

eRMM provides a virtualization layer between applications like Tivoli Storage Manager and the tape library hardware. Essentially, eRMM decouples tape resources from applications which simplifies the sharing of tape resources even in large heterogeneous environments.

eRMM provides these benefits:

- ▶ Decouples tape resources and applications
- ▶ Simplifies the sharing of resources even in large heterogeneous configurations
- ▶ Allows you to change the hardware without changing all applications
- ▶ Provides policy-based cartridge management
- ▶ Virtualizes the media changer interface (IEEE 1244 or IBM 3494)
- ▶ Provides centralized management and monitoring

Figure 11-15 shows the virtualization layer between applications and tape library hardware in an eRMM tape environments.

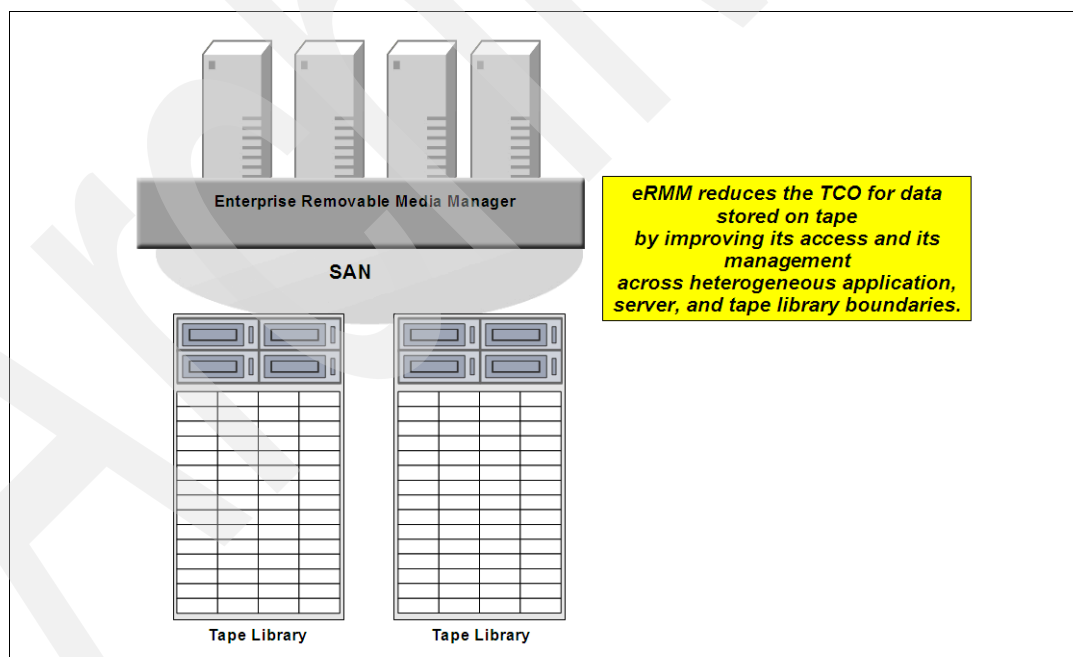


Figure 11-15 eRMM tape library hardware virtualization

11.5.2 eRMM central reporting and monitoring

eRMM enables central reporting and monitoring across heterogeneous application and tape library boundaries.

eRMM automatically detects which drive is available on which server and by which device handle (for example, \\.\Tape1, /dev/rmt2). Thus eRMM can provision the device handles to applications, eliminating the requirement to define all device handles by each application.

eRMM checks a device handle before it provisions it to an application. Thus eRMM detects and reports centrally broken paths from the application server to the tape drives. eRMM helps the administrator to answer questions such as: Is there a single path from one server to one drive broken? Are there multiple paths between the same server and multiple drives broken? Are there multiple paths between multiple server and the same drive broken? Are there many paths between multiple server and multiple drives broken?

eRMM collects historical data on cartridge mounts including throughput and errors. This is helpful for the administrator for answering questions such as: What is the best window to schedule an additional backup task? Which cartridge was mounted in which drive, and by which application? The historical data enables proactive management, for example, by identifying servers, HBAs, drives, and cartridges with downgraded performance.

Figure 11-16 shows an example of a report for drive utilization. This report shows long-term trends for drive utilization helping to plan purchases of additional drives to satisfy an increasing tape workload.

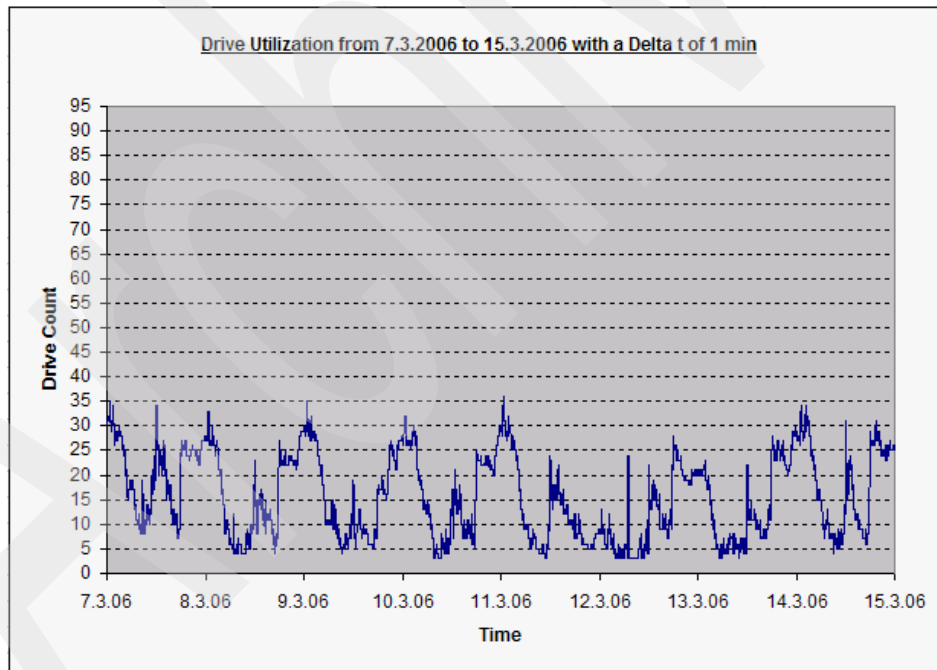


Figure 11-16 eRMM drive utilization

Figure 11-17 shows a drive histogram. This reports helps to determine whether additional backup jobs can be added to the existing tape infrastructure, and it helps to identify time frames for scheduling these new backup jobs.

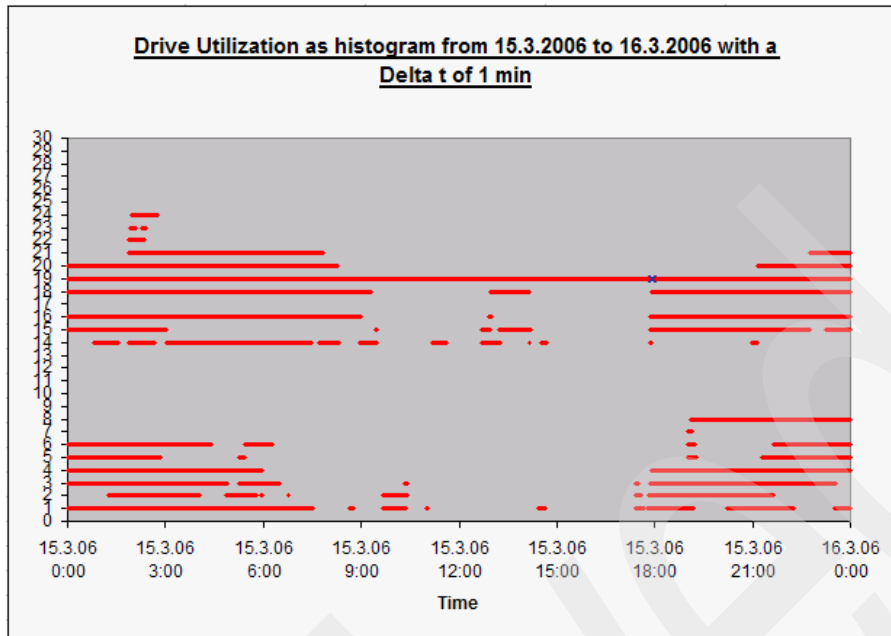


Figure 11-17 eRMM drive histogram

11.5.3 eRMM logical components and functions

This section describes the names and terms that are used in eRMM environments, as well as the key logical components and functions of eRMM.

DriveGroups and DriveGroupApplications

A DriveGroup object is a named group to which a drive can belong. DriveGroups are used to aggregate drives, and then to support both an access permissions model and a preferential usage policy. Each drive must belong to a DriveGroup.

DriveGroupApplication objects are used to allow applications to access a particular DriveGroup. A DriveGroup can span across Libraries.

CartridgeGroups and CartridgeGroupApplications

CartridgeGroups are logical collections of cartridges. They are used to control an application's access to cartridges. A single CartridgeGroup can contain cartridges from more than one library.

CartridgeGroupApplication objects are used to allow applications to access particular CartridgeGroups.

ScratchPools

ScratchPools are a special kind of CartridgeGroup. They are searched first for empty Cartridges before any other CartridgeGroup.

If a Volume is allocated on a Cartridge which belongs to a ScratchPool, the Cartridge is moved to another ordinary CartridgeGroup to which the application issuing the command also has access. If the application does not have access to another ordinary CartridgeGroup, the Cartridge is not moved out of the ScratchPool, but is set to the "not allocateable" state, which protects it from usage by another application.

In order to see how many Cartridges are currently available, we recommend that you create at least one ScratchPool and another ordinary CartridgeGroup so that the ScratchPool only contains empty cartridges. By default the eRMM installation creates a ScratchPool and an ordinary CartridgeGroup.

Figure 11-18 shows an example of DriveGroups, CartridgeGroups, and ScratchPools.

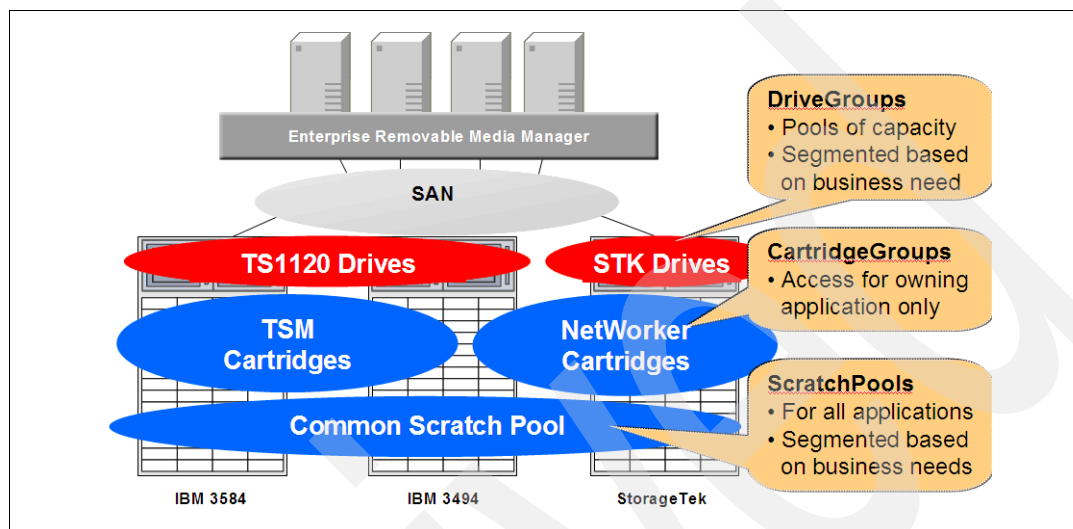


Figure 11-18 eRMM Groups & Pools

Media Manager (MM)

The Media Manager (MM) is the central “server” component which, among other tasks, coordinates access to drives and cartridges, handles volume allocation and deallocation requests, and stores a log of all activities. MM uses IBM DB2 for persistent storage.

Library Manager (LM)

The Library Manager (LM) provides MM access to library media changers. It reports all slots, tapes, and cartridges to the media manager, controls libraries on behalf of the media manager, and encapsulates (virtualizes) the library hardware. This allows you to integrate new library hardware without any changes to an already installed eRMM Media Manager.

Host Drive Manager (HDM)

The Host Drive Manager (HDM) reports all local device handles to MM, handles mount and unmount commands, checks the path when a cartridge is loaded, and reports statistical data to MM when a cartridge is unloaded.

Admin Console

The Admin Console offers a Command Line Interface (CLI) and a Web Graphical User Interface (WebGUI), which enable configuration and administration of eRMM.

External Library Manager (ELM)

The External Library Manager (ELM) Adapter for IBM Tivoli Storage Manager enables Tivoli Storage Manager to utilize eRMM for media management purposes.

Figure 11-19 shows the logical architecture of eRMM with the components Media Manager, Library Manager, Host Drive Manager, and Admin Console.

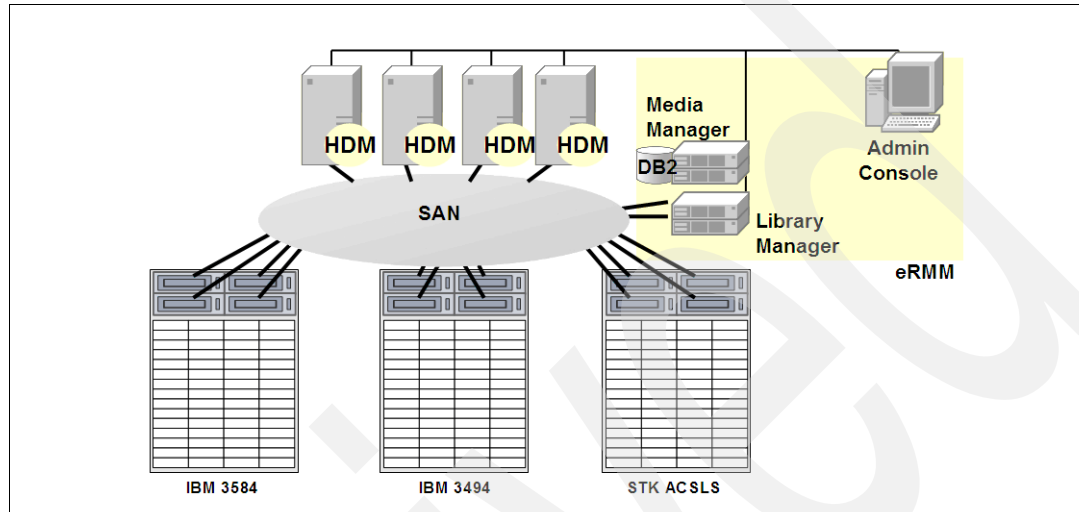


Figure 11-19 eRMM logical architecture

All eRMM logical components can run on the same or on different servers, therefore different options are available for scaling and high availability.

Application servers which require access to eRMM managed resources (libraries, drives, cartridges) must run the HDM. In addition, Tivoli Storage Manager servers require the ELM for eRMM.

11.5.4 eRMM control flow

The following three figures (Figure 11-20, Figure 11-21, and Figure 11-22) show the control flow for a tape mount in an eRMM environment. The workflow for Tivoli Storage Manager Storage Agents is similar to the workflow for Tivoli Storage Manager server.

Step 1: The Tivoli Storage Manager server wants to mount a new scratch volume of type 3592. The Tivoli Storage Manager server starts the ELM and sends it the mount request via the Tivoli Storage Manager External Library Manager Interface (see Tivoli Storage Manager Administrator's Guide, "Appendix B. External Media Management Interface Description" for further details).

Step 2: The eRMM ELM forwards the mount request via TCP/IP to the eRMM MM.

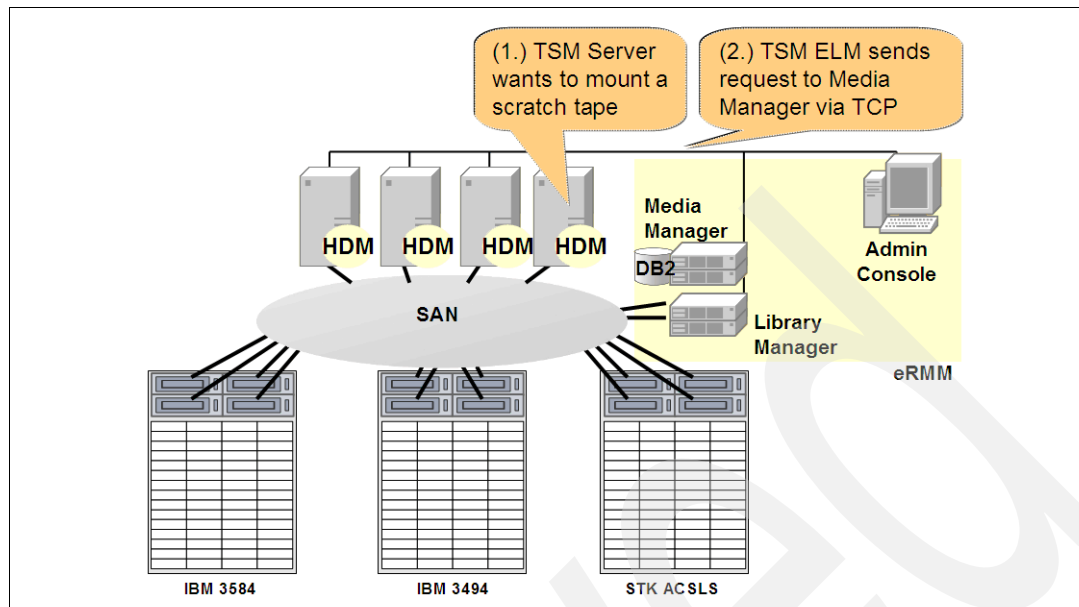


Figure 11-20 eRMM Control Flow - Steps 1 to 2

Step 3: The eRMM MM queries its database for idle drives and scratch cartridges and selects a drive and a cartridge according to the access rules. eRMM takes into account which drives are configured in the operating system of the requesting Tivoli Storage Manager server.

Step 4: The eRMM MM forwards the specific mount request via TCP/IP to the respective Host Drive Manager (running on the same server as the Tivoli Storage Manager server) and to the eRMM LM.

Step 5: The eRMM LM converts the mount request into a library specific command (SCSI for IBM 3584, Imcp for IBM 3494, ACSLS for STK) and loads the cartridge into the drive.

Step 6: The eRMM HDM queries the drive to ensure that the path to the drive is healthy.

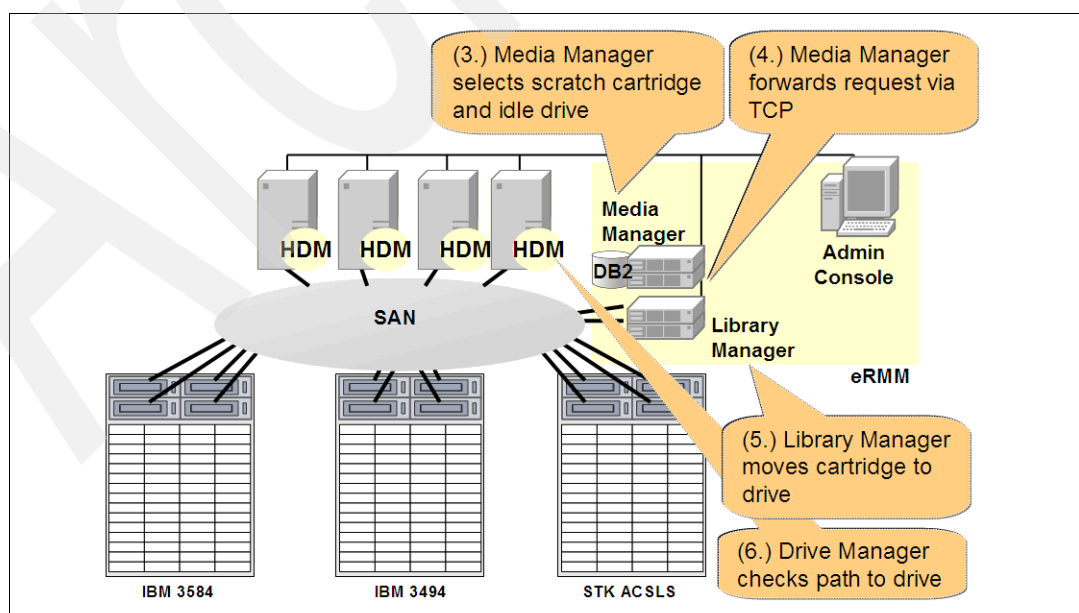


Figure 11-21 eRMM Control Flow - Steps 3 to 6

Step 7: The eRMM MM updates the status in the eRMM DB.

Step 8: The eRMM MM sends the device handle (for example, /dev/rmt5, \\.\Tape3) and the cartridge VOLSER via TCP to the eRMM ELM. The eRMM ELM returns the Device Handle and the VOLSER to the Tivoli Storage Manager server via the Tivoli Storage Manager External Library Manager Interface.

Step 9: The Tivoli Storage Manager server updates its volume inventory and directly accesses the drive. eRMM is not involved for read and write operations. This design allows to put eRMM into or pull eRMM out of an existing Tivoli Storage Manager environment without recopying the data on tape.

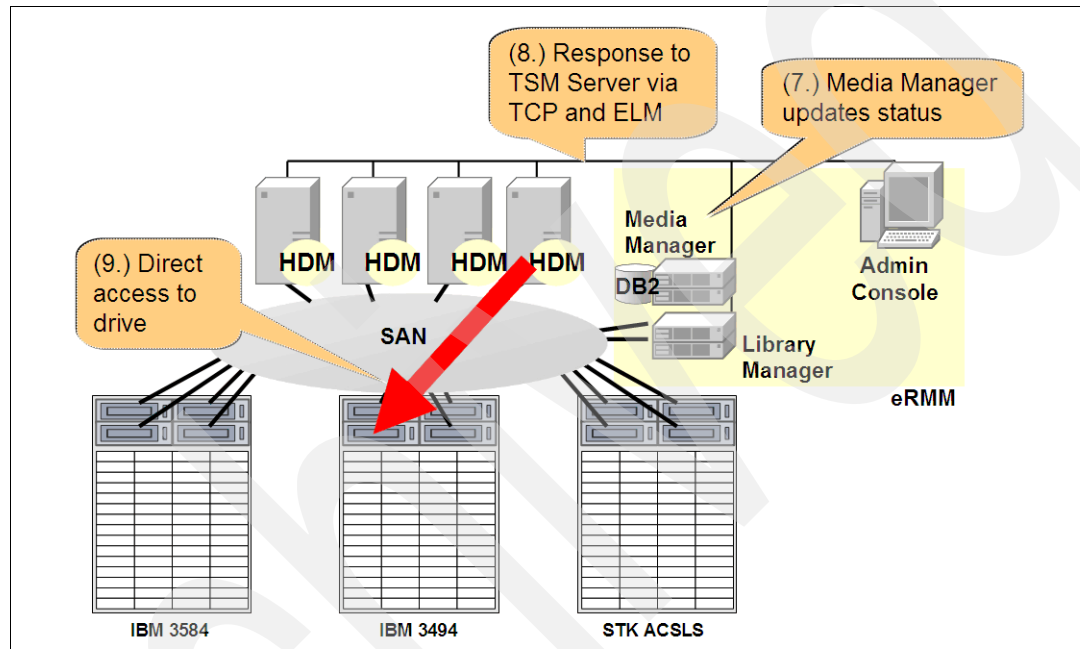


Figure 11-22 eRMM Control Flow - Steps 7 to 9

This example illustrates how Tivoli Storage Manager installations with Library sharing and LAN-free Backup can benefit from eRMM's provisioning of tape resources. It is no longer required to configure and maintain tape paths in Tivoli Storage Manager.

11.5.5 Supported features (eRMM 1.2.4)

The following features are supported by eRMM at the current release - eRMM v1.2.4:

- ▶ Automatic configuration of drives for Tivoli Storage Manager
- ▶ Centralized access control, administration, problem determination and reporting
- ▶ Policy-based drive and cartridge allocation
- ▶ Dynamic and heterogeneous drive sharing
- ▶ Dynamic and heterogeneous drive and cartridge pooling
- ▶ Mount request queuing
- ▶ Audit trails and statistical data for the complete cartridge life cycle
- ▶ Dynamic drive sharing
- ▶ Dynamic drive and cartridge pooling including common scratch pool management
- ▶ Tape library virtualization

The following features are considered to advance eRMM's value proposition:

- ▶ Policy based cartridge and life cycle management
- ▶ Offsite media management and tracking (vaulting)
- ▶ Advanced reporting and auditing

11.5.6 Supported platforms (eRMM 1.2.4)

The following platforms are supported by eRMM.

Applications

The following applications are supported:

- ▶ Tivoli Storage Manager on AIX, Solaris, HP-UX, Linux (Intel & System z) and Windows
- ▶ Native OS commands like tar, dd, and mksysb
- ▶ EMC Legato NetWorker (on request)

Tape libraries

The following tape libraries are supported:

- ▶ IBM System Storage TS3500 Tape Library Open Systems
- ▶ IBM TotalStorage 3494 Tape Library Open Systems
- ▶ ACSLS managed StorageTek libraries
- ▶ Other SCSI Libraries (for example, IBM 3583, ADIC on request)

Support for additional applications and tape libraries are under consideration as enhancements to eRMM.

11.5.7 Strategic fit and positioning

How does eRMM fit into the strategic framework of the business solution areas Infrastructure Simplification, Business Continuity, and Information Life Cycle Management?

eRMM and infrastructure simplification

This includes the following benefits:

- ▶ Automated detection and configuration of drives for Tivoli Storage Manager
- ▶ Provisioning and simplified sharing of drives and cartridges to Tivoli Storage Manager
- ▶ Policy based drive and cartridge utilization
- ▶ Policy based cartridge management
- ▶ Virtualization of the libraries' media changer interface

eRMM and Business Continuity

This includes the following benefits:

- ▶ High available library sharing for Tivoli Storage Manager on AIX, Solaris, HP-UX, Linux, and Windows
- ▶ High available provisioning of scratch cartridges to Tivoli Storage Manager

eRMM and ILM

This includes the following benefits:

- ▶ Policy based vaulting for tired storage management (considered for future releases of eRMM)
- ▶ Audit trails and statistical data of complete cartridge life cycle for regulatory compliance and for tape quality management (considered for future releases of eRMM)

Storage management software

In this chapter we provide a very high level view of the IBM comprehensive suite of storage management software solutions that complement and enable the disaster recovery solution. This series of products has been engineered to provide an end-to-end solution for an enterprise's ever-increasing demand for more data.

It is well known that data volumes are growing rapidly — both the amount and types of data being stored. The rate of change is overwhelming for most environments and the result is higher costs and more human error.

Regardless of format, every application's data must be acquired, stored, transferred, backed up, archived, and secured, secured from internal threats, external threats, man-made threats, and natural disasters.

In this chapter we describe a range of IBM Storage Management solutions that can provide your organization with multiple storage management benefits that provide a “defense in depth.”

12.1 Storage management strategy

Data growth and increasing management complexity continue unabated along parallel tracks for the foreseeable future. However, in some organizations, storage management is executed without a comprehensive strategy, thereby reducing cost-effectiveness, efficiency, and increasing the risk of data loss, corruption, or compromise.

A strategic approach to the management of data resources, utilizing the most appropriate software tools for *your organization's* environment, ensures not only that the data is secured, but also that the entire information grid can be recreated according to business priority in case of disaster. We want to emphasize *your organization* because it is the operative phrase to keep in mind as we describe the array of IBM storage software solutions in this chapter. When addressing the requirements of a storage management strategy, one size definitely does not fit all. However, no matter how dauntingly complex your storage environment has become, we are confident that a tailored solution can be implemented that exactly fits your organization's data availability and recovery requirements.

By implementing a strategy of software-based storage management solutions, we want to show how you can:

- ▶ Not only provide a disaster recovery solution, but reduce the likelihood of such occurrences as well.
- ▶ Meet or exceed your service level objectives for data availability.
- ▶ Measure a positive return on your storage dollar investment.

12.2 Key storage management areas

There are eight key management areas that must be addressed by comprehensive storage resource management solutions. Each of these management domains has implications for our discussion of Business Continuity support:

- ▶ **Asset management:** Identifies storage resources by parameters that include type/model/serial number, features, location, acquisition date, price, and maintenance information.

An accurate inventory of your computing assets is absolutely essential for purposes of data center reconstruction, should that become necessary. Also, if your organization's recovery plan calls for utilization of hot site, cold site, or mobile recovery facilities, an accurate and complete account of your storage infrastructure is essential.

- ▶ **Configuration management:** Views and modifies hardware configurations for host adapters, caches, and storage devices at the licensed internal code, operating system, and software application level.

Accurate schematic diagrams and narrative descriptions of your networked environment and storage infrastructure are crucially important in assuring recoverability. Like asset management accuracy, efficient configuration management assures the success of any off-site recovery.

- ▶ **Capacity management:** Views common capacity metrics (total capacity, utilization percentage and trends, fragmentation, and efficiency) across a variety of storage subsystems.

Accurate forecasting of storage requirements avoids costly and time-consuming *out of space* conditions, and manages your storage assets to avoid being in a continuous crisis mode.

- ▶ **Data/device/media migration:** Simplifies the movement of data to new device technologies and performs outboard data movement and replication within a storage subsystem.
A well-defined policy of data migration also has disaster recovery/availability implications. Such a policy assures that files necessary for recovery are appropriately positioned within the enterprise, and out of space conditions mitigated.
- ▶ **Event/alert management:** Monitors events, receives notification of exception conditions (such as thresholds), initiates exception handlers, and logs information.
Such warning of impending or actual equipment failure assists operations staff in employing appropriate contingency plans based on predefined organizational policies.
- ▶ **Performance management:** Monitors current and historical performance information such as path/device/cache utilization and I/O operations per second.
An effective real-time system performance monitoring capability assures a prompt response to system degradation. Trend analysis of isolation can mitigate developing bottlenecks.
- ▶ **Centralization of policies:** Ensures that enterprise-wide storage management policies are consistently followed. Propagates policies to all platforms and locations from a central repository.
Establishment of a consistent way of doing things ensures that response to system emergencies is measured and appropriate to the circumstance. When such policies are automated, mistakes or misinterpretation by operations staff is minimized.
- ▶ **Removable media management:** Tracks tape and optical media, and shares tape and optical devices among applications.
Effective management of tape or other like assets, for most organizations, is the first line of defense when confronted with any sort of unplanned outage. Policy-driven vaulting management provides additional assurances that your recovery strategy is viable.

12.3 Storage management software solutions

In the remainder of this chapter we provide a high-level view of some of the storage management products available from IBM. We do not usually think of some of these products as being related to an organization's disaster recovery effort. However, all of the IBM Storage Management products can either mitigate (that is, lessen the possibility of a data disaster) or at least lessen a disaster's adverse effects on your organization. Management of your organization's storage assets is the first step in prevention of disasters, and provides the means of recovery. The following storage products are available:

- ▶ IBM TotalStorage Productivity Center is a suite of infrastructure management software to centralize, automate, and simplify the management of complex and heterogeneous storage environments.
- ▶ IBM Tivoli Storage Manager provides a myriad of storage management solutions for disparate platforms, storage devices, and communication protocols. For our purposes we are going to focus on the capabilities of the Disaster Recovery Manager, Continuous Data Protection, and Bare Machine Recovery.
- ▶ Data Facility Systems Managed Storage (DFSMS) is a family of products that address every aspect of data management in the z/OS world, including hierarchical storage management, backup and restore, disaster recovery planning, tape management, and many others including the Mainstar Software suite of disaster recovery and data management products.

- ▶ The CICS/VSAM Recovery (CICSVR) is a feature whose forward-recovery capability can recover lost or damaged VSAM data sets.
- ▶ DFSMS Transactional VSAM Services (DFSMSStvs) is an optional z/OS feature that enables batch jobs and CICS online transactions to update shared VSAM data sets concurrently. Multiple batch jobs and online transactions can be run against the same VSAM data sets, and DFSMSStvs helps ensure data integrity for concurrent batch updates while CICS ensures it for online updates.
- ▶ The IBM TotalStorage Expert Enterprise Tape Library (ETL) Expert can also be helpful.
- ▶ The IBM TotalStorage Specialist family addresses management requirements for the TS7700, VTS, and PtP VTS.

12.4 IBM TotalStorage Productivity Center

The IBM TotalStorage Productivity Center is an open storage infrastructure management solution designed to help reduce the effort of managing complex storage infrastructures, to help improve storage capacity utilization, and to help increase administrative efficiency. It is designed to enable the storage infrastructure to have the ability to respond to *on demand* storage requirements.

IBM TotalStorage Productivity Center (TPC) consists of the following four products:

- ▶ **IBM TotalStorage Productivity Center for Data:** This product provides over 300 enterprise-wide reports, monitoring and alerts, policy-based action, and file system capacity automation in the heterogeneous environment.
- ▶ **IBM TotalStorage Productivity Center for Fabric:** This product provides automated device discovery, topology rendering, zone control, real-time monitoring and alerts, and event management for heterogeneous enterprise SAN environments.
- ▶ **IBM TotalStorage Productivity Center for Disk:** This product enables device configuration and management of supported SAN-attached devices from a single console. It can discover storage and provides configuration capabilities. Through the use of the data collection, setting of thresholds and use of performance reports, performance can be monitored for the DS4000, DS6000, DS8000, SVC, ESS, and any other storage system that supports the SMI-S block server performance subprofile.
- ▶ **IBM TotalStorage Productivity Center for Replication:** This product provides configuration and management of the Point-in-Time copy (FlashCopy), Metro Mirror (synchronous point-to-point remote copy) and Global Mirror (asynchronous point-to-point remote copy) capabilities of the DS8000, DS6000, SVC, and ESS

12.4.1 IBM TotalStorage Productivity Center for Data

IBM TotalStorage Productivity Center for Data (TPC for Data) helps discover, monitor, and create enterprise policies for disks, storage volumes, file systems, files, and databases. TPC for Data improves application availability by giving early warnings when file systems are running out of space and optionally automatically extending file systems. In today's server consolidation environments, TPC for Data helps efficiently utilize the accumulated storage resources. Architected for efficiency and ease-of-use, TPC for Data uses a single agent per server to provide detailed information without high consumption of network bandwidth or CPU cycles.

Figure 12-1 shows the concept of storage resource management from a life cycle perspective.

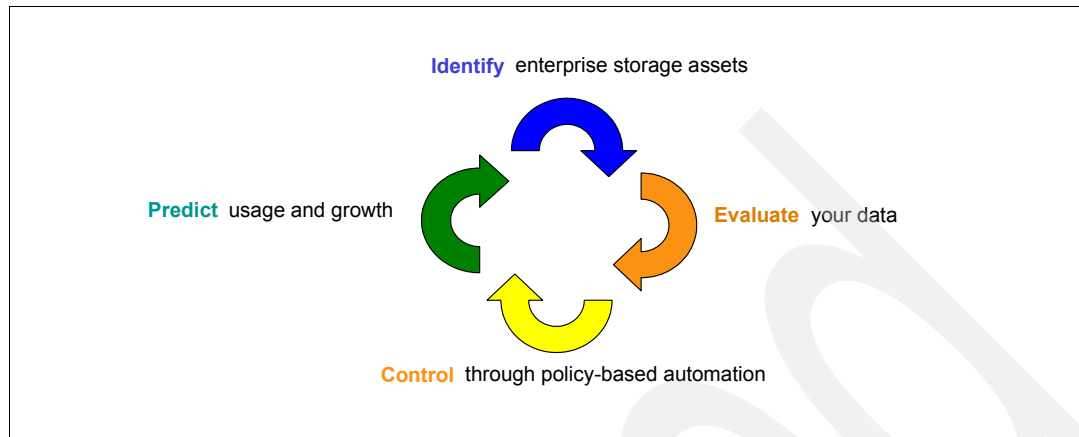


Figure 12-1 Storage life cycle resource management

The idea is to establish a base understanding of the storage environment, focussing on discovering areas where simple actions can deliver rapid return on investment. The steps required are to identify potential areas of exposure, evaluate the data currently stored on the servers, set up control mechanisms for autonomic management, and continue the capacity planning process by predicting growth.

TPC for Data monitors storage assets, capacity, and usage across an enterprise. TPC for Data can look at:

- ▶ Storage from a host perspective: Manage all the host-attached storage, capacity and consumption attributed to file systems, users, directories, and files, as well as the view of the host-attached storage from the storage subsystem perspective.
- ▶ Storage from an application perspective: Monitor and manage the storage activity inside different database entities including instance, tablespace, and table.
- ▶ Storage utilization: Provide chargeback information allowing for the ability to be able to justify, or account for, storage usage.

TPC for Data provides over 300 standardized reports (and the ability to customize reports) about file systems, databases, and storage infrastructure. These reports provide the storage administrator with information about:

- ▶ Assets
- ▶ Availability
- ▶ Capacity
- ▶ Usage
- ▶ Usage violation
- ▶ Backup

With this information, the storage administrator can:

- ▶ Discover and monitor storage assets enterprise-wide.
- ▶ Report on enterprise-wide assets, files and file systems, databases, users, and applications.
- ▶ Provide alerts (set by the user) on issues such as capacity problems and policy violations.
- ▶ Support chargebacks by usage or capacity.

In the following sections we describe some important features of TPC for Data.

Policy-based management

TPC for Data can enable you to define and enforce storage policies through user-defined alerts, quotas, and constraints. An administrator can be notified by e-mail, pager, the event log, or a systems management console for events like a quota has been exceeded or a constraint violated.

However, finding a problem is only the start — you also require a way to find and fix problems, or discover potential problems. TPC for Data can provide automated solutions through event management. Alerts, quotas that have been exceeded, or constraints that have been violated can result in notification and action, enabling you to fix or even prevent storage outages. For example, if TPC for Data discovers data that has not been accessed in more than a year, it can trigger Tivoli Storage Manager or another archive utility to save to cheaper storage, then delete the original, thus freeing up space.

Automatic file system extension

TPC for Data automated file system extension capability provides the ability to automatically extend a file system when a threshold has been reached. For example, if a file system's threshold is set at 78% and, through monitoring, TPC for Data identifies that this threshold has been exceeded, it can automatically initiate a file system extension to reduce the possibility of a storage-related outage. The feature supports both manual and automated initiated extension. Once you are comfortable with the manual process, you can turn over all the steps to TPC for Data. The agent runs a probe which sends file system statistics to the TPC server. The server compares the current utilization against the policy, and invokes provisioning and extension as necessary.

File sweep

TPC for Data can automatically invoke Tivoli Storage Manager to archive and delete files. This can free up space in a file system and allows more effective management of storage utilization. For example, a policy can be created to archive all files over 365 days old to tape using Tivoli Storage Manager, and then delete the files to free up the disk space.

Disk system reporting

TPC for Data gathers and reports on disk systems — including physical characteristics such as the drive's manufacturer, model, serial number, capacity, and rotational speed. Also included is how the allocation with the LUN, of logical volumes, snapshot copy volumes, and free space. This feature allows users to perform the following operations, subject to the vendor's implementation of Storage Management Initiative - Specification (SMI-S):

- ▶ Display the physical disks behind what the host sees as a disk drive
- ▶ Show the allocated and free capacity of systems in the network
- ▶ List disk volumes that have been allocated but are not in use
- ▶ Show which hosts have access to a given disk volume
- ▶ Show which hosts have access to a given disk drive (within the disk system)
- ▶ Show which volumes (and disks) a host has access to
- ▶ IBM SAN Volume Controller reporting

IBM SAN Volume Controller disk reporting

TPC for Data's general subsystem reporting includes support for SVC reporting:

- ▶ Provides a system view of allocated and free space
- ▶ Provides a view of storage from the host perspective
- ▶ Maintains a view of logical and physical drives
- ▶ Maintains a historical view of allocated and free capacity
- ▶ Allows grouping of subsystems

NAS support

TPC for Data can enable storage administrators to monitor, report on, and manage N series solutions:

- ▶ TPC for Data reports on the N series hardware configuration:
 - Provides RAID array information
 - Describes the physical relationship to the file system
 - Uniquely identifies volumes as Flexible Volumes in the TPC GUI
 - Shows the volume space in regular volumes
 - Shows the reported volume space in Flexible Volume (FlexVol)
 - Completes the discovery of Flexible Volumes via SNMP
- ▶ Quotas can be defined on N series storage, and predefined alerts can be raised.
- ▶ TPC for Data provides a network-wide view of NAS storage, including free versus used space, wasted space, file owners, and files not backed up.

Files by type reporting

This feature collects information about and reports on files by the file type (extension) so that you can see what kind of capacity is being utilized by files of each type (for example, JPEG, GIF, HTML, MPEG). Graphs can be produced showing the space used (and number of files) by each file type.

Push button integration with Tivoli Storage Manager

The following capabilities are available:

- ▶ You can define and configure a *constraint* (which defines acceptable and unacceptable uses of storage) to request a Tivoli Storage Manager archive or backup of the N largest violating files. For example, a constraint might be defined that reports on and automatically performs a Tivoli Storage Manager archive and delete of the largest MP3 files.
- ▶ File reports can be modified to archive or back up selected files directly from the reports. For example, these might apply to a file system's largest files, orphaned files, and duplicate files. A storage administrator might use this feature to quickly free up storage by archiving and deleting selected files.

Summary

The primary business purpose of TPC for Data is to help the storage administrator keep data available to applications. Through monitoring and reporting, TPC for Data helps the storage administrator prevent outages in the storage infrastructure. Armed with timely information, the storage administrator can take action to keep storage and data available to the application. TPC for Data allows administrators to use their existing storage more efficiently, and more accurately predict future storage growth. For more information about supported platforms and devices, see:

<http://www.ibm.com/servers/storage/software/center/data/interop.html>

12.4.2 IBM TotalStorage Productivity Center for Fabric

IBM TotalStorage Productivity Center for Fabric (TPC for Fabric) is a standards-based solution for managing heterogeneous SANs. It is a comprehensive solution that discovers, monitors, and manages SAN fabric components. By performing a SAN topology discovery and rendering of the components and storage resources, TPC for Fabric enables administrators to validate the intended connections between systems and storage devices.

TPC for Fabric provides automated device discovery, topology rendering, zone control, real-time monitoring and alerts and event management for heterogeneous enterprise SAN environments. The Fabric manager can monitor and report on SAN resources and switch performance. It provides a single location for zone control — TPC for Fabric discovers existing zones and zone members and allows you to modify or delete them. In addition, you can create new zones. Switch performance and capacity management reporting and monitoring can help determine if more bandwidth is required.

TPC for Fabric provides a view of events happening in the SAN environment and records state changes. The events are displayed in a color-coded fashion and can be further customized to reflect organizational priorities. TotalStorage Productivity Center for Fabric forwards events signaling topology changes or updates to the IBM Tivoli Enterprise™ Console, to another SNMP manager, or both.

TPC for Fabric supports heterogeneous storage environments. For more information about TPC for Fabric interoperability, refer to:

<http://www.ibm.com/servers/storage/software/center/fabric/interop.html>

TPC for Fabric has a manager and one or more managed hosts:

- ▶ The manager does the following tasks:
 - Gathers data from agents on managed hosts, such as descriptions of SANs, LUNs, and file systems and host information
 - Provides graphical displays of SAN topology
 - Generates Simple Network Management Protocol (SNMP) events when a change is detected in the SAN fabric
 - Forwards events to the Tivoli Enterprise Console® or an SNMP console
- ▶ An agent resides on each managed host. The agents on the managed hosts do the following tasks:
 - Gather information about the SAN by querying switches and devices for attribute and topology information
 - Gather host-level information, such as file systems and mapping to logical units (LUNs)
 - Gather event information and other information detected by host bus adapters (HBAs)

12.4.3 IBM TotalStorage Productivity Center for Disk

IBM TotalStorage Productivity Center for Disk (TPC for Disk) is designed to centralize management of networked storage devices that implement the Storage Management Interface Specification (SMI-S) established by the Storage Networking Industry Association (SNIA), including the IBM System Storage DS4000, SVC, DS6000, DS8000, and ESS. TPC for Disk performance management also is available for third-party devices that support the SNIA SMI-S standard 1.1, as well as provisioning of third-party storage arrays.

TPC for Disk:

- ▶ Helps reduce storage management complexity and costs while improving data availability
- ▶ Enhances storage administrator productivity
- ▶ Improves storage resource utilization
- ▶ Offers proactive management of storage devices

Discovery of IBM storage devices that are SMI-S enabled

Centralized access to storage devices information, information concerning the system attributes of connected storage devices, is available from the TPC for Disk console.

Centralized management of storage devices

The device configuration/manager console for the SMI-S enabled IBM storage devices can be launched from the TPC for Disk console.

Device management

TPC for Disk can provide access to single-device and cross-device configuration functionality. It can allow the user to view important information about the storage devices that are discovered by TPC for Disk, examine the relationships between those devices, or change their configurations.

It supports the discovery and LUN provisioning of DS4000, SVC, ESS, DS6000, and DS8000. The user can view essential information about the storage, view the associations of the storage to other devices, and change the storage configuration.

Performance monitoring and management

TPC for Disk can provide performance monitoring of ESS, SVC, DS4000, DS6000, and DS8000 storage devices. TPC for Disk can provide:

- ▶ Customization of thresholds based on the storage environment, and generation of events if thresholds are exceeded
- ▶ A “select a LUN” information for better performance
- ▶ Customization of both when, and how often the performance data should be collected
- ▶ Support for high availability or critical applications by allowing customizable threshold settings and generating alerts when these thresholds are exceeded
- ▶ Gauges to track real-time performance so that an IT administrator can:
 - Monitor performance metrics across storage subsystems from a single console
 - Receive timely alerts to enable event action based on individual policies
 - Focus on storage optimization through identification of best LUN

For more information about supported platforms and devices, see:

<http://www.ibm.com/servers/storage/software/center/disk/interop.html>

12.4.4 IBM TotalStorage Productivity Center for Replication

IBM TotalStorage Productivity Center for Replication (TPC for Replication) provides management of the advanced copy services provided by many of the IBM System Storage solutions. It is available in two complementary packages: TPC for Replication and TPC for Replication Two Site BC

IBM TotalStorage Productivity Center for Replication

The basic functions of TPC for Replication produce management of FlashCopy, Metro Mirror and Global Mirror capabilities for the IBM System Storage DS6000, DS8000, ESS, and SAN Volume Controller. These copying and mirroring capabilities help give users constant access to critical information during both planned and unplanned local outages. For more information about supported platforms and devices, see:

<http://www.ibm.com/servers/storage/software/center/replication/interop.html>

TPC for Replication provides:

- ▶ Automation of the administration and configuration of these services with wizard-based session and copy set definitions
- ▶ Operational control of copy services tasks, including starting, suspending and resuming

- ▶ Control of copy sessions to ensure that data on multiple, related, heterogeneous volumes are kept consistent by managing the volume pairs in the session as a consistent unit
- ▶ Tools for monitoring and managing copy sessions

IBM TotalStorage Productivity Center for Replication Two Site BC

The basic function of TPC for Replication Two Site Business Continuity is to provide disaster recovery management through planned and unplanned failover and automation for the IBM System Storage 8000, DS6000, DS8000, SVC, and ESS Model 800.

TPC for Replication Two Site BC helps manage replication to a remote backup site through Metro Mirror or Global Mirror. It monitors the progress of the copy services so you can verify the amount of replication that has been done as well as the amount of time required to complete the replication.

Automated failover is designed to keep critical data online and available to users even if the primary site fails. When the primary site comes back on, TPC for Replication Two Site BC manages failback to the default configuration as well.

12.4.5 Summary

The IBM TotalStorage Productivity Center family of products consist of software components which enable Storage Administrators and others to monitor, configure, and manage storage devices and subsystems primarily within a SAN environment. With graphical user interfaces, users can more easily maintain, customize and expand their storage networks. The IBM TotalStorage Productivity Center brings together the IBM storage management software, and consists of:

- ▶ IBM TotalStorage Productivity Center for Data
- ▶ IBM TotalStorage Productivity Center for Fabric
- ▶ IBM TotalStorage Productivity Center for Disk
- ▶ IBM TotalStorage Productivity Center for Replication
- ▶ IBM TotalStorage Productivity Center for Replication Two Site BC

For more information about TPC, see *IBM TotalStorage Productivity Center V3.1: The Next Generation*, SG24-7194.

12.5 IBM Tivoli Storage Manager

IBM Tivoli Storage Manager is developed from the ground up as a comprehensive storage management solution. IBM Tivoli Storage Manager is more than just a distributed backup product; it is a distributed data backup, recovery, AND storage management solution. It is built on a common code base that is consistent across all Tivoli Storage Manager platforms. This common code base enables Tivoli Storage Manager to scale to manage a single or thousands of desktops and database servers.

IBM Tivoli Storage Manager is designed as an enterprise-wide storage management application that focuses its resources on recovery. Key architectural features unique to Tivoli Storage Manager make it the industry's leading enterprise storage management solution. Tivoli Storage Manager, along with its complementary products, offers SAN-enabled integrated enterprise-wide network backup, archive, storage management, bare machine recovery and disaster recovery capabilities.

12.5.1 Backup methods

IBM Tivoli Storage Manager provides a rich set of tools for backing up data. In addition to the unique progressive backup methodology for backing up client data files, Tivoli Storage Manager also offers the following backup options:

- ▶ Image backup
- ▶ Backup sets
- ▶ Adaptive sub-file backup
- ▶ Database and application data protection
- ▶ Bare machine recovery
- ▶ LAN-free backup
- ▶ NDMP backups
- ▶ Open file backups
- ▶ Selective backups

Additional features allow you to control backup and restores to achieve the best solution for each unique client system. These are:

- ▶ Include/exclude list: Fine level of control for what files and directories are and are not backed up
- ▶ Point-in-time restore: Can restore to the most recent backup, or further back in time
- ▶ Collocation: Keeps related data together for faster restore
- ▶ Reclamation: Consolidates data on tapes for faster restore and more efficient media utilization
- ▶ Encryption: For safe data storage
- ▶ Migration through a hierarchy of storage: Improves backup times and allows use of tiered storage for ROI

These features combine to minimize tape resource requirements and improve restore times. For more information, see the IBM Redbooks, *IBM Tivoli Storage Manager Implementation Guide*, SG24-5416 and *IBM Tivoli Storage Management Concepts*, SG24-4877.

For complete details on Tivoli Storage Manager supported platforms, refer to the Web site:

<http://www.ibm.com/software/tivoli/products/storage-mgr/platforms.html>

In this section we briefly discuss progressive backup, image backups, backup sets, and adaptive sub-file backup. Later in the chapter we discuss database and application data protection, bare machine recovery, and LAN-free backup.

Progressive backup

Key to Tivoli Storage Manager's architecture is *progressive backup*. With the progressive technique, only new or changed files are ever backed up. Unlike other backup products, there is no requirement to do regular full backups and associated incremental or differential backups. Therefore significantly less backup data has to be transferred over the network and stored than with traditional full + incremental or differential backup methods. Restores are also more efficient.

The progressive backup feature is made possible by Tivoli Storage Manager's inbuilt relational database and recovery log, which provides file-level tracking. This file-level tracking drives the potential for other benefits as well. For example, automatic tape reclamation can reorganize data on tape to maximize the utilization of storage space and minimize restore time. An optional tape collocation feature keeps data grouped in logical sets, helping also to enable faster restores.

When a restore is required, unlike other products, it is not necessary to transfer the full backup plus the differential data (a combination that often contains multiple copies of the same files). Instead, the restore process transfers only the actual files required for a full restore. In effect, Tivoli Storage Manager makes it possible to assemble a full backup for almost any point-in-time the administrator wants to specify, thereby helping improve overall restore performance.

Image backups

An image backup is a block-by-block copy of data from the Tivoli Storage Manager client to the backup server. One important function of an image restore is to accelerate recovery in a disaster recovery scenario. With image backup, the Tivoli Storage Manager server does not track individual files in the file system image. File system images are tracked as complete individual objects, with a single retention policy. An image backup provides the following benefits:

- ▶ Can work together with progressive backup to expedite the restore; image backup restored first, then most recent versions of files backed up with the progressive backup
- ▶ Can provide a quicker backup and restore than a file-by-file backup because there is no overhead involved in creating individual files
- ▶ Conserves resources on the server during a backup because only one entry is required for the image
- ▶ Provides a point-in-time picture of a file system
- ▶ Restores a corrupt file system or raw logical volume, with data restored to the same state it was when the last logical volume backup was performed

On the Windows 2000 and 2003 client platform, a Logical Volume Snapshot Agent (LVSA) is included to provide open file support. This allows Tivoli Storage Manager to perform a snapshot backup or archive of files that are open (or locked) by other applications. The snapshot allows is taken from a point-in-time copy that matches the file system at the time the snapshot is taken. In this way the client is able to send a consistent image of the volume as it was at the start of the snapshot process to the Tivoli Storage Manager server.

Backup sets

Tivoli Storage Manager enables you to generate a copy of a backup client's most recent backup from the Tivoli Storage Manager server onto sequential media. This copy, known as a backup set or portable backup, is self-contained and can be used independently of the Tivoli Storage Manager server to restore the client's data from a locally attached device that can also read this media, such as a CD-ROM. This is a method of restoring client data files only, this does not restore a failed system. For more on restoring a failed system see 12.7.1, "Cristie Bare Machine Recovery (CBMR)" on page 427.

Adaptive sub-file backup

Adaptive sub-file backup is well suited for mobile and remote computers. These computers often have limited access to the infrastructure that serves the rest of the company. Some limitations include being attached to the corporate network with reduced bandwidth, limited connect time, and minimal assistance to perform the backup.

Adaptive sub-file backup reduces the amount of data transferred by backing up only the changed portion of a file, either on the byte level or on the block level, instead of transferring the whole file to the server every time. The changed file portion is backed up as a differential backup relative to the last complete backup of the file and it is called a delta file. All changes since the last complete backup of the file are included in this delta file. In the case of a restore, this allows for the restore of the whole file by restoring only two sub-file components, one delta file and the last complete backup of the whole file.

The adaptive sub-file backup, as well as the restore of a file consisting of a base file and the delta file, is completely transparent to the user. All necessary file data separations or reconstructions happen under the covers of the backup-archive client.

Archive

Tivoli Storage Manager also provides an inbuilt archive function. Unlike backup, which is managed by version, an archive is a stand-alone retained copy of a file or group of files, which is controlled by a retention period. For example, legal requirements might dictate that each year's financial results must be saved in an archive and kept for 7 years. Tivoli Storage Manager can do this, and optionally delete the files from the source server, freeing up space. Tivoli Storage Manager's storage pool hierarchy can also handle moving of the archived data to new media, as new technology becomes available.

12.5.2 Disaster Recovery Manager

IBM Tivoli Storage Manager Extended Edition includes the Disaster Recovery Manager (DRM) component. DRM provides disaster recovery for the tivoli storage manager server and assists in disaster recovery for clients.

DRM is designed to manage the restore of critical business systems and minimize the recovery time from any type of disaster that might affect on-site storage. It creates a Disaster Recovery Plan and facilitates the tracking of off-site volumes. The recovery plan is integrated with and maintained on the Tivoli Storage Manager server. It contains detailed recovery steps and automated computer scripts. It can be customized with relevant data like the location of the Disaster Recovery site, contact numbers for system administrators, hardware vendors, or what ever else is important to an individual company. The report is generated automatically, pulling data out of the Tivoli Storage Manager database, thus creating an up-to-date disaster recovery plan on a nightly basis. The plan can be sent off-site with your off-site tapes.

One of the key features of DRM is the ability to track the life cycle of an off-site tape. For example, an off-site copy moves from the on-site tape library, to a courier, to a vault. As the data on the off-site tapes expires and the tapes are reclaimed, DRM tracks the fact that the tapes should be returned on-site for reuse. This greatly simplifies the management of tape rotation, and assures that off-site copies of backup data are current and complete.

For a detailed explanation on volume tracking and other IBM Tivoli Storage Management Disaster Recovery Manager concepts, see the IBM Redbooks, *IBM Tivoli Storage Manager Implementation Guide*, SG24-5416, *IBM Tivoli Storage Management Concepts*, SG24-4877, and *Disaster Recovery Strategies with Tivoli Storage Management*, SG24-6844.

The core functions of IBM Tivoli Storage Manager Disaster Recovery Manager include:

- ▶ Automated generation of a customized server Disaster Recovery Plan
- ▶ Off-site recovery media management
- ▶ Inventory of machine information required to recover the server and its clients
- ▶ Centralized management of the disaster recovery process
- ▶ Executable scripts that assist in recovery automation
- ▶ Volume tracking for physical vaulting of tapes
- ▶ Electronic vaulting of storage pool and database backups
- ▶ Managing and identifying off-site media required for recovery
- ▶ Ability to customize

Figure 12-2 illustrates many of the functions of Disaster Recovery Manager.

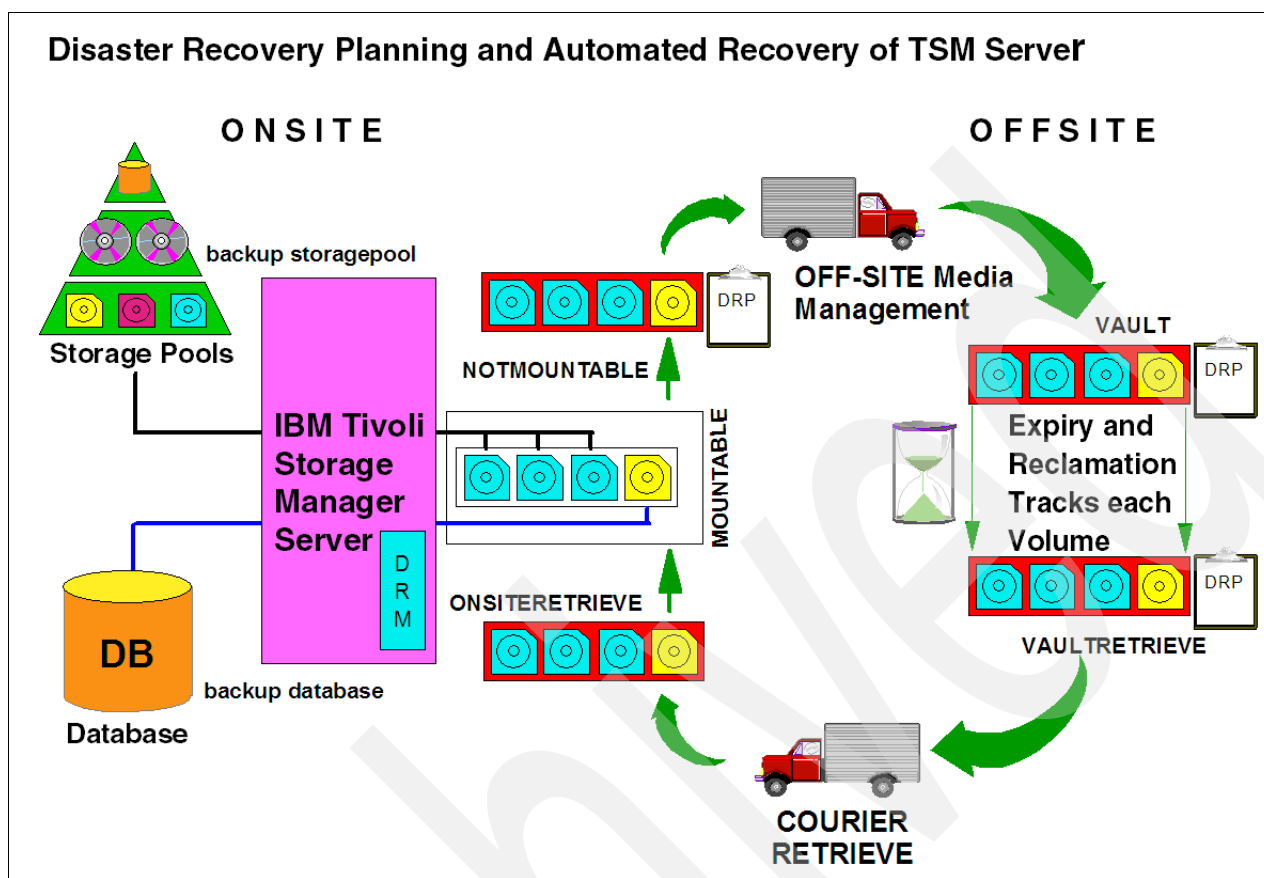


Figure 12-2 IBM Tivoli Disaster Recovery Manager functions

For more information about IBM Tivoli Storage Manager Disaster Recovery Manager, see the following Web site:

<http://www.ibm.com/software/tivoli/solutions/disaster/>

12.5.3 Disaster Recovery for the Tivoli Storage Manager Server

This section discusses ways of automating the protection of the Tivoli Storage Manager volumes. Electronically moving this data to a secondary site improves the ability to restore data in the event of a primary site failure. We review three methods to accomplish this, each offering a greater degree of preparation for Disaster Recovery. These methods include:

- ▶ Electronic vaulting over a SAN
- ▶ Electronic vaulting with virtual volumes
- ▶ Server-to-Server Hot-Standby using Incremental Export

Electronic vaulting over a SAN

Electronic vaulting consists of electronically transmitting and creating backups at a secure facility, moving business-critical data off-site faster and more frequently than traditional data backup processes allow. The receiving hardware must at a remote site, physically separated from the primary site, and the data stored at the remote site for recovery, should there be a disaster at the primary site.

Figure 12-3 shows a solution with an active Tivoli Storage Manager server that creates off-site copies to a SAN attached tape library which resides at a remote site. This configuration does not require an active Tivoli Storage Manager server at the remote site. The off-site vaulting location can have a standby server which can be used to restore the Tivoli Storage Manager Database and bring the server online quickly.

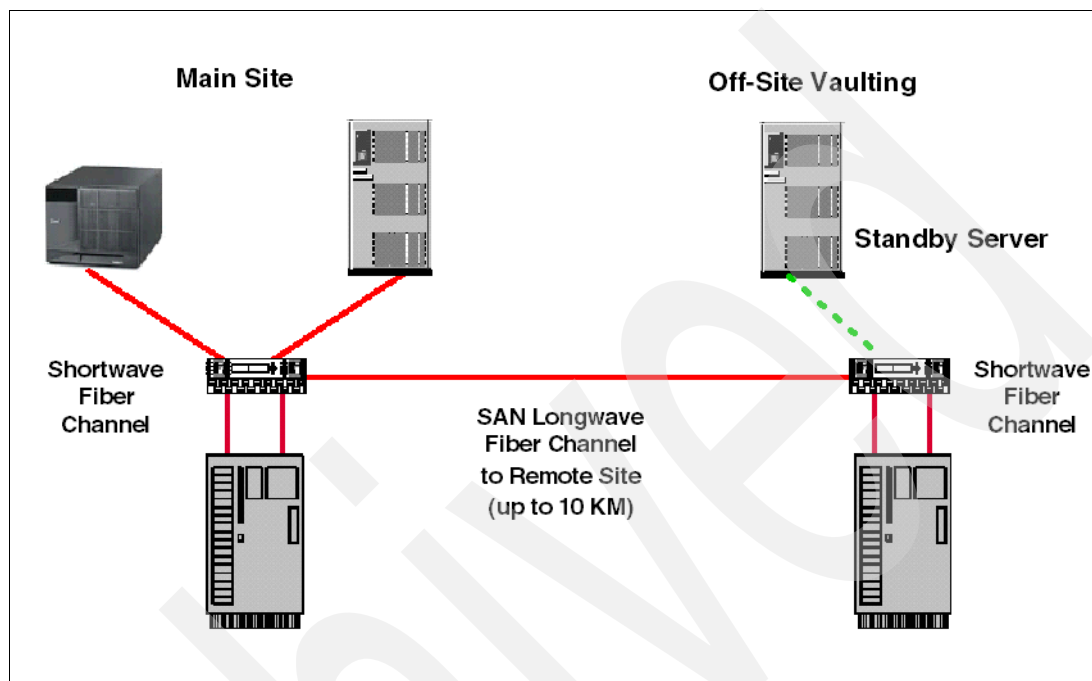


Figure 12-3 Remotely connected tape library

Electronic vaulting with virtual volumes

It is also possible, by using DRM to define server hierarchies or multiple peer-to-peer servers, to build a mutual vaulting solution. As shown in Figure 12-4, the Tivoli Storage Manager server APPLE backs up Tivoli Storage Manager data to virtual volumes. Virtual volumes are not physical volumes, they are definitions which send data to a target Tivoli Storage Manager server. In this case the target server is PECAN.

Tivoli Storage Manager data received from APPLE by PECAN is written to physical volumes in an Archive primary storage pool. By defining virtual volumes on PECAN to send data to APPLE, each system can back up the vital Tivoli Storage Manager data for the other. For details on setting up this method of electronic vaulting, see the appropriate Tivoli Storage Manager Administrator's Guide for your operating system.

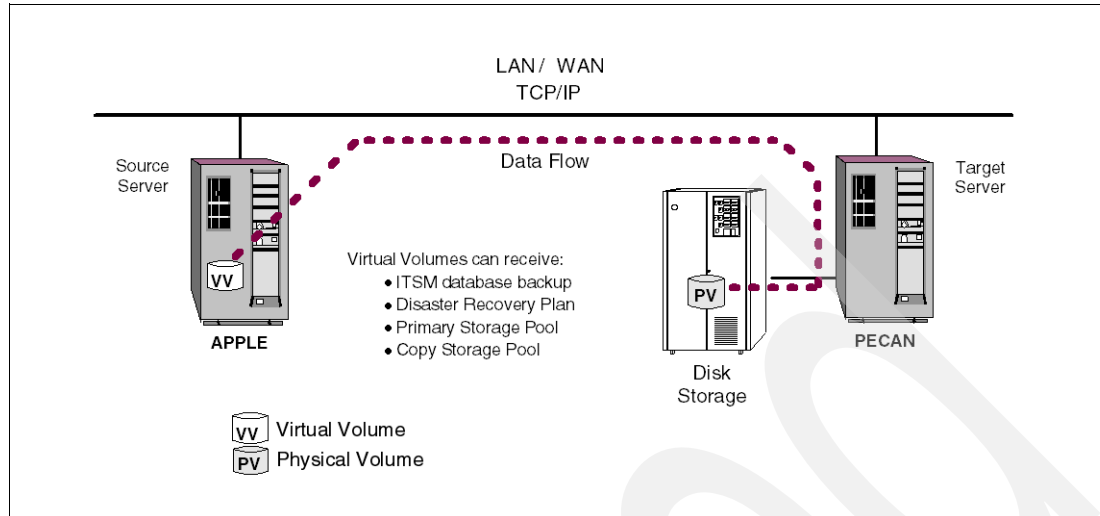


Figure 12-4 Server-to-Server virtual volumes

Although the foregoing configuration has an active Tivoli Storage Manager server at each site, the data written to the virtual volumes is saved as archived volumes on the target server. PECAN does not have an active instance of the APPLE Tivoli Storage Manager server running. These volumes must be restored to a separate instance of the Tivoli Storage Manager server on PECAN to activate the Tivoli Storage Manager server APPLE.

Server-to-server hot-standby using incremental export

Tivoli Storage Manager export and import processing allow a mutual hot-standby configuration, as follows:

- ▶ The export process can initiate an immediate import to another server over TCP/IP.
- ▶ The import process can merge existing client file spaces.
- ▶ The export process can limit client file data based on a date and time specification.

Together these enhancements allow the source server to incrementally update a local node's data to a target server. This allows server-to-server export and import operations to maintain duplicate copies of client data on two or more servers. This also allows the export/import process to include storage devices of different types as well as different server platforms at the source and target locations.

Figure 12-5 shows a sample hot-standby configuration. When the BIRCH and MAPLE servers are running the Tivoli Storage Manager server, the procedure for establishing the hot-standby includes the following steps:

1. Connecting the servers via a TCP/IP LAN/WAN
2. Defining the server BIRCH to MAPLE and MAPLE to BIRCH, including a server password and TCP/IP address
3. Exporting the client node definitions from one server to the other
4. Setting up a scheduled incremental export of the client node data between servers

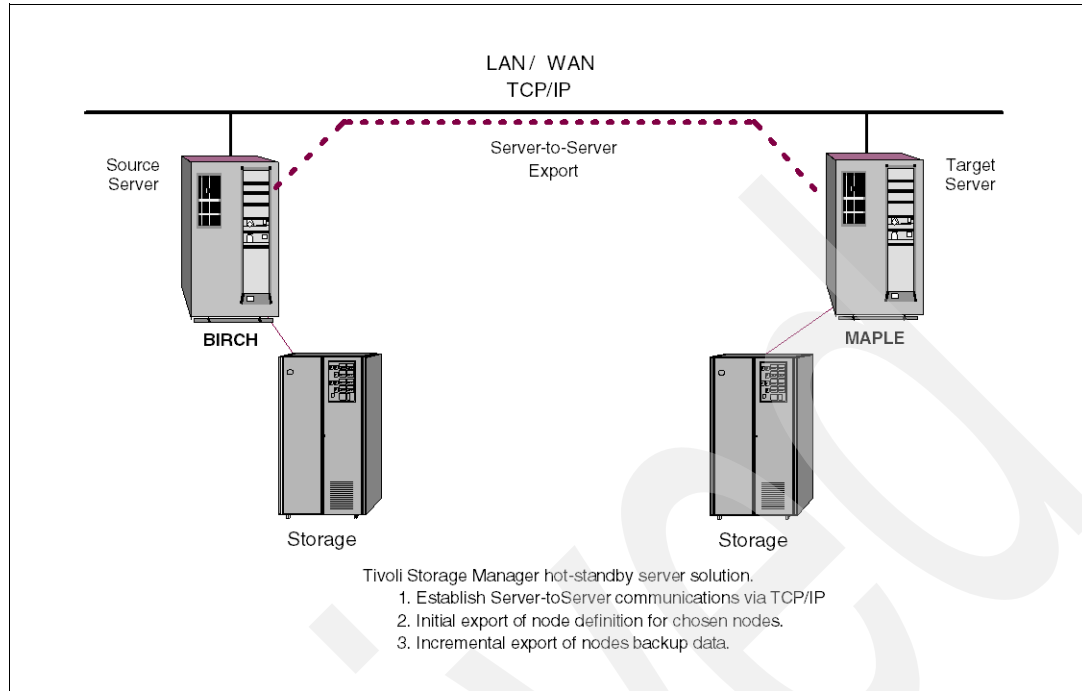


Figure 12-5 Tivoli Storage Manager hot-standby server

12.5.4 IBM Tivoli Storage Manager for Databases

IBM Tivoli Storage Manager for Databases provides integrated protection for a wide range of applications and databases. Tivoli Storage Manager for Databases exploits the backup certified utilities and interfaces provided by Oracle and Microsoft SQL Server databases. Its functionality is included in the IBM DB2 Universal Database™ and Informix® 10x packages and does not have to be purchased separately. For mySAP systems with their underlying databases, a separate product, IBM Tivoli Storage Manager for Enterprise Resource Planning Data Protection for mySAP is available.

Data Protection for IBM DB2 UDB

The Tivoli Storage Manager for Databases functionality is included in the IBM DB2 UDB package and does not have to be purchased separately.

As shown in Figure 12-6, the DB2 backup utilities (but not export/import) are fully integrated with Tivoli Storage Manager services because the DB2 utilities use the Tivoli Storage Manager API. This means that an intermediate file is not created during the backup operation — the backup is stored directly onto the Tivoli Storage Manager servers. Both online and offline backups can be performed with Tivoli Storage Manager, and DB2 data is automatically restored by using the DB2 restore utility.

Tivoli Storage Manager can also archive DB2 log files. DB2 provides a user exit program for backing up and restoring its log files directly to Tivoli Storage Manager. Log files are moved to the DB2 user exit program when they become inactive. The logs are also automatically retrieved from Tivoli Storage Manager for roll-forward recovery.

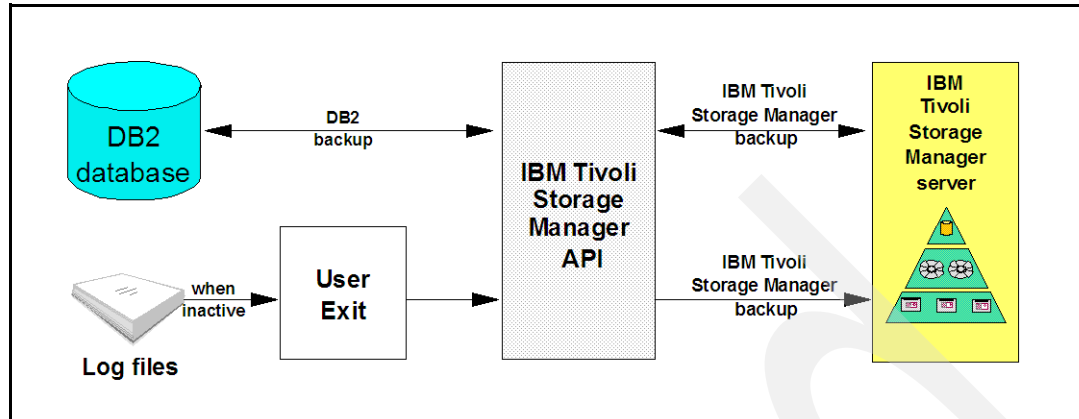


Figure 12-6 IBM Tivoli Storage Manager interface with DB2

Data Protection for DB2 UDB also provides FlashCopy backup of databases, when Tivoli Storage Manager for Advanced Copy Services is installed.

Data Protection for Microsoft SQL Server

Data Protection for SQL performs online backups and restores of Microsoft SQL Server databases to Tivoli Storage Manager Server. Data Protection for SQL helps protect and manage SQL Server data, and makes it easy to:

- ▶ Back up any SQL Server database to any Tivoli Storage Manager Server.
- ▶ Perform full and transaction log backups and restores of SQL databases.
- ▶ Perform backups with an expanded range of options, such as differential, file, and group operations.
- ▶ Perform operations from multiple SQL Server instances on the same machine as Data Protection for SQL.
- ▶ Perform any backup using data striping in parallel threads using parallel sessions.
- ▶ Automate scheduled backups using the Tivoli Storage Manager scheduler.
- ▶ Perform expanded restore operations on backup objects, such as relocating, restoring to named marks, and partially restoring full backups.
- ▶ Restore database backups to a different SQL Server.
- ▶ Retain, with a backup, the information required to recreate or move SQL databases or files, such as sort order, code page, and Unicode information, or file group and file logical and physical names. The meta object information is retained on the Tivoli Storage Manager Server separately from the backup data objects.
- ▶ Deactivate all active objects, all objects of a particular backup type, or specific objects.
- ▶ Deactivate objects older than a specified number of days.
- ▶ Set automatic expiration of backup objects based on version limit and retention period.
- ▶ Participate in MSCS failover clusters.
- ▶ Apply failover clustering (for maintenance or restoring the master database) without unclustering.

Data Protection for SQL supports LAN-free backup for Tivoli Storage Manager.

Data Protection for Oracle

IBM Tivoli Storage Manager Data Protection for Oracle integrates with Oracle Recovery Manager (RMAN) to backup and restore Oracle databases.

In addition, Tivoli software enables you to use the *duplex copy* feature available in RMAN, making it possible to send a backup to two separate storage tapes simultaneously.

Data Protection for Oracle also provides FlashCopy backup of databases, when Tivoli Storage Manager for Advanced Copy Services is installed.

Oracle RMAN and Data Protection for Oracle

RMAN provides consistent and secure backup, restore, and recovery performance for Oracle databases. While the Oracle RMAN initiates a backup or restore, Data Protection for Oracle acts as the interface to the Tivoli Storage Manager server. The Tivoli Storage Manager server then applies administrator-defined storage management policies to the data. Data Protection for Oracle implements the Oracle defined Media Management API, which interfaces with RMAN and translates Oracle commands into Tivoli Storage Manager API calls to the Tivoli Storage Manager Server.

With the use of RMAN, Data Protection for Oracle can perform the following functions:

- ▶ Full backup function for the following while online or offline:
 - Databases
 - Tablespaces
 - Datafiles
 - Archive log files
 - Control files
- ▶ Full database restores while offline
- ▶ Tablespace and datafile restore while online or offline

Backup using Data Protection for Oracle

The RMAN utilities are used by interfacing them with an external media manager (such as IBM Tivoli Storage Manager) via the library provided by Oracle. After the RMAN initiates a backup or restore, Data Protection for Oracle acts as the interface to Tivoli Storage Manager. It translates Oracle API calls into Tivoli Storage Manager API calls to the Tivoli Storage Manager server for the desired functions. This causes minimal disruption, so they can be carried out while users continue working.

Figure 12-7 shows how Oracle works in conjunction with Data Protection for Oracle and the Tivoli Storage Manager server.

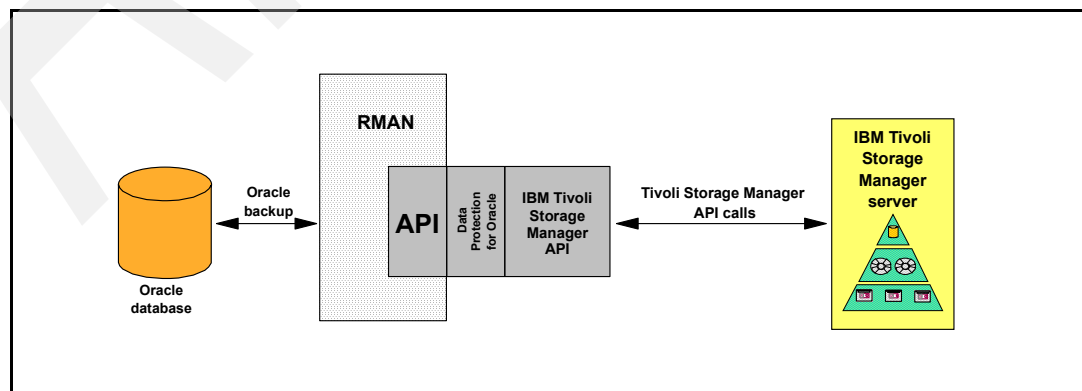


Figure 12-7 Oracle interfacing with IBM Tivoli Storage Manager

Oracle provides backup and restore functions for Oracle databases: full or partial, offline or online. When you have identified which database to back up, Oracle locates all of the necessary files and sends them to Tivoli Storage Manager. Recovery is managed similarly, in that Oracle issues the appropriate commands to Tivoli Storage Manager to restore the required objects (such as tables, control files, and recovery logs) and then performs the recovery.

Data Protection for Oracle supports backup and restore operations in a LAN-free environment.

Data Protection for Informix

IBM Informix Dynamic Server (IDS) v10.0 has built-in support for backup and restore of the IDS database to and from Tivoli Storage Manager. With this integration, you can backup and restore IDS databases directly to and from a Tivoli Storage Manager server using just the Tivoli Storage Manager API client.

In previous versions of IDS, backups and restores required the Data Protection for Informix module to be installed on the database server, that served as an interface between the database backup utilities and the Tivoli Storage Manager server.

Regardless of the interface used between IDS and Tivoli Storage Manager, Informix uses the ON-Bar (online backup archive) utility to manage database backups and restores.

The Tivoli Storage Manager for Informix client supports parallel sessions for both backups and restores. This can help use network resources efficiently during regular backups and can help reduce restore times for Informix databases.

Components

The Tivoli Storage Manager interface to ON-Bar consists of the following components, shown in Figure 12-8: the ON-Bar program, the X/OPEN Backup Services Application Programmer's Interface (XBSA), the Tivoli Storage Manager API client (or Data Protection for Informix), and the ON-Bar sysutils tables, message file, and emergency boot file.

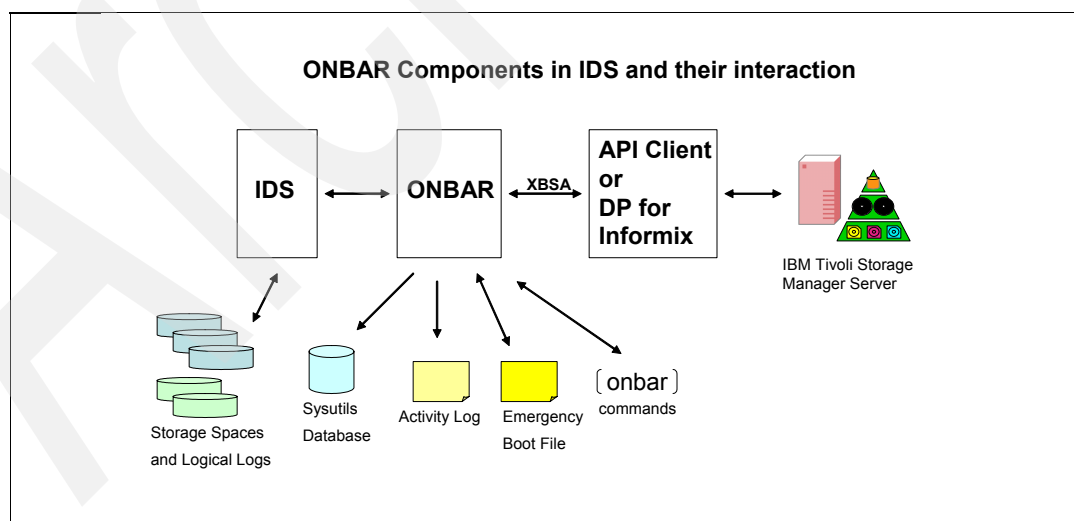


Figure 12-8 Informix backup with Tivoli Storage Manager

Emergency boot file

The ON-Bar emergency boot file contains enough information to cold-restore an online server instance. The boot file is no longer just for critical dbspaces. It is used instead of the sysutils tables during a cold restore of the database.

For more information about Tivoli Storage Manager for Databases, see the following Web site:

<http://www.ibm.com/software/tivoli/products/storage-mgr-db/>

12.5.5 IBM Tivoli Storage Manager for Mail

IBM Tivoli Storage Manager for Mail is a software module for Tivoli Storage Manager that automates the data protection of e-mail servers running either IBM Lotus Domino or Microsoft Exchange. This module utilizes the application program interfaces (APIs) provided by e-mail application vendors to perform online *hot* backups without shutting down the e-mail server and improve data-restore performance. As a result, it can help protect the growing amount of new and changing data that should be securely backed up to help maintain 24x7x365 application availability.

Data Protection for Lotus Domino

Tivoli Storage Manager provides a backup solution for a heterogeneous Lotus Domino environment, which includes the backup-archive client and the Data Protection component for Domino. Together, these provide a complete backup solution to fulfill the requirements of Domino storage management.

Data Protection for Lotus Domino helps protect and manage Lotus Domino Server data by making it easy to:

- ▶ Implement centralized, online, selective, and incremental backup of Lotus Domino databases
- ▶ Maintain multiple versions of Domino databases
- ▶ Maintain multiple versions of Domino database backups
- ▶ Archive Lotus Domino transaction log files, when archival logging is in effect
- ▶ Restore backup versions of a Lotus Domino database and apply changes made since the backup from the transaction log
- ▶ Restore Domino databases to a specific point-in-time
- ▶ Recover to same or different Domino server
- ▶ Expire database backups automatically based on version limit and retention period
- ▶ Expire archived transaction logs when no longer required
- ▶ Automate scheduled backups
- ▶ Recover one or more archived transaction logs independent of a database recovery
- ▶ Recover from the loss of the transaction log
- ▶ Archive the currently filling transaction log file
- ▶ Supports Lotus Domino *Individual Mailbox Restore*

Data Protection for Lotus Domino provides two types of database backup, incremental and selective, and a log archive function. Incremental backup provides a conditional backup function that creates a full online backup of Domino databases, when necessary. The specific conditions that determine when a new backup is necessary vary depending on whether the database is logged or not.

Selective backup unconditionally backs up the specified databases, unless they are excluded from backup through exclude statements. When archival logging is in effect, changes to logged databases can be captured in between full backups, by archiving the transaction log.

The logical components of Data Protection for Domino are shown in Figure 12-9.

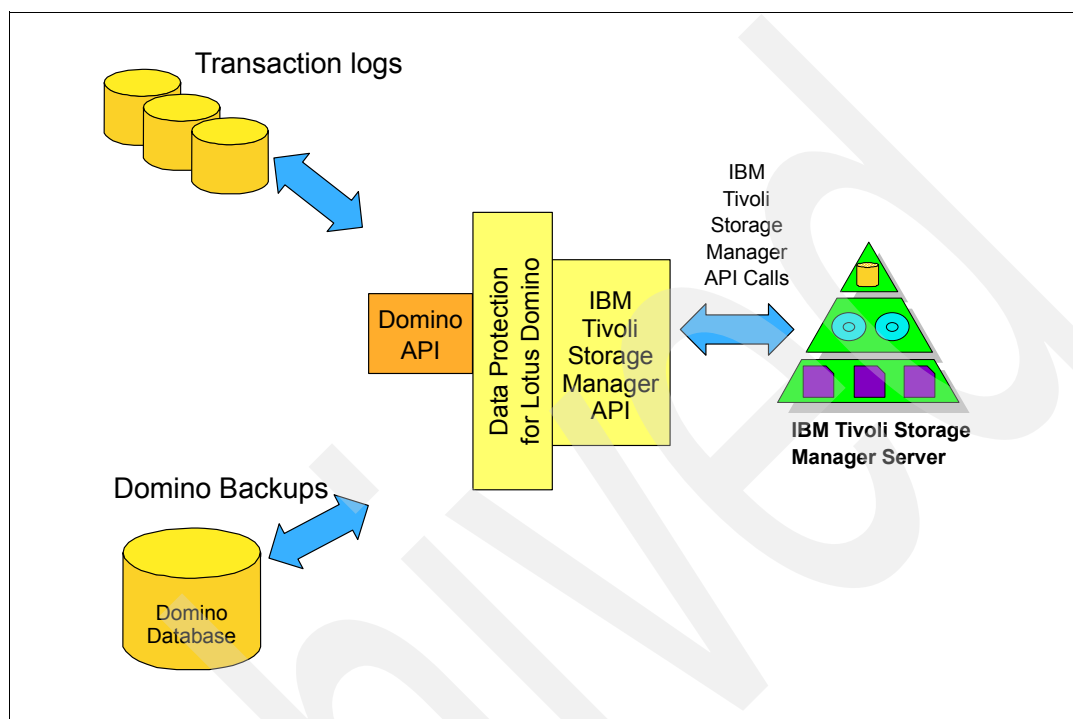


Figure 12-9 Logical components of Data Protection for Domino

Data Protection for Microsoft Exchange

Tivoli Storage Manager for Mail Data Protection for Microsoft Exchange exploits Microsoft-provided APIs to allow the online backup and restore of Microsoft Exchange Server data to Tivoli Storage Manager. These functions are initiated using automated scheduling, command-line, or graphical user interfaces on Windows environments.

Data Protection for Microsoft Exchange provides complete integration with Microsoft Exchange APIs by offering:

- ▶ Centralized online backups (full, copy, incremental, and differential) of Exchange Directory and Information Stores to Tivoli Storage Manager server storage
- ▶ Automatic expiration and version control by policy
- ▶ Failover for Microsoft Cluster Server (MSCS)
- ▶ Parallel backup sessions for high performance
- ▶ Automated transaction log file management
- ▶ LAN-free backup

Data Protection for Microsoft Exchange supports Microsoft Exchange Individual Mailbox Restore in combination with Tivoli Storage Manager backup-archive client and the Microsoft Exchange Mailbox Merge Program (ExMerge).

Data Protection for Microsoft Exchange also provides VSS backup of Exchange databases, when Tivoli Storage Manager for Copy Services is installed.

Data Protection for Microsoft Exchange helps protect and manage Exchange Server data by facilitating the following operations:

- ▶ Performing full, copy, differential, and incremental backups of the Microsoft Exchange Directory and Information Store databases
- ▶ Restoring a full Directory or Information Store database and any number of associated transaction logs
- ▶ Deleting a Directory or Information Store database backup from Tivoli Storage Manager storage
- ▶ Automating scheduled backups
- ▶ Automating deletion of old backups

Data Protection for Microsoft Exchange communicates with Tivoli Storage Manager using its API, and with an Exchange Server using the Exchange API, as shown in Figure 12-10.

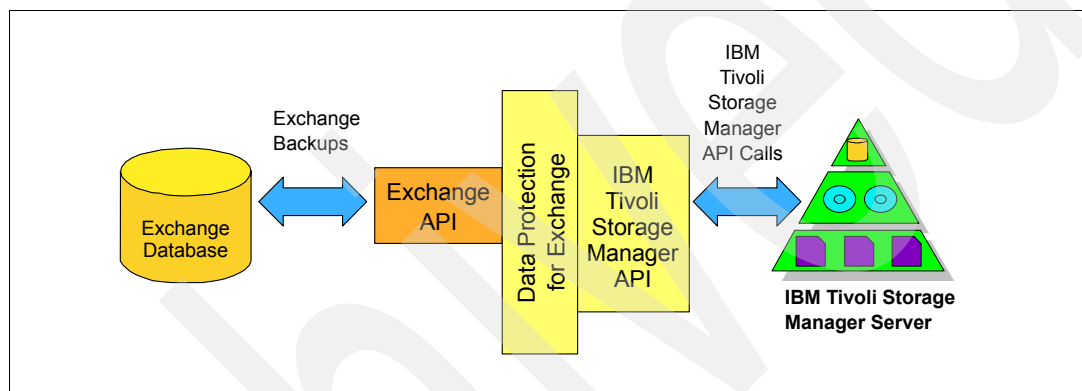


Figure 12-10 Overview of Data Protection for Microsoft Exchange operating environment

Data Protection for Microsoft Exchange also runs in an MSCS environment.

For more information about IBM Tivoli Storage Manager for Mail, see the following Web site:

<http://www.ibm.com/software/tivoli/products/storage-mgr-mail/>

12.5.6 IBM Tivoli Storage Manager for Application Servers

IBM Tivoli Storage Manager for Application Servers provides back up of stand-alone Application Servers, Network Deployment, Express, and Enterprise configurations of WebSphere Application Servers. A Network Deployment configuration is backed up from the node that contains the Network Deployment Manager. Tivoli Storage Manager for Application Servers can also back up multiple instances of the Network Deployment Manager and Application Server concurrently. However, multiple concurrent back up sessions of the same node or cell are not supported.

Tivoli Storage Manager for Application Servers backs up the following Network Deployment Manager and Application Server data:

- ▶ The properties directory
- ▶ WebSphere Application Server Version 5.x.x Web applications:
 - Java archive files (JAR)
 - Enterprise archive files (EAR)
 - Web archive files (WAR)
 - Class files
 - Configuration information from the configuration repository

For more information about IBM Tivoli Storage Manager for Application Servers, see the following Web site:

<http://www.ibm.com/software/tivoli/products/storage-mgr-app-servers/>

12.5.7 IBM Tivoli Storage Manager for Enterprise Resource Planning

IBM Tivoli Storage Manager for Enterprise Resource Planning (ERP) is a software module that works with Tivoli Storage Manager to optimally protect the infrastructure and application data and improve the availability of *mySAP* systems.

Tivoli Storage Manager for ERP builds on the *mySAP* database, a set of database administration functions integrated with *mySAP* for database control and administration. The Tivoli Storage Manager for ERP software module allows multiple *mySAP* database servers to utilize a single Tivoli Storage Manager server to automatically manage the backup of data. As the intelligent interface to the *mySAP* database, Tivoli Storage Manager for ERP is SAP certified in heterogeneous environments, supporting large-volume data backups, data recovery, data cloning, and Disaster Recovery of multiple *mySAP* systems.

Tivoli Storage Manager for ERP offers the following features:

- ▶ It can handle large amounts of data reliably to minimize downtimes and operational cost and flexibly adapts to changing requirements in a dynamic system infrastructure.
- ▶ Unlike other offerings, which merely provide an interface to a generic storage management tool, it is specifically designed and optimized for the SAP environment, delivering business value by focusing on automated operation, built-in productivity aids, optimum performance, and investment protection.
- ▶ Multiple management classes provide a library structure for sets of device parameters, which means allowing the device parameters to be called by class names.
- ▶ Multiple path/session support provides one path or session per tape device, maximizing backup and restore performance.
- ▶ Multiple server operations allow multiple Tivoli Storage Manager servers to be used in parallel for backup and restore, eliminating capacity bottlenecks.
- ▶ Multiplexing merges multiple data streams into one data stream, thereby exploiting the full write bandwidth of storage devices and minimizing backup window times.
- ▶ Multiple log files store log files in two management classes, providing additional security through redundancy of log files.
- ▶ SAN support and integration allows the use of SAN Fibre Channels with high bandwidth, freeing up the LAN.
- ▶ Centralized management with administration assistant enables policies and processes to be managed from a central point, achieving consistent backup and recovery of critical data, even in *lights out* operations.
- ▶ Policy controlled file migration across storage hierarchies uses the most cost effective storage media based on retention periods and resource usage, decreasing the cost of ownership.

The LAN-free support offered by IBM Tivoli Storage Manager for ERP allows disparate servers to access tape libraries through the SAN (library sharing feature). Therefore, data can be sent and retrieved directly from the production system to and from the tape. Especially in restore situations, the LAN-free feature allows administrators to retrieve data directly over the SAN with virtually no LAN impact.

The solution can be expanded with Tivoli Storage Manager for Advanced Copy Services to support *mySAP* databases on ESS, SVC, DS6000, and DS8000 by leveraging the FlashCopy function, which can virtually eliminate the impact of the backup on the production server. With the FlashCopy function, you can make an instant copy of a defined set of disk volumes, allowing immediate access to data on the target volumes and greatly reducing the impact of the backup on the production server.

IBM Tivoli Storage Manager for ERP uses native utilities to back up and restore Oracle database objects. These utilities pass the database objects as a temporary list of files. IBM Tivoli Storage Manager for ERP receives control with the parameters specified in its profile, sorts the files for optimum load balancing and throughput, and initiates the specified number of multiple sessions. These sessions execute in parallel and transfer data between the *mySAP* database and storage devices to maximize overall throughput.

IBM Tivoli Storage Manager for ERP connects to DB2 UDB and to the *mySAP*-specific BRArchive functions to provide the same automation and productivity functions in DB2 UDB environments. It supports all DB2 backup modes.

To reduce network-induced bottlenecks, IBM Tivoli Storage Manager for ERP can simultaneously use multiple communication paths for data transfer with the Tivoli Storage Manager server. The number of sessions can be set individually for each path depending on its bandwidth. Multiple Tivoli Storage Manager servers can also be used to back up and restore simultaneously, helping minimize bottlenecks in server capacity.

Data Protection for mySAP

This section explains the Data Protection *for mySAP* architecture and gives an introduction to the product features. Data Protection *for mySAP*, together with Tivoli Storage Manager, provides a reliable, high performance and production oriented backup and restore solution that allows you to back up and restore *mySAP*. It is integrated with DB2's backup and recovery facilities and, specifically *for mySAP* environments, SAP administration tools BRARCHIVE and BRRESTORE, so that SAP's recommended backup and recovery procedures are followed.

The Data Protection *for mySAP* is used for backup and restore of database contents, control files, and offline DB2 log files (see Figure 12-11).

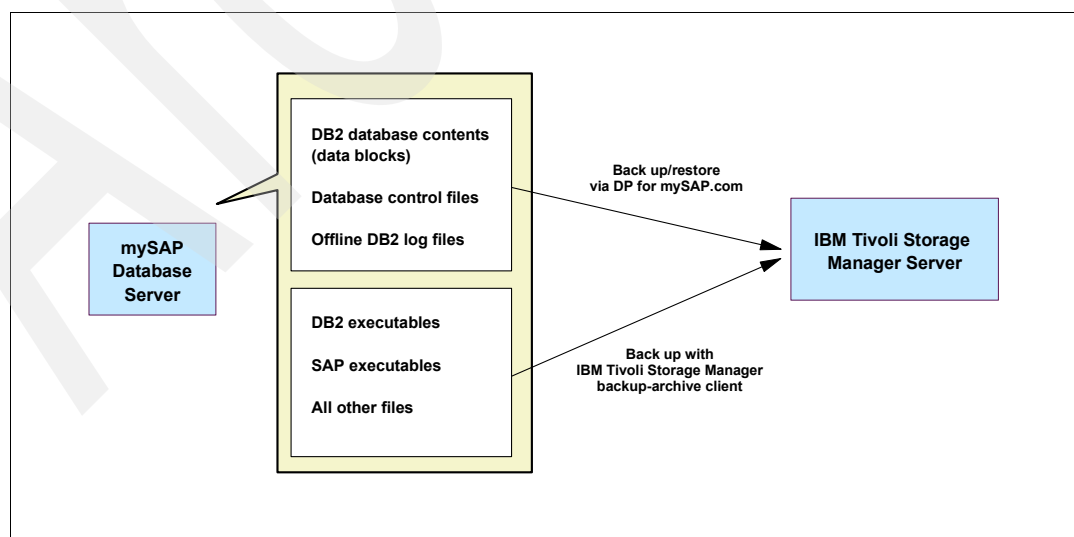


Figure 12-11 Scope of Data Protection for mySAP

Other files, such as SAP and DB2 executables, control files or user data, can be backed up using the IBM Tivoli Storage Manager backup-archive client. As a consequence, in the case of a Disaster Recovery, you have to make sure that all DB2 and *mySAP* executables are available before starting restore and recovery of your database using the Data Protection for *mySAP* and the DB2 utilities.

For more information about IBM Tivoli Storage Manager for Enterprise Resource Planning, see the following Web site:

<http://www.ibm.com/software/tivoli/products/storage-mgr-erp/>

12.5.8 Tivoli Storage Manager for Advanced Copy Services

Tivoli Storage Manager for Advanced Copy Services software provides online backup and restore of data stored in mySAP, DB2 and Oracle applications by leveraging the copy services functionality of the underlying storage hardware. Using hardware-based copy mechanisms rather than traditional file-based backups can significantly reduce the backup/restore window on the production server. Backups are performed through an additional server called the *backup server* which performs the actual backup. Since the backup operation is offloaded to the backup server, the production server is free from nearly all the performance impact. The production server's processor time is dedicated for the actual application tasks, so application users' performance is not affected during backup.

Specifically, Tivoli Storage Manager for Advanced Copy Services provides FlashCopy integration on the DS6000, DS8000, SVC, and ESS for split mirror backup and optional FlashBack restore of mySAP, DB2 UDB, and Oracle databases.

Tivoli Storage Manager for Advanced Copy Services is used in conjunction with some other products to interact with the applications and perform the backup from the backup server to Tivoli Storage Manager. The products which it interfaces with are Tivoli Storage Manager for Enterprise Resource Planning (Data Protection for mySAP), Tivoli Storage Manager for Databases (Data Protection for Oracle), and the inbuilt Tivoli Storage Manager interfaces for DB2 UDB.

Tivoli Storage Manager for Advanced Copy Services has the following modules currently available:

- ▶ Data Protection for IBM Disk Storage and SAN Volume Controller for mySAP with DB2 UDB - FlashCopy integration for mySAP with DB2 on SVC, DS6000, DS8000
- ▶ Data Protection for IBM Disk Storage and SAN Volume Controller for mySAP with Oracle - FlashCopy integration for mySAP with Oracle on SVC, DS6000, DS8000
- ▶ Data Protection for IBM Disk Storage and SAN Volume Controller for Oracle - FlashCopy integration for Oracle on SVC, DS6000, DS8000
- ▶ DB2 UDB Integration Module and Hardware Devices Snapshot Integration Module - FlashCopy integration for DB2 on ESS, SVC, DS6000, DS8000
- ▶ Data Protection for ESS for Oracle - FlashCopy integration for Oracle on ESS
- ▶ Data Protection for ESS for mySAP - FlashCopy integration for mySAP with DB2 or Oracle on ESS

For more details, see the IBM Redbook, *IBM Tivoli Storage Manager for Advanced Copy Services*, SG24-7474.

12.6 Tivoli Storage Manager for Copy Services

IBM Tivoli Storage Manager for Copy Services helps protect business critical Microsoft Exchange databases that require 24x7 availability. It offers options to implement high-efficiency backup of business-critical applications while virtually eliminating backup-related impact on the performance of the MS Exchange production server. This is done by integrating the snapshot technologies of the storage system with Tivoli Storage Manager's database protection capabilities for Microsoft Exchange to support a "near zero-impact" backup process. The product also works with IBM SAN Volume Controller in a virtualized storage environment and provides a unique feature leveraging SVC's FlashCopy function to allow for Instant Restores.

Tivoli Storage Manager for Copy Services takes advantage of a supported storage system's inbuilt snapshot capabilities, together with the Volume Shadowcopy Service provided in Microsoft Windows 2003 to provide fast online backups of Exchange databases. The snapshot backups can be retained on disk, or optionally copied to Tivoli Storage Manager for longer term storage. Because the snapshot operation itself is rapid, the impact of backup on the database is minimal. To further reduce the overhead of backup, the copy operation to Tivoli Storage Manager can optionally be performed by a separate server — known as an offloaded backup server.

If the database has to be restored, it can use the backup on the snapshot (target) volumes, or a backup stored on the Tivoli Storage Manager server. If the snapshot and original database are using SAN Volume Controller, then an "Instant Restore" is possible, where the target volumes are copied directly back to the source volumes, using the SVC's volume-level copy facility (FlashCopy).

Tivoli Storage Manager for Copy Services requires Tivoli Storage Manager for Mail Data Protection for Exchange as a prerequisite, which provides the command-line and GUI for VSS backup operations.

For more details, see the IBM Redbook, *Using IBM Tivoli Storage Manager to Back Up Microsoft Exchange with VSS*, SG24-7373.

12.6.1 IBM Tivoli Storage Manager for Space Management

IBM Tivoli Storage Manager for Space Management, also known as hierarchical storage management or HSM (see Figure 12-12) provides hierarchical storage management to automatically migrate rarely accessed files to alternate storage without disrupting the most frequently used files in local storage. The files are automatically migrated to lower storage according to the applicable migration policy. When they are required by applications or users, they are recalled transparently without administrators and users tasks.

Some percentage of data is inactive, that is, it has not been accessed in weeks, if not months. Tivoli Storage Manager for Space Management can automatically move inactive data to less-expensive offline storage or near-line storage, freeing online disk space for more important active data. Tivoli Storage Manager for Space Management frees administrators and users from manual file system pruning tasks, and defers the necessity to purchase additional disk storage, by automatically and transparently migrating rarely accessed files to Tivoli Storage Manager storage, while the files most frequently used remain in the local file system.

Figure 12-12 shows a screen capture of the Tivoli Storage Manager for Space Management interface.

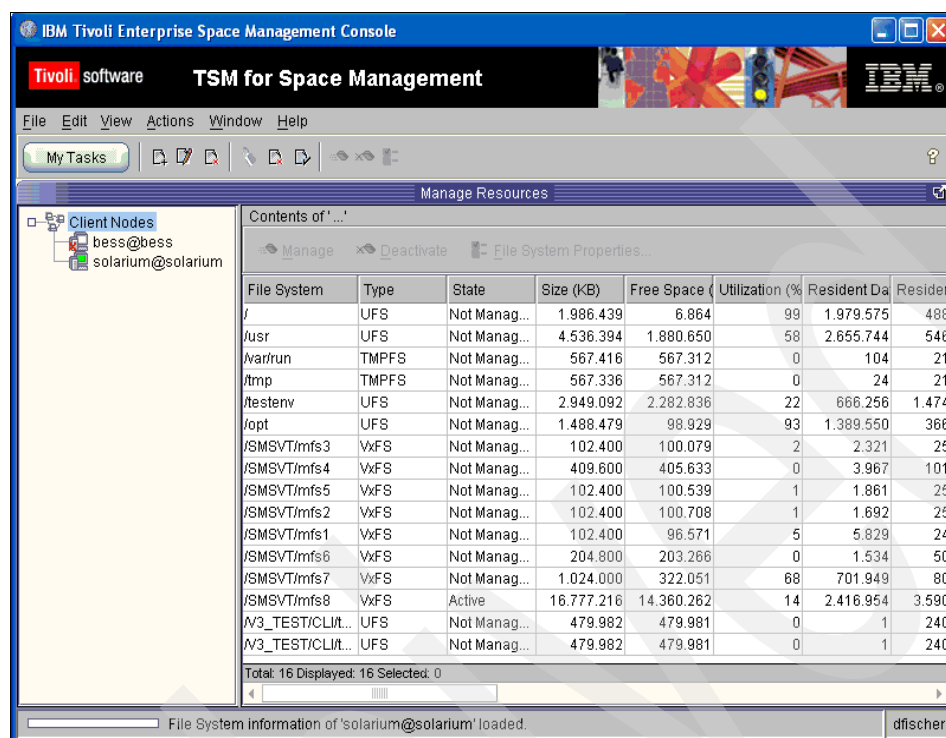


Figure 12-12 IBM Tivoli Storage Manager for Space management

Files are migrated by Tivoli Storage Manager for Space Management from the original file system to storage devices connected to a Tivoli Storage Manager server. Each file is copied to the server and a stub file is placed in the original file's location. Using the facilities of storage management on the server, the file is placed on various storage devices, such as disk and tape.

There are two types of HSM migration: automatic and selective.

Automatic migration

With automatic migration, Tivoli Storage Manager for Space Management monitors the amount of free space on the file systems. When it notices a free space shortage, it migrates files off the local file system to the Tivoli Storage Manager server storage based on the space management options that have been chosen. The migrated files are replaced with a stub file. Tivoli Storage Manager for Space Management monitors free space in two ways: threshold and demand.

Threshold

Threshold migration maintains your local file systems at a set level of free space. At an interval specified in the options file, Tivoli Storage Manager for Space Management checks the file system space usage. If the space usage exceeds the high threshold, files are migrated to the server by moving the least-recently used files first. When the file system space usage reaches the set low threshold, migration stops. Threshold migration can also be started manually.

Demand

Tivoli Storage Manager for Space Management checks for an out-of-space condition on a file system every two seconds. If this condition is encountered, Tivoli Storage Manager for Space Management automatically starts migrating files until the low threshold is reached. As space is freed up, the process causing the out-of-space condition continues to run. You do not receive out-of-space error messages while this is happening.

Selective migration

Tivoli Storage Manager for Space Management can selectively migrate a file immediately to the server's storage. As long as the file meets the space management options, it is migrated. The file does not have to meet age criteria, nor does the file system have to meet space threshold criteria.

Pre-migration

Migration can take a long time to free up significant amounts of space on the local file system. Files have to be selected and copied to the Tivoli Storage Manager server, (perhaps to tape) and a stub file must be created in place of the original file. To speed up the migration process, Tivoli Storage Manager for Space Management implement a pre-migration policy.

After threshold or demand migration completes, Tivoli Storage Manager for Space Management continues to copy files from the local file system until the pre-migration percentage is reached. These copied files are not replaced with the stub file, but they are marked as pre-migrated.

The next time migration starts, the pre-migrated files are chosen as the first candidates to migrate. If the file has not changed since it was copied, the file is marked as migrated, and the stub file is created in its place in the original file system. No copying of the file has to happen, as the server already has a copy. In this manner, migration can free up space very quickly.

Recall

Recall is the process for bringing back a migrated file from Tivoli Storage Manager to its original place on the local file system. A recall can be either transparent or selective.

Transparent

From a user or running process perspective, all of the files in the local file system are actually available. Directory listings and other commands that do not require access to the entire file appear exactly as they would if the HSM client was not installed. When a migrated file is required by an application or command, the operating system initiates a transparent recall for the file to the Tivoli Storage Manager server. The process temporarily waits while the file is automatically copied from the server's storage to the original file system location. Once the recall is complete, the halted process continues without requiring any user intervention. In fact, depending on how long it takes to recall the file, the user might not even be aware that HSM is used.

After a recall, the file contents are on both the original file system and on the server storage. This allows Tivoli Storage Manager for Space Management to mark the file as pre-migrated and eligible for migration unless the file is changed.

Selective

Transparent recall only recalls files automatically as they are accessed. If you or a process have to access a number of files, it might be more efficient to manually recall them prior to actually using them. This is done using *selective recall*.

Tivoli Storage Manager for Space Management batches the recalled file list based on where the files are stored. It recalls the files stored on disk first, then recalls the files stored on sequential storage devices, such as tape.

Advanced transparent recall

Advanced transparent recall is available only on AIX platforms. There are three recall modes: normal, which recalls a migrated file to its original file system, migrate-on-close, and read-without-recall.

Migrate-on-close

When Tivoli Storage Manager for Space Management uses the migrate-on-close mode for recall, it copies the migrated file to the original file system, where it remains until the file is closed. When the file is closed and if it has not been modified, Tivoli Storage Manager for Space Management replaces the file with a stub and marks the file as migrated (because a copy of the file already exists on the server storage).

Read-without-recall

When Tivoli Storage Manager for Space Management uses read-without-recall mode, it does not copy the file back to the originating file system, but passes the data directly to the requesting process from the recall. This can happen only when the processes that access the file do not modify the file, or, if the file is executable, the process does not execute the file. The file does not use any space on the original file system and remains migrated (unless the file is changed; then Tivoli Storage Manager for Space Management performs a normal recall).

Backup and restore

Tivoli Storage Manager for Space Management is not a replacement for backup — it enables extension of local disk storage space. When a file is migrated to the HSM server, there is still only one copy of the file available, because the original is deleted on the client and replaced by the stub. Also, Tivoli Storage Manager for Space Management maintains only the last copy of the file, giving no opportunity to store multiple versions. Tivoli Storage Manager for Space Management allows you to specify that a file is not eligible for HSM migration unless a backup has been made first with the Tivoli Storage Manager backup-archive client. If the file is migrated and the same Tivoli Storage Manager server destination is used for both backup and HSM, the server can copy the file from the migration storage pool to the backup destination without recalling the file.

Archive and retrieve

The Tivoli Storage Manager backup-archive client enables you to archive and retrieve copies of migrated files without performing a recall for the file first, providing the same Tivoli Storage Manager server is used for both HSM and the backup archive. The file is simply copied from the HSM storage pool to the archive destination pool.

Platform support

Platform support for various different platforms and file systems, includes AIX JFS, AIX GPFS, Linux GPFS, HP, and Solaris. A detailed list can be found at:

<http://www.ibm.com/software/tivoli/products/storage-mgr-space/platforms.html>

For more information about IBM Tivoli Storage Manager for Space Management, see the following Web site:

<http://www.ibm.com/software/tivoli/products/storage-mgr-space/>

12.6.2 Tivoli Storage Manager for HSM for Windows

IBM Tivoli Storage Manager for HSM for Windows is a companion solution to IBM Tivoli Storage Manager for Space Management in the Windows environment. It tracks low-activity, inactive files, parts of NTFS files or complete NTFS file systems and migrates the data to lower-cost media. A migrated file leaves a small piece of the file (stub file) on the Windows system, with sufficient metadata so that the name, directory path, owner, creation date, last access date, and last modification date are all visible. Files are migrated transparently to Windows users and application; this means that Windows users see and access migrated files like any file physically stored on the file system. The only difference is that when a migrated file is opened, it is transparently retrieved and copied back to the local Windows system.

Figure 12-13 illustrates how the HSM client acts as a Tivoli Storage Manager client exploiting the Tivoli Storage Manager client's archiving API. Migrated files from the HSM client are stored in *archive* pools on the Tivoli Storage Manager server, not HSM pools. Migrated files can be sent via a variety of LAN and LAN-free methods for storage in a variable cost hierarchy of storage managed by the Tivoli Storage Manager server.

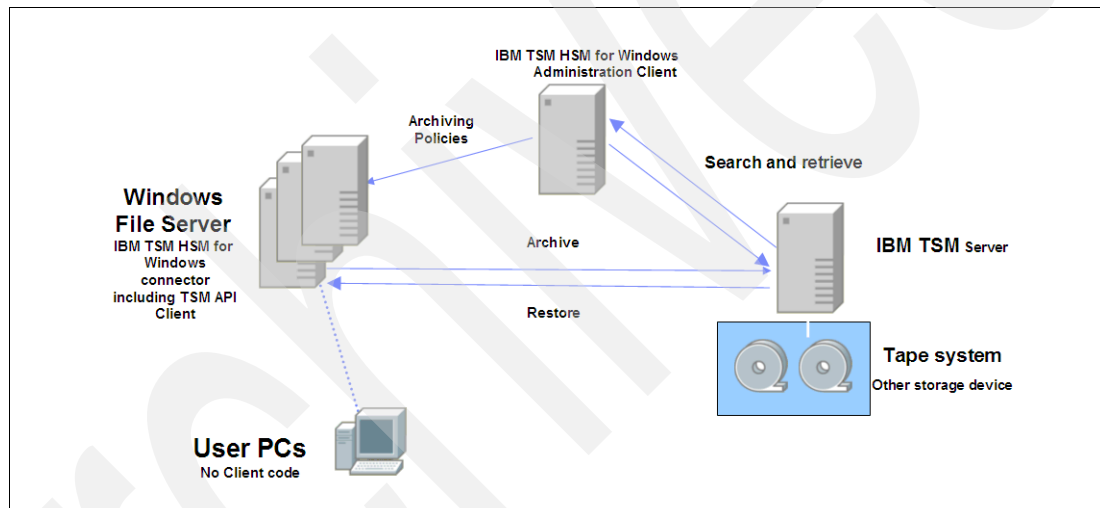


Figure 12-13 Tivoli Storage Manager HSM for Windows architecture

The HSM client supports NTFS file systems under Windows 2000 with Service Pack 3 and later, and Windows 2003. Windows NT® and FAT partitions are not supported.

See the IBM Redbook, *Using the Tivoli Storage Manager HSM Client for Windows*, REDP-4126, for more details.

12.6.3 Tivoli Continuous Data Protection for Files

In many companies, about 60-70% of corporate data resides on desktops, mobile computers, and workstations, which are rarely backed up or not backed up at all. Almost half of small and medium sized business admit to having no formal data protection process. One reason for this lack of protection is that these kinds of workstations might not always be connected to the network backbone, making it difficult to run scheduled traditional backups. Or, for smaller companies with no dedicated IT department, implementing backup might be seen as too complex. Virus and data corruption on file servers is increasing, yet traditional backup solutions might not offer sufficiently granular point in time recovery capabilities. In many cases, the classic nightly scheduled backup is not sufficient to protect what might be the most important data of all — what the user is working on right now.

Consider a typical “road warrior” who is working on critical presentations for the next day’s client call from the airport or hotel. How can this data be protected? The solution for this situation is to ensure recoverability through the automated creation, tracking, and vaulting of reliable recovery points for all enterprise data.

With all these issues in mind, Tivoli Continuous Data Protection for Files is designed to provide simple, effective, and efficient data protection and integrity. Tivoli Continuous Data Protection for Files is a real-time, continuous data-protection solution for mobile computers, workstations, and personal computers. It is specifically designed to work well even if network connections are intermittent. But Tivoli Continuous Data Protection for Files also provides continuous protection for file servers, reducing or eliminating backup windows and the amount of data potentially lost in a failure.

Tivoli Continuous Data Protection for Files can back up the most important files the moment they change instead of waiting for a scheduled backup. Non-critical files are backed up periodically on a scheduled basis. It works in the background, much like a virus scanner, and is therefore totally transparent to the user.

Since Tivoli Continuous Data Protection for Files has a single end-point architecture, there is no requirement for additional components, for example, a server component. It only requires a single installation on the system with files to be protected.

Tivoli Continuous Data Protection for Files keeps the protected instances of files in their natural format and does not modify them or encode them in a proprietary format. The advantage of maintaining files in their native format is that they are directly accessible and available by any application.

To protect files and make them available for date-based restore, Tivoli Continuous Data Protection for Files creates up to three separate copies of files:

- ▶ On local disk for protection, even when not connected to a network
- ▶ On a network file system for remote machine protection
- ▶ On a IBM Tivoli Storage Manager server for use in more sophisticated enterprises

Table 12-1 demonstrates the differences between Tivoli Continuous Data Protection for Files and traditional backup approaches.

Table 12-1 Comparison of Tivoli Continuous Data Protection for Files and traditional backup solutions

	Tivoli Continuous Data Protection for Files	Traditional backup solutions
When to protect	Continuous for highly important files, scheduled for others	Scheduled, full system
How to detect	Journal-based on <i>all</i> file systems	Journal-based on some file systems
Where copies are stored	Disk only, locally or remote; Tivoli Storage Manager	Typically on tape
Storage format	Left “native”, online as files	Wrapped into a proprietary format
Management / administration complexity	Simplified per-client administration only	Client-server concept; server component typically more expensive/complex

So, overall, Tivoli Continuous Data Protection for Files provides simple, effective and real-time file protection for:

- ▶ Accidental file deletion
- ▶ File corruption
- ▶ Unwanted file alteration
- ▶ Disk crashes
- ▶ Other unforeseen disasters

How does Tivoli Continuous Data Protection for Files work?

Figure 12-14 gives a general overview on how Tivoli Continuous Data Protection for Files works.

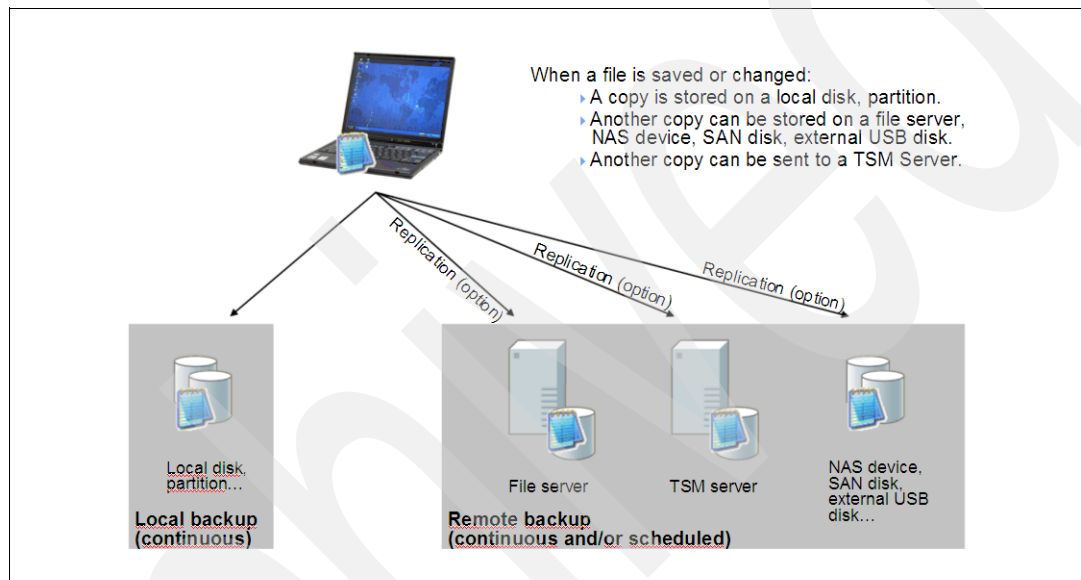


Figure 12-14 General overview of Tivoli Continuous Data Protection for Files

Whenever a file is changed or created, Tivoli Continuous Data Protection for Files notices it. If this file type is tagged as a high-priority continuous type (using settings, such as a Microsoft Word or PowerPoint® file), an immediate copy is made into a designated backup area (a separate directory tree) on the local disk. Tivoli Continuous Data Protection for Files can store many versions of each file (typically up to 20) subject to a configurable “pool size”. When the pool is full, the oldest copies (versions) are removed to make room for newer ones.

The same file can also be sent to a remote storage area, such as a file server, NAS device, or SAN disk for off-machine protection. If the remote file server is not currently available (perhaps due to not being in the network at the time), then the changed file is recorded and sent as soon as the network appears to be functioning. The files sent to the remote file server in this mode have only a single instance stored (that is, not versioned), since they are versioned locally.

Another copy of the file can be sent to a Tivoli Storage Manager server, as Tivoli Continuous Data Protection for Files has built-in Tivoli Storage Manager support. Traditionally, Tivoli Storage Manager is often used in larger business environments. Those clients might find Tivoli CDP for Files useful as a real-time client solution for mobile computers and workstations, yet still want most of the protected data to ultimately be managed by a Tivoli Storage Manager server.

If scheduled protection has been enabled, then all other “non-important” changing files are recorded by Tivoli Continuous Data Protection for Files and queued for transmission to the remote file server based on the interval that has been selected. When the interval expires, Tivoli Continuous Data Protection for Files copies all of the changed files to the remote file server, or wait if the file server is not currently available.

All those types of protection offered by Tivoli Continuous Data Protection for Files (continuous or scheduled, local or remote) can be easily configured by the user in any combination. This allows the user to protect his assets in a highly flexible manner.

For more information about Tivoli Continuous Data Protection for Files, see *Deployment Guide Series: Tivoli Continuous Data Protection for Files*, SG24-7235, and the Web site:

<http://www.ibm.com/software/tivoli/products/continuous-data-protection/>

12.6.4 IBM System Storage Archive Manager

IBM System Storage Archive Manager (formerly IBM Tivoli Storage Manager for Data Retention) extends IBM Tivoli Storage Manager functionality to help meet regulatory retention requirements.

IBM System Storage Archive Manager facilitates compliance with the most stringent regulatory requirements in the most flexible and function-rich manner. It helps manage and simplify the retrieval of the ever increasing amount of data that organizations must retain for strict records retention regulations.

Your content management and archive applications can utilize the IBM Tivoli Storage Manager API to apply business policy management for ultimate deletion of archived data at the appropriate time.

Figure 12-15 gives an overview of the functionality.

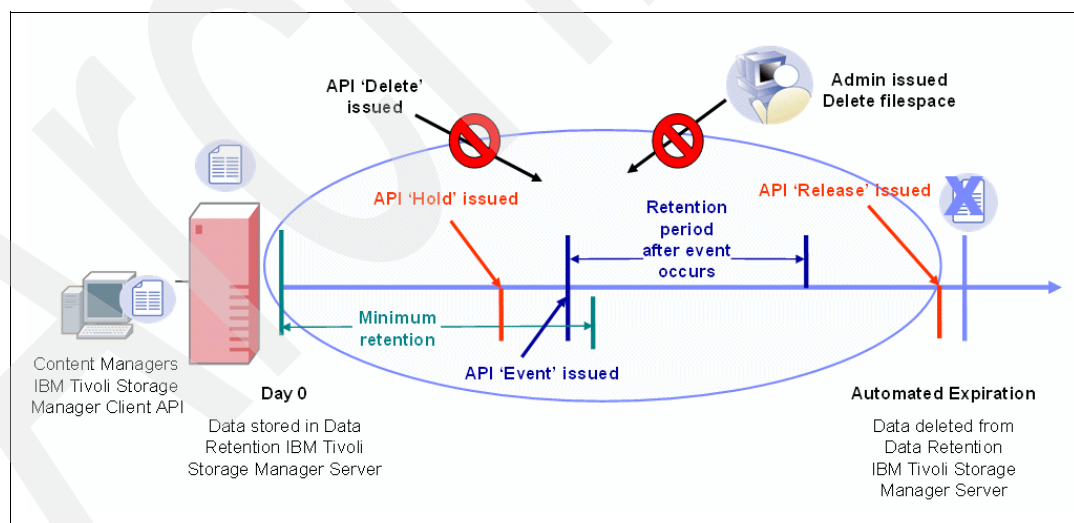


Figure 12-15 IBM System Storage Archive Manager functionality

Tivoli Storage Manager's existing policy-based data management capabilities helps organizations meet many of the regulatory requirements of various government and industry agencies. But some new regulations require additional safeguards on data retention. System Storage Archive Manager enhances the already function-rich data retention capabilities of Tivoli Storage Manager Extended Edition to provide data retention policies that help meet these new regulations.

Figure 12-16 illustrates a more detailed function of the software package.

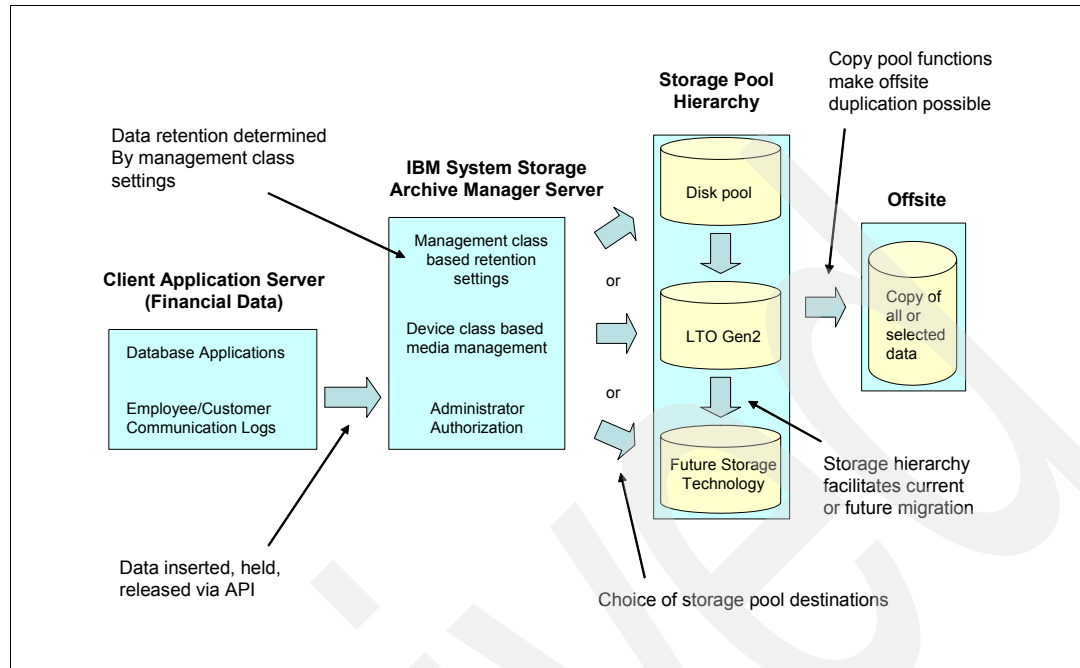


Figure 12-16 IBM System Storage Archive Manager function

Data Retention Protection

Data Retention Protection is controlled via the Tivoli Storage Manager archive client or Tivoli Storage Manager API by a variety of content management or archive applications, such as DB2 Content Manager CommonStore. System Storage Archive Manager makes the deletion of data before its scheduled expiration extremely difficult. Short of physical destruction to storage media or server, or deliberate corruption of data or deletion of the Tivoli Storage Manager database, System Storage Archive Manager does not allow data on the storage managed by the Tivoli Storage Manager Extended Edition server to be deleted before its scheduled expiration date. Content management and archive applications can apply business policy management for ultimate expiration of archived data at the appropriate time.

Event-based retention management

Should an event occur within an organization that requires some regulatory agency intervention, System Storage Archive Manager can suspend the expiration and deletion of data for a specified length of time.

For more information about IBM System Storage Archive Manager, see the following Web site:

<http://www.ibm.com/software/tivoli/products/storage-mgr-data-reten/>

12.6.5 LAN-free backups: Overview

SAN technology provides an alternative path for data movement between the Tivoli Storage Manager client and the server. Figure 12-17 shows an example of a simple SAN configuration. The solid lines indicate data movement, while the broken lines indicate movement of control information and metadata.

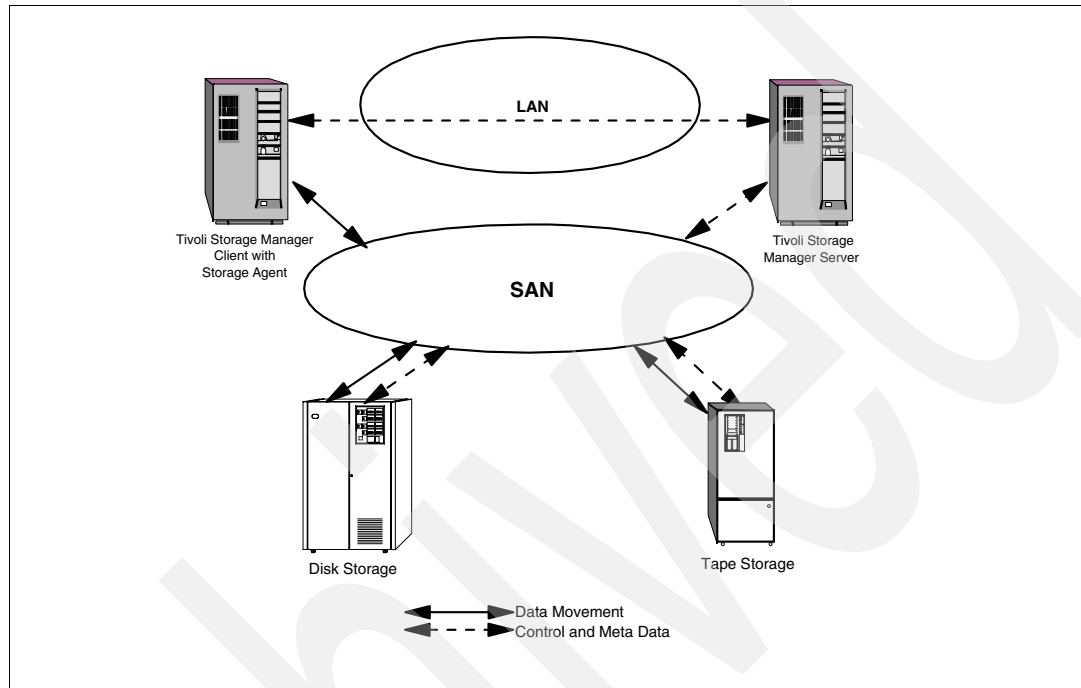


Figure 12-17 SAN data movement

Shared storage resources (disk, tape) are accessible to both the client and the server through the Storage Area Network. This gives the potential to off-load data movement from the LAN and from the server processor, allowing for greater scalability. This potential is achieved by the use of managing software, such as the Storage Agent in Tivoli Storage Manager for Storage Area Networks.

LAN-free backup and restore

Tivoli Storage Manager for Storage Area Networks is a feature of Tivoli Storage Manager that enables LAN-free client-data movement. This feature allows the client system to directly write data to, or read data from, storage devices attached to a storage area network (SAN), instead of passing or receiving the information over the network, making more network bandwidth available for other uses.

A LAN-free backup environment is shown in Figure 12-18. As shown in this example, the LAN is only used for metadata sent between the Tivoli Storage Manager client and the server. Also, the Tivoli Storage Manager server is relieved of the task of writing the backup data as in a traditional LAN backup. The Tivoli Storage Manager client reads that data, then instead of sending it over the LAN, it writes the data over the SAN.

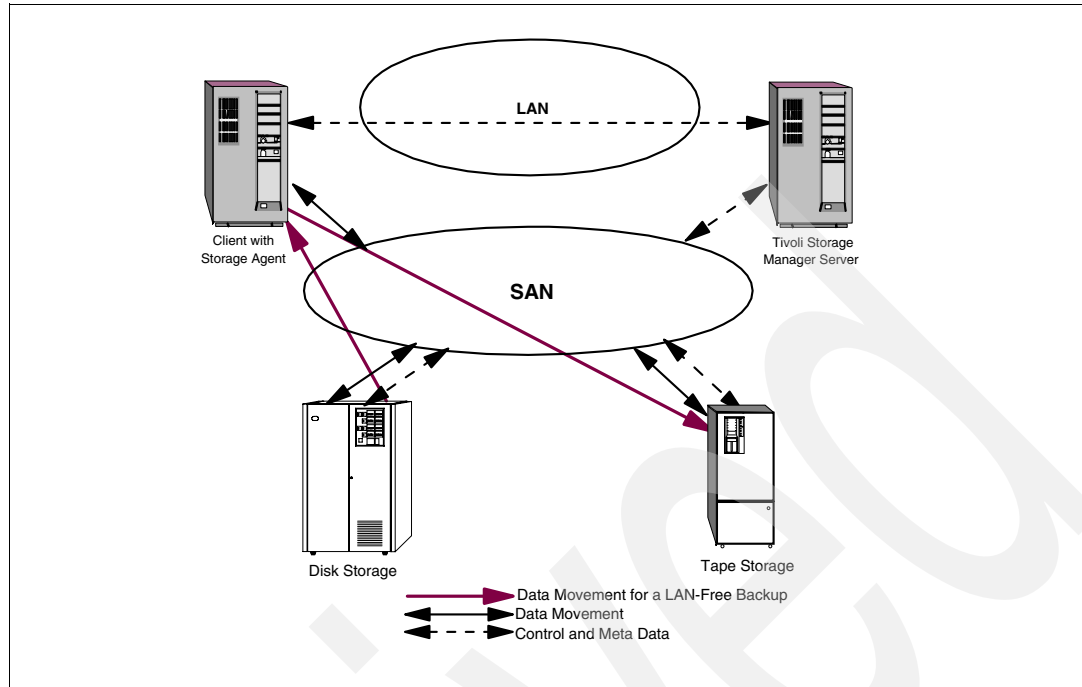


Figure 12-18 LAN-free backup

12.7 Bare Machine Recovery

In this section we discuss the various Bare Machine Recovery (BMR) options that integrate with Tivoli Storage Manager. The central focus of Bare Machine Recovery is recovering the operating system (OS) of a hardware system that cannot boot due to hardware failure or software corruption, corruption of the operating system by a virus, erasure or damage to files due to human error, and many other reasons. In addition to this main focus, we point out where cloning and migrating to different hardware is possible.

There are several Bare Machine Recovery solutions available on the market each with varying degrees of automation and robustness. The focus of this section is on those solutions which directly integrate with IBM Tivoli Storage Manager. These solutions are Cristie Bare Machine Recovery, Tivoli Storage Manager support for Microsoft Automated System Recovery (ASR), and Tivoli Storage Manager for System Backup and Recovery.

The advantage of using solutions that integrate with Tivoli Storage Manager, is that there is no additional hardware required to support the Bare Machine Recovery and much of the Tivoli Storage Manager interface and automation is used.

12.7.1 Cristie Bare Machine Recovery (CBMR)

Cristie Data Products integrates with Tivoli Storage Manager to provide a Bare Machine Recovery solution known as Cristie Bare Machine Recovery (CBMR). CBMR is a software package which automates bare metal recovery of a system to a new hard disk drive or array.

CBMR operating system support

The combined functionality of CBMR and Tivoli Storage Manager enables recovery of Windows NT, 2000, XP, 2003, x86 Linux, Solaris, and HP-UX operating system to a new disk drive. CBMR uses a special installation CD to boot during a restore, a configuration file from a floppy disk, memory key, network share, or from the operating system backup, and an operating system backup stored in the Tivoli Storage Manager server. This is illustrated in Figure 12-19.

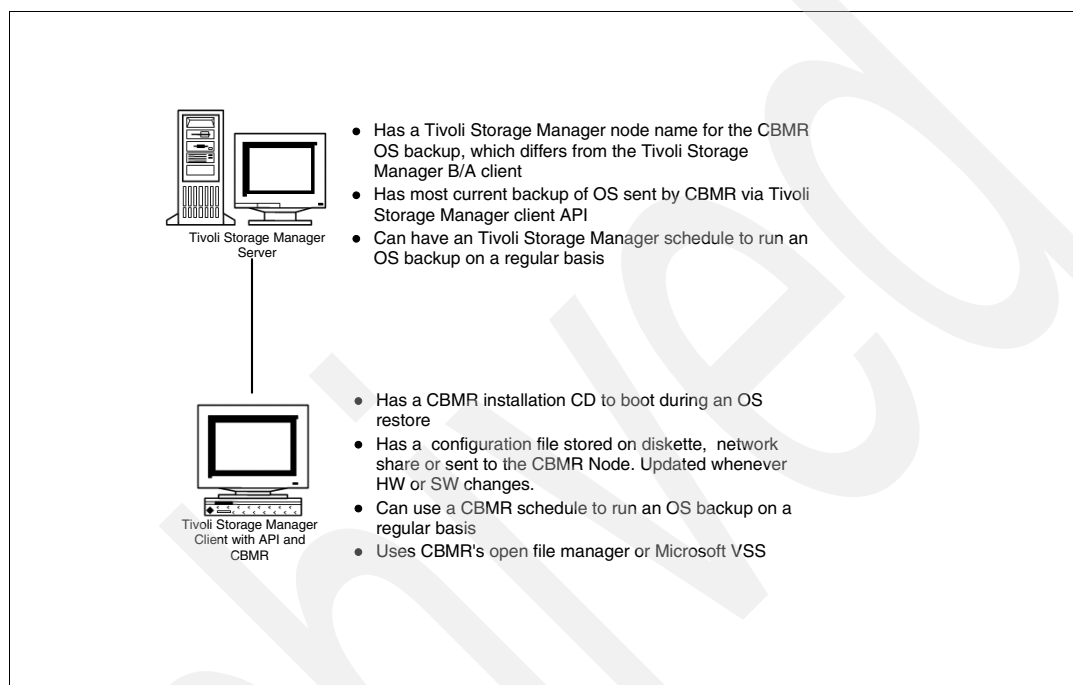


Figure 12-19 Sample Cristie configuration

CBMR advantages

CBMR's integration into Tivoli Storage Manager provides several advantages. These include:

- ▶ Reduced hardware requirements, other solutions require a separate image server.
- ▶ Simplified administration, Tivoli Storage Manager handles tape management and backup version control.
- ▶ Live backup of the operating system, with CBMR's Open File Module, backups can be made to Tivoli Storage Manager at any time.
- ▶ Dissimilar hardware restore for Windows.

Latest CBMR features

At the time of writing, CBMR V5 for Windows is the latest release. New features include:

- ▶ Windows PE or Linux recovery CD can be used
- ▶ Support 64-bit Windows
- ▶ System State support
- ▶ Full Citrix and Microsoft Terminal Server support
- ▶ Support for Windows installed on a Dynamic Volume which is also mirrored.
- ▶ Full Windows dynamic disk support

For the steps to set up your systems to prepare for system recovery and more information about Cristie Bare Machine Recovery, with Tivoli Storage Manager, see the paper:

http://www.cristie.com/fileadmin/DATA/Download/cbmr/Documents/Guide_to_using_CBMR_4.30_with_ITSM_-_update_2.pdf

Also refer to this Web site:

<http://www.cbmr.info>

For more information about supported environments for Cristie Bare Machine Recovery, see:

<http://www.ibm.com/software/tivoli/products/storage-mgr/cristie-bmr.html>

12.7.2 Tivoli Storage Manager support for Automated System Recovery

IBM Tivoli Storage Manager currently supports Microsoft Automated System Recovery on Windows Server 2003 and Windows XP. Automated System Recovery (ASR) is a feature built into Windows XP and 2003. With ASR you can create full operating system backups that can be used to restore a system under two conditions:

- ▶ If the operating system is still functioning, but a complete system restore is required to return to a previous state
- ▶ If your system has experienced a catastrophic failure that keeps your system from booting

ASR restore should be used only as a last resort when other system recovery methods have failed. For example, if possible, first, try Safe Mode Boot or booting with the Last Known Good Configuration feature. For more information about these methods refer to the *Microsoft Knowledge Base* at:

<http://support.microsoft.com/>

The purpose of ASR is to recover the boot and system drives plus system state, not the application and user data. The application and user data is recovered using the Tivoli Storage Manager Backup/Archive client.

ASR is not meant for cloning. The target system must be identical with the exception of hard disks, video cards, and network interface cards.

Tivoli Storage Manager interface to Automated System Recovery

IBM Tivoli Storage Manager includes a direct interface for use of the Windows Automated System Recovery feature. A brief description of some of the changes in this interface includes:

- ▶ **backup asr** — This command generates Automated System Recovery (ASR) files required by the Tivoli Storage Manager server to restore the files necessary during ASR recovery mode.
- ▶ **asrmode** — This option specifies whether to perform a restore operation in system ASR recovery mode.
- ▶ **restore asr** — This command restores the Automated System Recovery (ASR) files to a specified location.

A sample panel of Windows XP recovery by using Automated System Recovery with IBM Tivoli Storage Manager is shown in Figure 12-20.

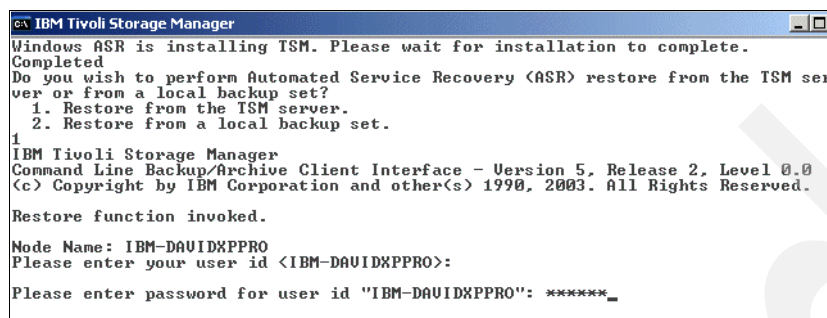


Figure 12-20 Windows XP recovery by using Automated System Recovery with IBM Tivoli Storage Manager

For more information about using IBM Tivoli Storage Manager with Automated System Recovery, see the IBM Redpaper, *IBM Tivoli Storage Manager: Bare Machine Recovery for Microsoft Windows 2003 and XP*, REDP-3703.

12.7.3 Tivoli Storage Manager for System Backup and Recovery

IBM Tivoli Storage Manager for System Backup and Recovery is based on a former IBM Global Services offering known as SysBack™. It offers the following features:

- ▶ It is a comprehensive system backup, restore, and reinstallation utility for AIX.
- ▶ It allows for backup and recovery options that range from:
 - A single file, to a full system Bare Machine Recovery
 - A local system or a remote system in the network
- ▶ It provides the ability to clone an AIX system backup to a wide range of System p servers.

An example of a SysBack/Tivoli Storage Manager configuration is illustrated in Figure 12-21.

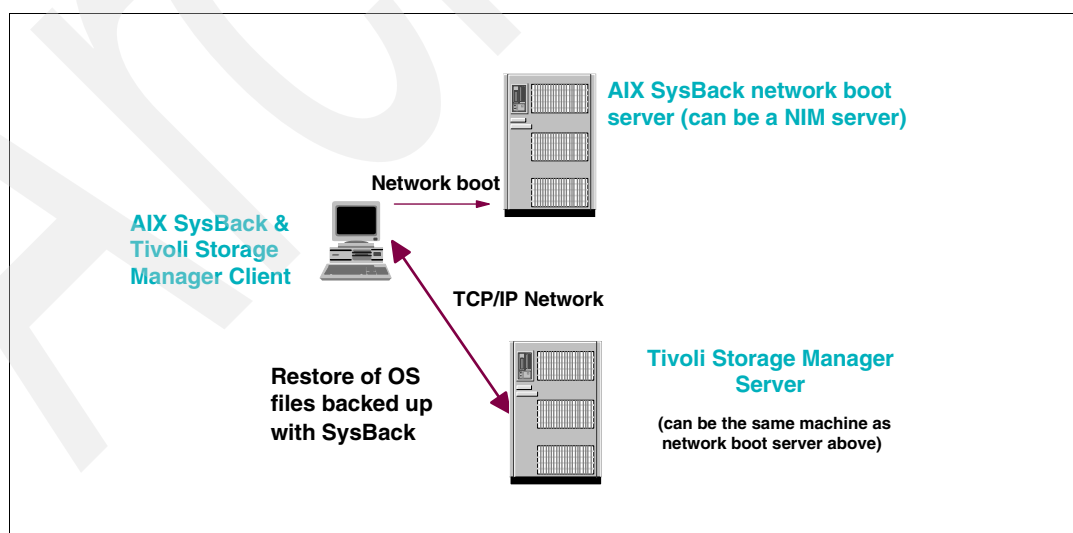


Figure 12-21 Sample SysBack/Tivoli Storage Manager integration

A sample panel of IBM Tivoli Storage Manager for System Backup and Recovery is shown in Figure 12-22.

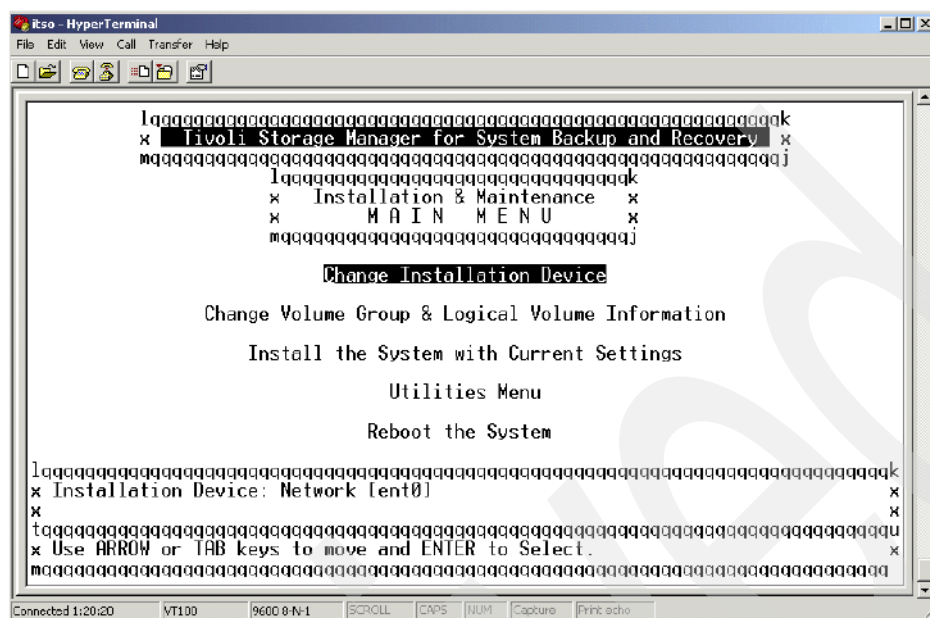


Figure 12-22 IBM Tivoli Storage Manager for System Backup and Recovery

For more information about Tivoli Storage Manager for System Backup and Recovery, refer to *IBM Tivoli Storage Manager: Bare Machine Recovery for AIX with SYSBACK*, REDP-3705 and to the Web site:

<http://www.ibm.com/software/tivoli/products/storage-mgr-sysback/>

12.8 DFSMS family of products

DFSMS is a z/OS software suite that automatically manages data from creation to expiration in the z/OS environment. DFSMS provides:

- ▶ Allocation control for availability and performance
- ▶ Backup/recovery and disaster recovery services
- ▶ Space management
- ▶ Tape management
- ▶ Data migration
- ▶ Archiving
- ▶ Copy services
- ▶ Reporting and simulation for performance and configuration tuning
- ▶ Storage pools with automatic, policy-based assignment of data, effectively tiered storage
- ▶ Automated policy-based data migration
- ▶ Information Lifecycle Management, from categorization, data movement to different tiers, to expiration

12.8.1 DFSMSdfp (data facility product)

DFSMSdfp provides storage, data, program, and device management functions and DFSMS Copy Services capabilities. DFSMSdfp is the heart of the storage management subsystem; it provides the logical and physical input and output for z/OS storage, it keeps track of all data and programs managed within z/OS, and it provides data access both for native z/OS applications and other platforms such as LINUX, AIX/UNIX, the Windows family, or System i.

Data management

DFSMSdfp stores and catalogs information about DASD, optical, and tape resources so that it can be quickly identified and retrieved from the system. You can use the Catalog Search Interface, now part of DFSMSdfp, to access the catalog.

Device management

DFSMSdfp defines your input and output devices to the system and controls the operation of those devices in the z/OS environment.

Storage management

DFSMSdfp includes ISMF, an interactive facility that lets you define and maintain policies to manage your storage resources.

Program management

DFSMSdfp combines programs into executable modules, prepares them to run on the operating system, stores them in libraries, and reads them into storage for execution.

12.8.2 DFSMSdfp Copy Services

The IBM Data Facility Storage Management Subsystem (DFSMS) SDM (System Data Mover) Copy Services are enterprise-level data duplication and disaster recovery capabilities which include: z/OS Global Mirror (XRC, Extended Remote Copy), Coupled z/OS Global Mirror, Metro Mirror (PPRC), SnapShot, CC (Concurrent Copy), and FlashCopy.

The continuous availability/disaster recovery capabilities of Copy Services are explained in detail in 7.5, “Advanced Copy Services for DS8000/DS6000 and ESS” on page 267 and following sections. These include z/OS Global Mirror (XRC), Metro Mirror (Synchronous PPRC), Global Copy (PPRC-XD), Global Mirror (Asynchronous PPRC), and FlashCopy.

Concurrent Copy (CC)

CC enables users to copy or dump data while applications are updating the data. Both IMS and DB2 users can use CC for their backups. CC is supported on the IBM ESS, DS6000, DS8000, and on most other vendor storage systems that support the OS/390 and z/OS environments.

SnapShot

The SnapShot function enables clients to produce almost instantaneous copies of data sets, volumes, and VM minidisks, all without data movement. SnapShot is automatically initiated by DFSMSdss on storage subsystems with the SnapShot feature.

12.8.3 DFSMShsm (Hierarchical Storage Manager)

DFSMShsm is an optional feature that provides space, availability, and tape mount management functions in a storage device hierarchy for both system-managed and non-system-managed storage environments. DFSMShsm provides automation of storage management tasks, which improves the productivity by effectively managing the storage devices.

Data management

DFSMSHsm is a disk storage management and productivity tool that manages low-activity and inactive data. It provides backup, recovery, migration, and space management functions as well as a full function disaster recovery facility, Aggregate Backup and Recovery Support (ABARS). DFSMSHsm improves disk use by automatically managing both space and data availability in a storage hierarchy.

A new Compatible Disk Layout for volumes under Linux on S/390 allows DFSMSHsm to perform dump and restore from OS/390 or z/OS operating systems.

Availability management

Through management policies, HSM automatically ensures that a recent copy of required disk data sets exists and can be restored as required.

Space management

Improves disk storage subsystem utilization by automatically deleting unnecessary data sets and migrating unused data sets to less expensive media so it reduces the total cost of disk storage ownership.

See the IBM Redbook, *DFSMSHsm Primer*, SG24-5272, for more details.

12.8.4 DFSMSHsm Fast Replication

DFSMSHsm Fast Replication provides DFSMSHsm management for the use of volume-level fast replication. Fast replication is made possible by exploiting the FlashCopy capability of the IBM Enterprise Storage Server and DS Series.

With this capability, a set of storage groups can be defined as a copy pool. The volumes in this pool are processed collectively, creating, via fast replication, backup versions that are managed by DFSMSHsm. Recovery can be performed at the volume or copy pool level.

While the focus of the DFSMSHsm Fast Replication function is its interaction with DB2 to back up and recover databases and logs, its implementation and benefit is not DB2-specific. DFSMS provides a volume-level fast replication that enables you to create a point-in-time backup that is managed by DFSMSHsm with minimal application outage. The Fast Replication function only supports the FlashCopy and SnapShot functions.

DFSMSHsm Fast Replication works on the volume level, not the data set level. When using Fast Replication, an application must be stopped or must flush its in-storage buffers to force the application data to be written to disk before the point-in-time backup by DFSMSHsm Fast Replication can be taken. Fast Data Replication occurs so fast because it builds a map, with pointers, to the source volume tracks or extents.

There is no longer a requirement to wait for the physical copy to complete before applications can resume the access to the data. Both the source and target data are available for read/write access almost immediately, while the copy process continues in the background. This process guarantees that the contents of the target volume are an exact duplicate of the source volume at that point in time. You can back up, recover to, and resume processing from that point in time. When the Fast Replication backup is complete, the application can again update its data sets.

See the IBM Redbook, *DFSMSHsm Fast Replication Technical Guide*, SG24-7069, for more details.

12.8.5 DFSMSHsm Disaster Recovery using ABARS

ABARS facilitates a point-in-time backup of a collection of related data in a consistent manner. This group of related data is defined to ABARS as an *aggregate*. An aggregate is user-defined and is usually a collection of data sets that have some sort of relationship, such as all the specific data sets required for an application (for example, payroll). ABARS backs up the data directly from disks or from HSM migration level 2 (tape), without the necessity for intermediate staging capacity. The backup copies are created in a device-independent format.

ABARS saves all the information associated with the backed up data, such as allocation information, catalog information, Generation Data Group base information and DFSMSHsm control data set records for migrated data sets; this information is necessary for restoring the backed up data. ABARS is generally used for Disaster Recovery or to move applications across non-sharing z/OS images. At the recovery site, ABARS is used to recover the application's data sets.

ABARS and Mainstar

IBM also provides the following software products from Mainstar Software Corporation to enhance your ABARS environment:

- ▶ Automated Selection and Audit Process (ASAP):

ASAP provides a set of tools for storage administrators and applications development personnel implementing or managing an installation's recovery strategy. The primary function of ASAP is to automatically identify the files that must be restored in a disaster, populate the ABARS selection data set, and automatically maintain the list of critical data sets as the application changes. ASAP is a companion product to IBM DFSMSHsm ABARS.

- ▶ Backup and Recovery Manager Suite: ABARS Manager:

ABARS Manager enhanced and simplified the ABARS function of DFSMSHsm. The focus of ABARS Manager was to provide easy aggregate recovery, online monitoring of the ABARS process, selective data set restore, online and batch reporting and additional functionality not provided by DFSMSHsm ABARS.

Backup and Recovery Manager Suite: ABARS Manager is a companion product to DFSMSHsm ABARS. See the IBM Redbook, *DFSMSHsm ABARS and Mainstar Solutions*, SG24-5089, for more details.

DFSMS/zOS Network File System (NFS)

The z/OS Network File System (NFS) is a network file system product that brings IBM system-managed storage to the network environment. It lets you optimize efficiency in a distributed network while still capitalizing on the capacity, security and integrity of z/OS multiple virtual storage (MVS).

z/OS NFS is used for file serving (as a data repository) and file sharing between platforms supported by z/OS. NFS can be used to remotely access both MVS data sets and UNIX hierarchical mounted file systems (HFS) files. These remote MVS data sets or Open HFS files are mounted from the mainframe to appear as local directories or files on the client system. This brings the resources of an MVS system, such as system-managed storage, high-performance storage access, file access security, and centralized data access, to client platforms.

In addition, the z/OS NFS server can access DASD data sets other than virtual storage access method (VSAM) linear data sets, multi-volume non-striped data sets, and generation data sets. However, z/OS NFS does not support tape and optical drives.

See the IBM Redbook, *Data Sharing: Using the OS/390 Network File System*, SG24-5262, for more details.

DFSMSdss (Data Set Services)

Data Set Services (DFSMSdss) is an optional component of DFSMS (Data Facility Storage Management Subsystem), and is used to:

- ▶ **Move and replicate data:**

DFSMSdss offers powerful, user-friendly functions that let you move or copy data between volumes of like and unlike device types. It can also copy data that has been backed up.

- ▶ **Manage storage space efficiently:**

DFSMSdss can increase performance by reducing or eliminating DASD free-space fragmentation.

- ▶ **Backup and recover data:**

DFSMSdss provides host system backup and recovery functions at both the data set and volume levels. It also includes an enhanced Stand-alone Restore Program that restores vital system packs during disaster recovery - without a host operating system.

- ▶ **Convert data sets and volumes:**

DFSMSdss can convert data sets and volumes to system-managed storage, or return data to a non-system-managed state as part of a backed up.

DFSMSdss is an external interface to Concurrent Copy, SnapShot, and FlashCopy:

- ▶ **FlashCopy and SnapShot:**

DFSMSdss automatically detects if a storage system supports the FlashCopy or the Snapshot function and if possible executes this function.

- ▶ **Concurrent Copy (CC):**

Concurrent copy is both an extended function in cached IBM Storage Controllers, and a component of DFSMSdss. CC enables you to copy or dump data while applications are updating the data. Both IMS/ESA® and DB2 can use CC for their backups.

- ▶ **DFSMSdss Stand-Alone Services:**

The DFSMSdss Stand-Alone Services function is a single-purpose program designed to allow the system programmer to restore vital system packs during Disaster Recovery without having to rely on a z/OS environment.

DFSMSrmm (Removable Media Manager)

Removable Media Manager (DFSMSrmm) is a functional component of DFSMS for managing removable media resources including automatic libraries (such as the IBM Virtual Tape Server), for the z/OS environment.

DFSMSrmm is an integral part of DFSMS and is shipped as part of DFSMS with the z/OS operating system. DFSMSrmm can manage all the tape volumes and the data sets on those volumes. It protects tape data sets from being accidentally overwritten, manages the movement of tape volumes between libraries and vaults over the life of the tape data sets, and handles expired and scratch tapes, all according to user-defined policies. DFSMSrmm also manages other removable media that are defined to it. For example, it can record the shelf location for optical disks and track their vital record status.

DFSMSrmm volume retention and movement are specified interactively with ISPF panels. This allows authorized application owners to alter existing values without contacting the tape librarian. DFSMSrmm is functionally compatible with existing tape management systems and runs in parallel during conversion.

For more details, see the IBM Redbook, *DFSMSrmm Primer*, SG24-5983.

12.9 IBM System Storage and TotalStorage Management Tools

This is a collection of products designed to provide a comprehensive menu of solutions to meet the storage management requirements of z/OS clients. These tools are positioned to complement and extend the storage management capabilities of DFSMS that we described in the preceding sections. They also provide storage administrators with a more simple, straightforward environment, increasing their productivity while reducing mistakes and potential outages. The tools includes the products described in the following sections.

12.9.1 DFSMStvs Transactional VSAM Services

DFSMS Transactional VSAM Services (DFSMStvs) is a feature that enables batch jobs and CICS online transactions to update shared VSAM data sets concurrently. Multiple batch jobs and online transactions can be run against the same VSAM data sets and DFSMStvs helps ensure data integrity for concurrent batch updates while CICS ensures it for online updates.

DFSMStvs is designed to offer the following benefits:

- ▶ Contributes to the reduction or elimination of the batch window for CICS applications and other VSAM applications by allowing concurrent batch and online updating of VSAM recoverable data. Delivers increased system availability with simpler, faster, and more reliable recovery operations for VSAM storage structures.
- ▶ Simplifies scheduling batch jobs, because multiple batch jobs that access the same files, can be run concurrently on one or more z/OS images in a Parallel Sysplex instead of serially on one image.
- ▶ Delivers increased system availability with simpler, faster and more reliable recovery operations for VSAM storage structures.
- ▶ Provides the ability to share VSAM data sets at record-level with integrity and commit and rollback functions for non-CICS applications.
- ▶ Offers backup-while-open backups to be taken using DFSMSdss and DFSMSHsm.
- ▶ Enables batch applications to use the same forward recovery logs.
- ▶ Helps enable 24x7 CICS Transaction Server (TS) applications.
- ▶ Supports System-Managed CF Structure Duplexing.

DFSMStvs is an extension to the function provided by VSAM RLS and requires RLS for accessing VSAM recoverable data sets. VSAM RLS (and therefore, DFSMStvs) requires the use of a system coupling facility for caching, locking, and buffer coherency.

For more details, see the IBM Redbook, *DFSMStvs Overview and Planning Guide*, SG24-6971.

12.9.2 CICS/VSAM Recovery

CICS/VSAM Recovery (CICSVR) recovers lost or damaged VSAM data sets for OS/390 and z/OS clients. It determines which CICS logs and VSAM backups are required and then constructs the recovery jobs. CICSVR uses an ISPF dialog interface that complies with the Common User Access® (CUA®) standards.

Starting from V3.3, CICSVR also provides recovery capabilities for batch VSAM updates. Two kinds of recovery are available in the batch environment, forward recovery and backout.

To support these, CICSVR VSAM batch logging performs two types of logging, forward-recovery logging and undo logging (for batch backout). CICSVR performs batch logging without making any updates to your batch jobs.

CICSVR forward-recovery logging records all changes made to eligible VSAM files. In the event of catastrophic loss of data, you can run the forward recovery process for the data sets updated by the batch job.

Without a forward recovery or back out utility, lost or damaged data sets must be recovered manually. If each update has to be reentered manually, errors can increase, which threatens the integrity of your VSAM data. If you have high transaction volumes between backups, you want to avoid long work interruptions due to failed or lost data sets. CICSVR minimizes your recovery labor and can eliminate data errors in your recovery of data sets.

CICSVR helps you recover your CICS VSAM data sets in these situations:

- ▶ Physical loss or damage of VSAM data
- ▶ Loss of logical data integrity because a failure occurred during CICS transaction back out or during emergency restart back out
- ▶ Loss of logical data integrity because a CICS application incorrectly updated your VSAM spheres

CICSVR recovers all updates made before the problem occurred, and if either dynamic transaction back out (DTB) or emergency restart back out has failed, CICSVR backs out partially completed transactions.

These features are supported in all CICS versions:

- ▶ Determines the backups and logs required for recovery automatically.
- ▶ Has an ISPF dialog interface that is CUA compliant.
- ▶ Performs forward recovery of VSAM data sets.
- ▶ Accepts backup-while-open (BWO) backups taken while a data set is open and being updated by CICS.
- ▶ Recovers multiple data sets in a single run.
- ▶ Logs batch updates without changes to batch jobs, allowing VSAM recovery.
- ▶ Monitors batch jobs and removes VSAM updates made by failing job.

For more details, see the IBM Redbooks, *CICSVR Update for Release 3.2*, SG24-7022, and *CICSVR Usage Guide*, SG24-6563.

12.9.3 IBM TotalStorage DFSMSHsm Monitor

The IBM TotalStorage DFSMSHsm Monitor is a client/server solution that provides automation support and real-time monitoring of major HSM functions, thresholds, and events on multiple systems. It enables you to keep informed as to the status of HSM activities such as:

- ▶ Real-time monitoring of HSM automatic migration, backup, and dump functions
- ▶ Volume data set processing and monitoring
- ▶ Recall and processing statistics, such as CPU utilization, tape usage, and tasking levels

For more details, see the IBM Redbook, *DFSMS Optimizer: The New HSM Monitor/Tuner*, SG24-5248.

12.9.4 Tivoli Storage Optimizer for z/OS

IBM Tivoli Storage Optimizer for z/OS automates the monitoring of storage resources across the enterprise which increases the efficiency of existing storage to yield a higher return on storage investment. Storage Optimizer automatically discovers the z/OS environment. Storage administrators can then use the Java-based interface to set up rules, thresholds, responses, and corrective actions to manage disks resources.

With extensive filtering capabilities, trend analysis, and root-cause analysis, administrators and operators can focus on storage issues before they become storage problems. While Storage Optimizer manages resources automatically, based on site policies and best practices established by the administrator, computer operators and administrators are free to address other critical problems.

Through a Java-based graphical user interface, Storage Optimizer generates a true graphical visualization of storage assets and current status. This includes extensive drill-down capabilities to the actual metrics concerning the storage object in question, giving advanced support for storage management. In addition, Storage Optimizer provides an optional ISPF-based interface to assist administrators in generating JCL for storage management tasks.

Storage Optimizer operates 7x24x365. Based on site policies, Storage Optimizer regularly collects measurements about your z/OS storage environment, then compares the results to previous metrics and identifies exceptions. Whether you tell Storage Optimizer to simply notify you when exceptions occur or have it automatically carry out corrective actions is your call. Storage Optimizer can do as much or as little as you require.

With Storage Optimizer, a storage administrator can perform all of the following tasks:

- ▶ View volume pool reports and trends for any system from one logon.
- ▶ View usage history reports that are categorized by volume, by storage group or by application group.
- ▶ Respond to alerts for conditions that have exceeded thresholds, such as out of space and media errors.
- ▶ Make changes to the system environment, such as adding devices, formatting volumes, and cataloging volumes and data sets.
- ▶ Trends for any pool or storage group. Select a view for that particular day and receive a trend report for a specific volume over time. You can also view trends for the total system, meaning all volumes.

- ▶ Perform dataplex maintenance. You can see collected information, then dynamically sort, scroll, or enter commands to manage storage groups or individual volumes. For example, if you select a storage group and it contains 100 volumes, values are displayed for each of those volumes, and you can perform actions against a specific volume.
- ▶ Set thresholds for every pool or storage group, and receive alerts when those thresholds are crossed.
- ▶ Use the GUI or the optional ITSO Workbench to perform DASD and storage management functions. There is no requirement to remember command sequences. For example, if you want to analyze a disk, you indicate what you want to do, and Storage Optimizer creates the JCL for you.
- ▶ Schedule a variety of actions to occur at a predetermined time or interval, such as volume space defragmentation and deletion of old data. This process is easily established through Storage Optimizer's GUI front-end.
- ▶ Perform a number of catalog management functions, such as define catalogs, manage GDGs (generation data groups), or diagnose catalog structure.
- ▶ Send the query ACTIVE command to HSM dynamically and receive an immediate reply.
- ▶ View and create numerous reports through the easy-to-use GUI:
 - Forward all messages (alerts, events) to a database or to a comma separated value (CSV) file for use in Excel® or other spreadsheet program.
 - Dump current contents of a message browser to a CSV file, which enables you to create numerous configurable, filtered reports.
 - Dump table views to a CSV file. Table views can include any number of different kinds of information including, but not limited to, summaries and statistical reports.

12.9.5 Mainstar Mirroring Solutions/Volume Conflict Rename (MS/VCR)

MS/VCR is an innovative tool that solves the cloned data access dilemma, giving users access to data sets on target volumes created with FlashCopy or SnapShot by renaming and cataloging them. MS/VCR resolves catalog conflicts of like-named data sets that are created by *cloning* a volume. It also solves internal conflicts that are created when copying a volume to a different VOLSER. It is designed to minimize the time required to rename and catalog target volumes, data sets and to manage DB2 subsystem cloning. The window of time required to copy and rename a target data set is a critical factor for many clients. Additional benefits of MS/VCR are:

- ▶ Fixes volume conflicts (VTOC, VTOCIX, and VVDS) and then renames and recatalogs the data sets.
- ▶ Allows users to change any qualifier or add/delete qualifiers with the Rename capability.
- ▶ Tracks the behind-the-scenes FlashCopy, enables a user to know when they can start another cloning operation or withdraw from the current copy.
- ▶ Provides automatic pairing of volume characteristics.
- ▶ Provides the capability to FlashCopy or SnapShot by storage groups.
- ▶ Enables unique support for users to keep only the data sets that are really wanted at the conclusion of a volume level FlashCopy or SnapShot. This can be helpful to clients who require data set level support under FlashCopy V1.
- ▶ Enables several SMS options to determine how the SMS class constructs are applied to their cloned data sets.

- ▶ Enables with the catalog search facility the user to easily identify, in advance, all ICF catalogs that are necessary for the rename and recatalog function. In addition it can clean ICF catalogs from orphaned entries and check for catalog integrity.
- ▶ Clones DB2 subsystems and by using an online or offline method can customize another DB2 subsystem to access the new volumes. It can manage the number of data sharing members and go from a sharing to non-sharing DB2 environment.
- ▶ Simulates Copy, Rename and DB2 cloning operations, in this way you can check the results of your activities before executing the real commands.

For more details, see the IBM Redbook, *Mainstar MS/VCR: The Tool for Efficient Device Cloning*, SG24-6860.

12.9.6 Mainstar Catalog RecoveryPlus

Catalog RecoveryPlus provides repair, backup and restore capability, as well as a forward recovery facility for BCS (Basic Catalog Structure) and VVDS (VSAM Volume Dataset) catalog structures, using SMF data merged with backup records. It updates the BCS and VVDS from the time of the backup to the current time, additionally, Catalog RecoveryPlus is rich in valuable features including:

- ▶ A backup and restore facility for VSAM KSDS (Key-Sequence Data Set) files.
- ▶ Expanded Diagnose and Mergecat facilities.
- ▶ Additional commands and facilities that ensure the fast and easy recovery and daily maintenance of ICF catalogs.
- ▶ A backup and restore facility for VSAM KSDS files that is specifically designed for situations where the index is damaged and the records must be unloaded from the current file.
- ▶ An expanded Diagnose facility that enables frequent health checks of the BCS, VVDS, and VTOC structures. This facility contains an audit-check and Fix capability to re-synchronize entries that are in error.
- ▶ A fast and safe Mergecat facility that enables moving or copying of individual BCS entries, alias groups of BCS entries, or entire catalogs.
- ▶ A Master Catalog Alias audit facility that compares master catalogs to determine if they contain consistent and appropriate alias entries. This facility also contains a Fix capability to resynchronize the catalogs if necessary.
- ▶ A Zap command that provides an easy-to-use AMASPZAP-like facility to delete, print, and patch BCS and VVDS record structures.
- ▶ A powerful Superclip facility that allows users to change the VOLSER of online DASD volumes in a single command, without requiring the user to unload and reload all data on the volume.
- ▶ An **Alter** command that provides several BCS/VVDS facilities that include VOLSER changes to BCS records and BCS back pointer correction in the VVDS.
- ▶ An **Explore** command that provides a powerful search facility with extensive data set naming and file attribute filters to identify data sets of interest.
- ▶ A Simulate feature on several commands that enables users to check critical catalog changes in advance to determine the effect of the command before it is run.
- ▶ A Journalling facility for critical commands that can enable a Restart or Backout option in the event of a failure or interruption during the execution of the command.

- ▶ **Reorg and Repair while open** allow you to reorg or fix your catalog related problems without quiescing the applications accessing the catalog. The catalog can remain open to the Catalog Address Space from any sharing system.
- ▶ Maps the internal structure of VSAM files or catalog to identify potential Control Area, Key or free space problems.

For more details, see the IBM Redbook, *ICF Catalog Backup and Recovery: A Practical Guide*, SG24-5644.

12.9.7 Mainstar FastAudit/390

FastAudit/390 is a *set* of products that offers comprehensive and extremely fast audit support in the following areas: DFSMSHsm Migration, Backup, and Offline Control DataSets (MCDS, BCDS, and OCDS), Tape, and Catalog and DASD.

Catalog and DASD Audit

Catalog and DASD Audit evaluates all master and user catalog data set entries, and reconciles them against all DASD VTOC information about currently mounted volumes. This audit finds such problems as unused alias entries, unconnected nodes, uncataloged data sets, and data sets cataloged to the wrong volumes. It also provides tools to automatically reconcile multiple master catalogs.

Tape Audit

Tape Audit correlates entries in the TMC (Tape Management Catalog) with the ICF (Integrated Catalog Facility) catalog status, assuring consistency within and between these critical structures. Successful tape management depends on the accurate synchronization of TMC information and ICF catalog information. Tape Audit's superior performance makes this reconciliation fast and painless.

HSM FastAudit

FastAudit evaluates the Migration, Backup, and Offline Control DataSets (MCDS, BCDS, and OCDS) to resolve structural and logical discrepancies that could have prevented migrated data from being recalled, or backup data from being restored. HSM FastAudit checks the ICF catalog relationships to assure that migrated data is cataloged, and that data cataloged as migrated is actually migrated. This function duplicates the DFSMSHsm command audit facilities, but provides improved performance and flexibility.

HSM FastAudit-MediaControls

HSM FastAudit-MediaControls provides the ability for you to be proactive and audit your HSM Migration and Backup tapes as you see fit. It works outside the HSM environment to correct the errors it finds and has documented bench speeds from 5 to over 180 times faster than the standard HSM Command audit mediacontrols.

For more details, see the IBM Redbook, *DFSMSHsm Audit and Mainstar: FastAudit/390*, SG24-6425.

12.9.8 Mainstar disaster recovery utilities

This is a set of products that complement and extend the value of DFSMSHsm's ABARS disaster recovery solution. ABARS is a DFHSM option described in 12.8.3, "DFSMSHsm (Hierarchical Storage Manager)" on page 432, that provides disaster backup at the application level. The Mainstar utilities add to the ABARS capability, making it more efficient and easier to use.

Mainstar Backup and Recovery Manager

Backup and Recovery Manager provides several benefits that enhance the value of DFSMSHsm's ABARS disaster recovery solution:

- ▶ All backup information and backup status is accessible through one central point.
- ▶ Backup status and details can be viewed online.
- ▶ Online access to the full activity log and all error messages, as well as a condensed status summary, makes ABARS management simple.
- ▶ ABACKUP and ARECOVERY activity is continuously monitored in real-time.
- ▶ Detail data set inventory provides recovery capacity requirements, recovery tape VOLSER, and other key planning and operational details.

Mainstar ASAP

ASAP is a sophisticated tool that automates the identification of critical application data sets required for Disaster Recovery. It uses as input data your procedures, schedule information, and SMF data and then creates and forwards this critical application data set list to the Backup and Recovery Manager for backup processing by ABARS.

Mainstar Catalog BaseLine

Catalog Baseline supports Disaster Recovery on z/OS and OS/390 systems with empty ICF catalogs by reducing catalog scrubbing and synchronization issues. Catalog BaseLine provides the facility to synchronize multiple Master Catalogs. A common problem during many disaster recoveries is invalid content in the production ICF master and user catalogs. A powerful and preferred strategy is to provide a complete, but empty, ICF user catalog environment with aliases intact, and let data recoveries repopulate the catalog contents. Catalog BaseLine allows users to implement this strategy.

Backup & Recovery Manager Suite: ABARS Manager

The focus of ABARS Manager was to provide easy aggregate recovery, online monitoring of the ABARS process. It automatically builds backup and recovery jobs keeping track of volume and dataset information. ABARS manager is a centralized management point for your ABARS backup, restore and query operation. It has both an interactive powerful ISPF interface and a batch interface for routine operations. It has selectable features, such as Incremental ABARS and CATSCRUB.

Incremental ABARS allows users to combine the disaster recovery solution provided by ABARS with a regular incremental backup strategy. Incremental ABARS allows user to combine their disaster recovery backup with their regular incremental backups. The benefits of this are: A reduced backup size, improved backup and recovery performance, and reduced resources to vault and maintain the backup. With Incremental ABARS, the value of ABARS is also expanded with the capability to provide a solution in a local recovery circumstance as well as in a disaster recovery situation.

CATSCRUB is a catalog tool able to clean up catalogs from unwanted entries and synchronize them with the recovered volumes. It accesses data directly and not by using IDCAMS services and in this way can reduce catalog fixing time.

Backup and Recovery Manager Suite: All/Star

All/Star is a tool that interfaces to all your non ABARS backups without JCL changes and lets you find the backup you require. It can inventory your backups, in this way you can easily find what is backed up, what is missing from your backup and what is being backed up by different utilities. It supports Innovation's Fast Dump Restore (FDR), SORT, DFSMSdss, DFSMSHsm Incremental, AUTODUMP, ARCINBAK, IDCAMS, IEBGENER, ICEGENER, and IEBCOPY utilities.

More details about the Mainstar Disaster Recovery utilities can be found at the Mainstar Web site:

<http://www.mainstar.com>

12.10 IBM TotalStorage Enterprise Tape Library (ETL) Expert

The ETL Expert collects capacity and performance information and stores it in a relational database. The ETL Expert is used to format the collected data into real-time viewable ETL asset, capacity, and performance reports. The reports contain key measurements regarding the Tape Library and Virtual Tape Server.

This innovative software tool provides administrators with the storage asset, capacity, and performance information they require to centrally manage IBM tape libraries (3494 and VTS) located anywhere within the enterprise. As a result, the ETL Expert can simplify the management of tape libraries while helping to increase administrator productivity and reduce overall storage costs.

Asset management reports

The ETL asset management report details the type and number of ETLs installed. The report displays:

- ▶ The assigned library name and library ID
- ▶ The library type (composite, VTS, library)
- ▶ The library associated with a VTS
- ▶ The model
- ▶ The number of physical drives (installed and available)

Capacity reports

The ETL capacity reports are only available for a VTS. They are not available for composite libraries or native libraries. Two capacity reports are provided. The Capacity Summary report displays the following for each VTS:

- ▶ The number of logical volumes
- ▶ The average size of a logical volume in megabytes
- ▶ The amount of data associated with the VTS in gigabytes
- ▶ The free storage associated with the VTS in gigabytes
- ▶ The number of stacked volumes (scratch and private)
- ▶ The reclaim threshold
- ▶ The backstore compression ratio

From the summary report, a specific VTS can be selected for more detailed information that by default is displayed in an hourly format. The hourly cycle can be changed to daily, weekly, monthly, or real-time.

Performance management reports

The performance management reports are also available as either a Performance Summary for all VTSs or as an hourly Performance Report for a specific VTS. There are also performance reports for native libraries. Examples of generated performance information include:

- ▶ Average and maximum virtual mount times
- ▶ Average and maximum cache-miss virtual mount times
- ▶ Average and maximum fast-ready virtual mount times
- ▶ Cache miss percentage
- ▶ Least average age in cache
- ▶ Throttle values
- ▶ Data transfer statistics
- ▶ Average and maximum drives mounted
 - Total mounts
 - Mount time

VTS health monitor

The TotalStorage Expert also includes a feature called the VTS health monitor. The monitor enables administrators to view key statistics from a desktop window on an ongoing basis. It helps increase administrator productivity by tracking a variety of performance, asset, and capacity information. The monitor checks the status of the VTS by querying the TotalStorage Expert every ten minutes. Based on predetermined thresholds and values, the monitor displays a color coded signal to indicate the VTS health in normal (green), warning (yellow), critical (red) status.

12.11 IBM TotalStorage Specialists

The IBM TotalStorage Specialists are a family of tools used for reporting and monitoring IBM tape products.

12.11.1 IBM TotalStorage Tape Library Specialist

This tool does not provide reports, but can be used for online queries about the status of the Peer-to-Peer VTS, its components, and the distributed libraries. In a Peer-to-Peer VTS configuration, you have two TotalStorage Specialists available: The TotalStorage Tape Library Specialist and the TotalStorage Peer-to-Peer VTS Specialist. You can access the Peer-to-Peer Specialist Web pages from the Tape Library Specialist and vice versa, as there is a link between the two products that enables you to switch between them seamlessly and hence easily find the required information. The TotalStorage Tape Library Specialist is the Web interface to the Library Manager and the VTS of each distributed library.

Library Manager information includes:

- ▶ System summary
- ▶ Operational status and operator interventions
- ▶ Component availability
- ▶ Performance statistics
- ▶ Command queue
- ▶ LAN hosts' status and LAN information
- ▶ VOLSER ranges and cleaner masks

12.11.2 IBM TotalStorage Peer-to-Peer VTS Specialist

The Peer-to-Peer VTS Specialist is a Web server that is installed in the Virtual Tape Controllers (VTC). You have to perform some installation steps to access the Web pages from your workstation browser. Although this IBM TotalStorage Specialist is not required for monitoring of the Peer-to-Peer VTS, we highly recommend that you install it, because it is a single source of current information about the complete hardware and logical components of the Peer-to-Peer VTS. VTS information includes:

- ▶ Active data and active data distribution
- ▶ Data flow
- ▶ Logical mounts per hour and mount hit data
- ▶ Physical device mount history
- ▶ Category attributes
- ▶ Management policies
- ▶ Real time statistics

Archived



Tape and Business Continuity

In this chapter we discuss the positioning of tape for Business Continuity. We also look at the recent tape technology advancements and the effect of these tape advancements affecting the tape positioning in Business Continuity.

The purpose is to help you to choose and position tape technology deployment in a Business Continuity perspective.

13.1 Positioning of tape in Business Continuity

Tape has been in the IT industry for more than 50 years. Tape has always been a key technology deployed in Business Continuity for backup, restore and archive. Automated tape libraries together with recent tape advancements enable tape to take part in the three BC segments.

Table 13-1 summarizes the positioning of IBM tape products and technologies in the BC segments. We expand on this table throughout this chapter.

Table 13-1 Tape positioning in BC Segments

	Tape drive	Tape library	Backup management software	BC solution	Availability impact for backup	Extent of tape automation	Expect RTO
Continuous Availability (CA)	<ul style="list-style-type: none"> ▶ LTO3 ▶ TS1120 	<ul style="list-style-type: none"> ▶ TS3500 ▶ 3494 ▶ TS7740 ▶ 3494 (VTS) 	<ul style="list-style-type: none"> ▶ DFSMS 	<ul style="list-style-type: none"> ▶ VTS PtP with GDPS ▶ TS7740 (Virtual Tape Grid) with GDPS 	Concurrent or non-concurrent	Automated Tape operation with failover and failback between sites capability	Within an hour
Rapid Data Recovery (RDR)	<ul style="list-style-type: none"> ▶ LTO3 ▶ TS1120 	<ul style="list-style-type: none"> ▶ TS3500 ▶ 3494 ▶ TS7740 ▶ 3494 (VTS) ▶ TS7510 ▶ TS3310 ▶ TS3210 ▶ TS3110 	<ul style="list-style-type: none"> ▶ DFSMS ▶ Tivoli Storage Manager 	<ul style="list-style-type: none"> ▶ TS7750 with DFSMS ▶ VTS with DFSMS ▶ TS3500 with DFSMS ▶ Tivoli Storage Manager with <ul style="list-style-type: none"> – TS3500 – TS7510 – TS3310 – TS3210 – TS3310 	Concurrent or non-concurrent	Automated Tape Operation, but without automated failover nor failback between site capability	Within a few hours
* Backup and Restore (B&R)	<ul style="list-style-type: none"> ▶ LTO3 ▶ TS1120 	<ul style="list-style-type: none"> ▶ TS3500 ▶ 3494 ▶ TS7740 ▶ 3494 (VTS) ▶ TS7510 ▶ TS3310 ▶ TS3210 ▶ TS3110 	<ul style="list-style-type: none"> ▶ DFSMS ▶ Tivoli Storage Manager 	<ul style="list-style-type: none"> ▶ TS7750 with DFSMS ▶ TS3500 with DFSMS ▶ TSM with <ul style="list-style-type: none"> – TS3500 – TS7510 – TS3310 – TS3210 – TS3310 	Non-concurrent	Automated or non-automated tape operation	More than a few hours, might be days

Note: * Backup and Restore - BC Segment might not require a tape library, although it is highly recommended.

13.2 Why tape requires a longer term commitment than disk

Even though tape technology is ubiquitous in the data center, it is a general misconception that commitment to tape technology is less important than disk. While disk and tape commitment are both important, in the perspective of technology change, you normally would commit to a specific disk technology for 3 to 5 years. For tape technology, on the other hand, a 10 to 15+ years commitment is not uncommon.

The main reasons for this are that tape data is normally expected to be retained for longer time periods, and also that changing tape technology is not so simple. You have to consider how can you move or migrate data that is stored on anywhere from hundreds to hundreds of thousands of existing tape cartridges.

This affects BC operations, when you have to retrieve production data that has been backed up or archived. Therefore, when the time comes to change tape technology, you must plan for it so that it does not affect BC.

13.2.1 Typical tape media migration plans

There are various strategies for migrating from one tape media to another.

Copy data from old to new tape technology format

In this strategy, copying tape data is labour intensive, time consuming, and costly — it can take months to complete the migration, depending on how many tape cartridges have to be migrated.

Add new technology but coexist with the old tape technology

In this strategy, all new data is saved to the new tape format, but existing data on the old media is retained for an extensive period of time - say 1 year to 7 years. The old tapes can be migrated over time or discarded when no longer required for retention.

- ▶ This allows old tape data to be restored or retrieved if required.
- ▶ However, it means drives and libraries have to be maintained which can read the old media (unless the new drives/libraries are read-compatible with the older media).
- ▶ Keeping old tape drive technology has issues of its own:
 - The drives and libraries might no longer be supported by the vendor.
 - It might not be possible to attach the devices to new servers.
 - The devices might no longer be supported by your application software.

Therefore, when choosing a tape technology, you have to think ahead about how easy it can be to ultimately move from that technology. The next section provides some guideline on choosing the right tape strategy.

13.3 What are the key attributes to a tape strategy?

IBM provides two current core tape media technologies - LTO (Linear Tape Open - currently in its third generation) and TS1120 (formerly known as 3592). Historically, IBM Magstar® 3590 tape technology has achieved wide acceptance in client environments.

Table 13-2 compares LTO and TS1120 key attributes to help you choose an appropriate tape strategy.

Table 13-2 Comparison of LTO3 and TS1120 tape technologies

Criteria	IBM LTO (Generation 3)	TS1120
Proprietary or Open Standard	Open Standard	Proprietary
Proven Technology	Yes	Yes
Performance	80 MBps	100 MBps
Capacity	400GB	500GB and 700GB
Reliability	Provide SARS support	Provide extensive SARs support
Scalability	Supported by TS3500 and TS7500	Supported by TS3500, TS7700 and TS7500
Investment Protection	<ul style="list-style-type: none"> ▶ Backward write compatible to last generation ▶ Backward read compatible to last two generation 	<ul style="list-style-type: none"> ▶ Backward write compatible to last generation ▶ Backward read compatible to last two generation ▶ PLUS, reformat media to improve media capacity to current
Application support	Support most of the backup popular management software	Support most of the backup popular management software
Simple to Manage	Support Policy Based tape management software like Tivoli Storage Management	Support Policy Based tape management software like Tivoli Storage Management and DFSMS
TCO	<ul style="list-style-type: none"> ▶ Reduction on power consumption with Sleep mode ▶ IBM unique compression implementation to store more data on same LTO media versus other vendors' LTO 	<ul style="list-style-type: none"> ▶ Provide IBM unique compression implementation ▶ Able to re-format same media of previous generation to TS1120 format, improved capacity by 67% without purchase new media ▶ Simplify the management of two generation technology to same format

Let us now consider these attributes in more detail.

13.3.1 Proprietary or Open Standard

Should you choose proprietary or open standard tape technology?

- ▶ An open standard tape technology gives more choices of vendor. It might also have wider application support.
- ▶ Proprietary is good for non-commodity usage, where tape technology development is not restricted by an Open Standard or a specific consortium. Therefore, applicable business technology which is required by the market can be quickly developed and integrated to provide easy business use. An example of this is the IBM TS1120 Enterprise Tape technology which provides encryption. The TS1120 provide seamless Encryption application use, where Encrypted Key Management can be flexibly deployed according to the client's business requirements.

13.3.2 Proven technology endorsement

What support is available for the tape technology?

- ▶ Since tape technology commitment is normally for around a ten year period, you have to ensure that the technology is around for a long time. Even though no one can be sure what will happen in the future, you should look at your chosen technology partners' track record of commitment to products over this time span.
- ▶ Your chosen tape partner should provide a clear roadmap and technology demonstration to demonstrate the roadmap.
- ▶ To commit to a tape technology, you want to ensure that this tape technology will be supported by most popular tape backup/management software besides your existing software. This should ensure that you have options to change software vendors or applications in the future.

13.3.3 Performance

What level of performance can you expect from the tape technology?

- ▶ To commit to a tape technology, you require to ensure that it can provide adequate performance. You might have different performance requirements for different applications.
 - For data streaming large applications, you require high throughput tape.
 - For writing smaller files, or applications where data is frequently retrieved, you require a tape drive that is more robust. IBM TS1120 (3592-E05) drive can perform virtual backhitch and high speed search with high resolution directory to facilitate the above tape application requirements. IBM LTO3 drives can perform backhitch.

13.3.4 Capacity

What is the capacity of the tape technology?

- ▶ You require a proven capacity scalability roadmap for the tape media; for example:
 - LTO second to third generation doubled capacity from 200GB to 400GB per cartridge
 - 3592-J1A to 3592-E05 increased capacity from 300 GB → 500 GB per cartridge

13.3.5 Reliability

What is the reliability of the tape technology?

- ▶ You have to ensure that the tape technology is reliable, especially as tape cartridge capacity scales to hundreds of GB and beyond. You have to feel confident that the tape vendor has a provide track record and technology demonstration for reliability to ensure:
 - That your data is securely protected
 - That your tape operation can be running on schedule without much impact due to reliability issue

13.3.6 Scalability

How scalable is the tape technology?

- ▶ Vertical and horizontal scalability is important:
 - Scaling up (horizontally) is to increase the overall capacity of the system. IBM provides a range of tape libraries to meet growing capacity requirements.
 - To scale horizontally means that there is backward compatibility for the media. Both LTO and TS1120 can read previous tape generations, and also, TS1120 can reformat older media to higher capacity.
- ▶ You should also look at your application support to see how tape media change can be accommodated. IBM Tivoli Storage Manager's tape pool concept allows multiple tape libraries with different tape media to coexist — and also provides automatic migration of data from one device to another.

13.3.7 Investment protection

How can you protect your investment in the tape technology?

- ▶ You want to make sure that your long-running tape investment is protected. This includes:
 - Proven investment protection in previous tape technology, therefore, your investment is likely protected after another 10 years.
 - You can scale up without changing tape technology.
 - Proven technology demonstration, that is, IBM demonstrated 1TB tape capacity and 8 TB tape capacity on April 5, 2002 and May16, 2006 respectively.
 - Support for new tape technology advancements, such as:
 - WORM tape support, such as LTO3 and TS1120
 - Encrypted tape support, such as TS1120
 - Virtual Tape Library support, such as TS7510 and TS7740

13.3.8 Application support

What application support is available for the tape technology?

- ▶ The tape technology should be supported by common applications. It should be easy to support different tape technologies, or generations within a technology (for example, LTO2 and LTO3).

Note: More discussion on the above tape technology advancement is covered in 13.6, “Recent tape technology advancements” on page 461

13.3.9 Simple to manage

Is the tape technology simple to manage?

- ▶ The tape technology must be simple to manage. If there are errors, you want to easily identify the problem — whether the source is the tape media, tape drive or another reason. Comprehensive tape error reporting is critical in today's IT operations. IBM tape technology provides comprehensive SARS (Statistical Analysis and Reporting System) to provide problem determination.

13.3.10 Total cost of ownership

What is the total cost of ownership for the tape technology?

- ▶ When calculating TCO of tape, make sure to consider:
 - Initial purchase cost plus ongoing maintenance cost
 - Power requirements, for example, different LTO vendors have different power consumption on their LTO drives
 - Floor space/footprint of tape libraries
 - Actual storage capacity on the tape — different LTO vendors use different compression methods. This can lead to a lower tape capacity.

See the IBM whitepaper, *Not All LTO Drives are Created Equal - LTO Data Compression* - by Glen Jaquette. For more detail, refer to the Web site:

http://www.ibm.com/servers/storage/tape/whitepapers/pdf/whitepaper_compression.pdf

In summary, to ensure that you have the right tape strategy for your business, you must map your business requirements, including your Business Continuity plan, with the foregoing important considerations.

13.4 Tape strategies

Tape (backup/restore, and archive/retrieve) is an integral part of business continuity, especially if you do not have a high-availability solution implemented at a remote site.

In this section we describe the various IBM tape options and tape recovery strategies that are available today.

13.4.1 Why tape backups are necessary

When data loss is experienced, for whatever reason, the impact can be catastrophic. Whether the data loss is caused by hardware failure, human error, or a natural disaster, no storage system is immune.

The risk of data loss can, however, be greatly reduced by planning and implementing a reliable, uncorrupted, tape backup strategy. The combination of increased options, enhanced functionality, and lower media costs of data storage will continue to keep tape backup at the forefront for older system replacement and upgrades as well as for new installations of backup systems.

Tape still makes more sense for backup than mirroring or other disk based options for protecting data integrity. If a mirrored server is attacked by a virus or worm, so is its replica. Any kind of disk storage can be affected on a real time basis as data is added, updated, deleted or corrupted. However, properly scheduled timely backups to tape allow users to undo all the damage that took place on a given day, or shorter time period (depending on the backup frequency).

Enterprise backup is no trivial task, especially in large multi-platform environments, and most organizations cannot even begin to measure the cost of lost data. So the value of a good tape backup and recovery strategy easily outweighs its purchase costs.

It is important to note that the many high-availability solutions today, including RAID disks, redundant power supplies, and ECC memory and clustering solutions, reduce the risk of downtime but do not prevent disasters from occurring. Additionally, since most enterprises are required by company or government regulation to keep multiple versions of data, the sheer volume of backing up multiple versions to disk can put a strain on any IT budget.

13.4.2 Tape applications

The major tape applications are backup, archive, batch, and HSM. Tape also plays an important part in system recovery and disaster recovery.

Backup

Backups of a company's data are like an insurance policy. You might not often require it, but when you do — you really must have it. There are several possible incidents that should be covered by your data backups:

- ▶ **Hardware failures of the active disk storage.** You can restore the lost data from the last point-in-time backup or in combination with more sophisticated software such as IBM Tivoli Storage Manager for Databases you can recover to within moments of an incident occurring.
- ▶ **Mistaken deletion or overwriting of files.** You can restore selected files/datasets to a specific point-in-time.
- ▶ **Logically corrupted data, such as by a virus or an application error.** You can restore the data to a specified point-in-time before the data corruption.

Regularly scheduled backups are key. In some cases data corruption might occur and go undetected for some time. Creating several backup versions of the data is the starting point for recovery from such a scenario.

Archive

Due to statutory or regulatory requirements, a company might have to archive, maintain, and reliably reproduce data for extended periods of time. This data might be considered active data and depending on the requirements, all or part of this data might have to be included in the disaster recovery storage pool.

Batch

Especially in mainframe environments, batch processing is still a part of the daily production cycle. In this environment, tape data is often active enterprise data, which must be included in the disaster recovery planning. The IBM Virtual Tape Server (VTS) as described later in this chapter is an ideal tool to optimize batch processing in mainframe environments. The automatic copy function of Peer-to-Peer VTS (PtP VTS), makes it an ideal disaster recovery solution for this kind of data.

Hierarchical storage management

Hierarchical storage management has been a basic function in mainframe environments and is becoming more prevalent in open systems. To reduce storage costs, hierarchical storage management can be used to automatically or selectively migrate files from high performance storage subsystems to lower-cost storage subsystems, usually tape. Often this data is still active data that might be recalled to the high performance storage subsystem at any time. This data must also be included in the disaster recovery planning. Applications like DFSMSHsm provide for efficient management of data in a hierarchy. DFSMSHsm also includes facilities for making simultaneous backups of data as it is migrated from disk to a lower cost storage media.

System recovery

Many Bare Machine Recovery solutions depend on tape for regularly scheduled backups of system files. This allows several versions of a server configuration to be saved, see 12.7, “Bare Machine Recovery” on page 427.

Disaster recovery

Depending on a systems RTO and RPO, it is often viable to use tape for disaster recovery. In this case backup tapes must be regularly taken off-site either physically or electronically, see 12.5.3, “Disaster Recovery for the Tivoli Storage Manager Server” on page 404, for methods of sending backups off-site electronically.

Conclusion: When designing a Business Continuity Plan, you must also consider tape data, because this tape data might not just be another copy of the original disk data.

There are many types of data on tape media, more than you would at first expect. Some or even all of this tape data might belong to the disaster recovery storage pool. The decision on impact tolerance of each application must be made independently of whether the data is stored on disk or tape.

13.5 Available tape technologies

Figure 13-1 shows a high level summary of the IBM Tape Portfolio.



Figure 13-1 IBM tape product portfolio

For details of the IBM System Storage tape product portfolio, see:

<http://www.ibm.com/servers/storage/tape>

Also, refer to the IBM Redbooks, *IBM System Storage Tape Library Guide for Open Systems*, SG24-5946 and *IBM TS3500 Tape Library with System z Attachment: A Practical Guide to TS1120 Tape Drives and TS3500 Tape Automation*, SG24-6789.

13.5.1 Tape automation

As you can see, you have a wide variety of options from which to choose with regard to backup and restore management. Single tape devices are adequate for small installations. This is a manual device that requires that an operator insert and remove tapes manually.

If you have a larger installation with several hosts and a large amount of data, then autoloaders and libraries are ideal for managing large backups that span several tape volumes and allow multiple servers to be backed up simultaneously.

Autoloaders

Autoloaders store and provide access to multiple tape cartridges through one or two drives using a robotic mechanism allowing for a hands-free backup and restore operation. Typical autoloaders have from 6 up to 44 slots.

Tape libraries

Libraries are offered in many sizes and allow many tape cartridges to be stored, catalogued, and referenced by multiple tape drives simultaneously through the use of robotic hardware. Tape libraries boost the performance and ease the headache of managing large numbers of backup volumes regardless of the type of tape media used. In many cases, human handling of the tapes is also eliminated which greatly increases the availability and reliability of the tape media. Another key point of tape libraries is that they are shareable. For example, the IBM tape libraries feature patented multi-path architecture, which enables a single tape library to be shared among different hosts and host types, and also among multiple homogenous or heterogeneous applications.

13.5.2 Tape drives

How many tape drive technologies are valid today? That really depends on your requirements. Wading through all the characteristics of every tape technology available today is an enormous task. Combine the available tape technology with the available autoloader and library options and the task becomes overwhelming. For example, IBM offers over twenty tape products many of which offer two or more tape options. The libraries also have options for the number of slots, number of drives, size of the input/output station for importing or exporting, and the list of options goes on. On top of all this there is the issue of compatibility. Does the product work with your software? Currently Tivoli Storage Manager supports over 700 devices for removable media. For a list of the devices supported by platform, see this URL:

<http://www.ibm.com/software/tivoli/products/storage-mgr-extended/platforms.html>

In the following sections we briefly review these topics:

- ▶ Tape drive characteristics
- ▶ Criteria for selecting your tape drive
- ▶ Tape cartridge usage considerations
- ▶ Tape drive technology

While these topics are important, in today's world, with the continual growth of data storage, make sure that the solution you choose also meets your requirements for:

- ▶ Reliability
- ▶ Support
- ▶ Scalability
- ▶ Compatibility

The last three points are very important for applications that process data on tape for more than just backup and restore, such as HSM, which migrates data for production use. These characteristics are also important for data reorganization tasks such as tape reclamation with HSM, Tivoli Storage Manager or VTS.

13.5.3 Criteria for selecting a tape drive

As we discussed earlier, there are a large number of tape solutions available. Deciding on tape technology normally is a long term decision and should be made carefully.

Here is a summary of the things you should consider:

- ▶ **Software:**
 - Check the compatibility list of your backup and recovery software to be sure that it supports the tape drive you are considering.
- ▶ **Application:**
 - Different tape based applications demand different behaviors of data streaming, quick load and positioning, high capacity, and high bandwidth.
- ▶ **Compatibility:**
 - Is the tape drive compatible with your servers?
 - Is the tape drive compatible with the SAN components (Host Bus Adapter, switches)?
 - Which Fibre Channel protocol does the tape drive use in a SAN-attached environment? Does the switch hardware support this protocol?
 - If purchasing a new tape drive technology, can the new tape drive read the older tape media, especially to be able to retrieve archived data? You might have to set up a data migration plan.
 - Check all dependencies between the disaster recovery infrastructure with the local site, especially if the disaster recovery site is provided by another company.
- ▶ **Bandwidth:**
 - How much data are you going to be backing up?
 - How much time do you have to back the data up?
 - The effective bandwidth in a client environment might be different than the possible bandwidth of the tape device.
- ▶ **Capacity:**
 - How much data do you require to backup?
 - What is the sum of all valid tape data? This determines the minimum number of cartridges and library storage cells. Tape reclamation methodologies require additional capacity for work space on the cartridges.
- ▶ **Reliability:**
 - What tape drive, cartridge, and robotic reliability levels are required to meet the service level agreements?
- ▶ **Attachments:**
 - Open systems:
 - SCSI or Fibre Channel

Be aware of the different SCSI speeds, interfaces (LVD/HVD) and the implications for the connection distance which is possible, and for device sharing.

For Fibre Channel, note the speed (1Gbps, 2Gbps, 4Gbps) and connections (GBIC, Small Form-factor Pluggable or Mini-GIBC (SFP)).
 - Mainframe:
 - FICON operates at 100 MBps, 200 MBps, or 400 MBps (bidirectional). FICON can go up to 250km without significant performance loss. Because of the higher transfer rates, higher channel throughput rates can be achieved as compared to ESCON. Therefore, depending on the individual requirements, 2 FICON channels can replace 4 to 8 ESCON channels.

13.5.4 Tape cartridge usage considerations

Tape cartridges have a long, but finite life expectancy. The life expectancy also varies widely between tape manufactured for the low end of the market and the high end of the market. Subjecting a low end type of tape to the usage characteristics of a high end tape, could lead to a failure during backup or recovery. When a tape cartridge is over used, the tape begins to break down, leaving particles that can contaminate or damage the tape system or other tape cartridges. Software such as Tivoli Storage Manager and tape management software have the ability to track tape usage and read/write errors. It is important to follow the recommended usage guide of each media type. Tapes that are involved in tape related errors, physically damaged, or exposed to extreme temperatures should be replaced immediately.

Note: Do not underestimate the impact of cartridge handling in manual environments. Inappropriate handling of cartridges can lead to tape cartridge or tape drive error or failure. For example, the improper storage of tapes at a temperature higher than recommended by the manufacturer can lead to intermittent errors which are hard to resolve. Another example of mishandling is the use of adhesive paper labels that are not approved for use on cartridges. These unapproved labels can come loose jamming the tape drive and possibly voiding the warranty on the drive and the library. Follow closely the cartridge handling instructions for your media type. You can find a detailed discussion of tape media handling in Chapter 2, “Overview of IBM Tape Technology” in the IBM Redbook, *IBM System Storage Tape Library Guide for Open Systems*, SG24-5946.

13.5.5 Tape drive technology

Although there are many technologies, only a few meet the demands of enterprise solutions requirements. Next we discuss the two primary solutions offered by IBM.

Linear Tape Open (LTO)

Linear Tape Open is, as the name implies, an open standard for linear tape storage. Proposed and developed initially by a consortium of Hewlett-Packard, IBM, and Seagate (now Quantum), LTO technology combines the advantages of linear multi-channel, bi-directional formats with enhancements in servo technology, data compression, track layout, and error correction code to maximize capacity, performance, and reliability.

LTO uses the Ultrium standard. Ultrium is a single-reel format targeted for users requiring ultra-high capacity backup, restore, and archive capabilities. The Ultrium format is currently on its third generation. A published road map of the capacity and performance characteristics of each generation is displayed below in Table 14-1. Note that although 2:1 compression is stated in the table as per the LTO consortium, real-world applications might achieve greater compression - ratios of 3:1 are not uncommon.

Table 13-3 Ultrium road map

Ultrium	Generation 1	Generation 2	Generation 3	Generation 4
Capacity (2:1 compression)	200 GB	400 GB	800 GB	1.6 TB
Transfer rate (2:1 compression)	20-40 MBps	40-80 MBps	80-160 MBps	160-320 MBps
Recording method	RLL 1, 7 Run Length Limited	PRML Partial Response Maximum Likelihood	PRML	PRML
Media	Metal Particle	Metal Particle	Metal Particle	Thin film

Reliability

In addition to considering today's speed and capacity requirements, the definition of the LTO standard takes into account, that the drives and the cartridges are implemented in automated libraries. From the cartridge design to the stable metal leader pin the LTO media is designed for automated processing. These design features make LTO media very reliable.

Because of the open format of LTO, clients have multiple sources of product and media, as well as enabling compatibility, between different vendor offerings. The IBM LTO implementation is the best-of-breed of LTO tape storage products available today.

For details of the IBM LTO Ultrium tape drive, refer to the IBM Redbook, *IBM System Storage Tape Library Guide for Open Systems*, SG24-5946.

IBM System Storage Enterprise Tape

IBM offers a range of tape drives that use the half-inch linear format. There have been three generations of 3590 tape drives. The most recent 3592 tape drive became available in 2003. The 3590 and 3592 tape cartridges use the same form factor of the 3480 and 3490 tape cartridges, making them compatible with existing automation and most cartridges storage.

The TS1120 (3592-E05 - second generation 3592) has dual-ported 4Gbps native switched FC interfaces for direct attachment to open system servers with Fibre Channel. Attachment to ESCON or FICON servers is provided via TS1120 Tape Controller Model C06. The TS1120 is supported in a wide range of mainframe, open, Linux, and Windows environments. For more information for TS1120 supported servers and interfaces, see:

http://www-03.ibm.com/servers/storage/tape/compatibility/pdf/ts1120_interop.pdf

The 3590 tape drives also support attachment to various environments. For information regarding 3590 supported servers and interfaces, refer to:

<http://www-1.ibm.com/servers/storage/tape/compatibility/pdf/3590opn.pdf>

3592 cartridges are offered in two capacities:

- ▶ Enterprise Tape Cartridge 3592 Model JA (3592JA)
- ▶ Enterprise Tape Cartridge 3592 Model JJ (3592JJ)

3590 cartridges are offered in two capacities:

- ▶ 3590 High Performance Cartridge Tape (3590J)
- ▶ 3590 Extended High Performance Cartridge Tape (3590K)

Table 13-4 provides native performance and capacity for the 3590 and 3592 drives and cartridges.

Table 13-4 3590 and 3592 native performance and capacity data

	3590-B1A	3590-E1A	3590-H1A	3592-J1A	3592-E05 (TS1120)
Data rate	9 MB / second	14 MB / second	14 MB / second	40 MB / second	100 MB / second
3590J	10 GB	20 GB	30 GB	NA	NA
3590K	20 GB	40 GB	60 GB	NA	NA
3592JA	NA	NA	NA	300 GB	500 GB/700 GB
3592JJ (TS1120)	NA	NA	NA	60 GB	100 GB

Reliability

A unique feature of the tape drive is the way data is striped horizontally and vertically. If a media error occurs, even one that covers several tracks, the error correction code can reconstruct the data for the application. The 3590 has proven to pass the hole in the tape test.

Figure 13-2 shows a picture of a test made on the Magstar 3590 E tape in 1999. Today the IBM System Storage TS1120 Tape Drive is an enterprise class tape drive that is enriched with a number of built-in capabilities together with the proven high reliability technology of its predecessor.

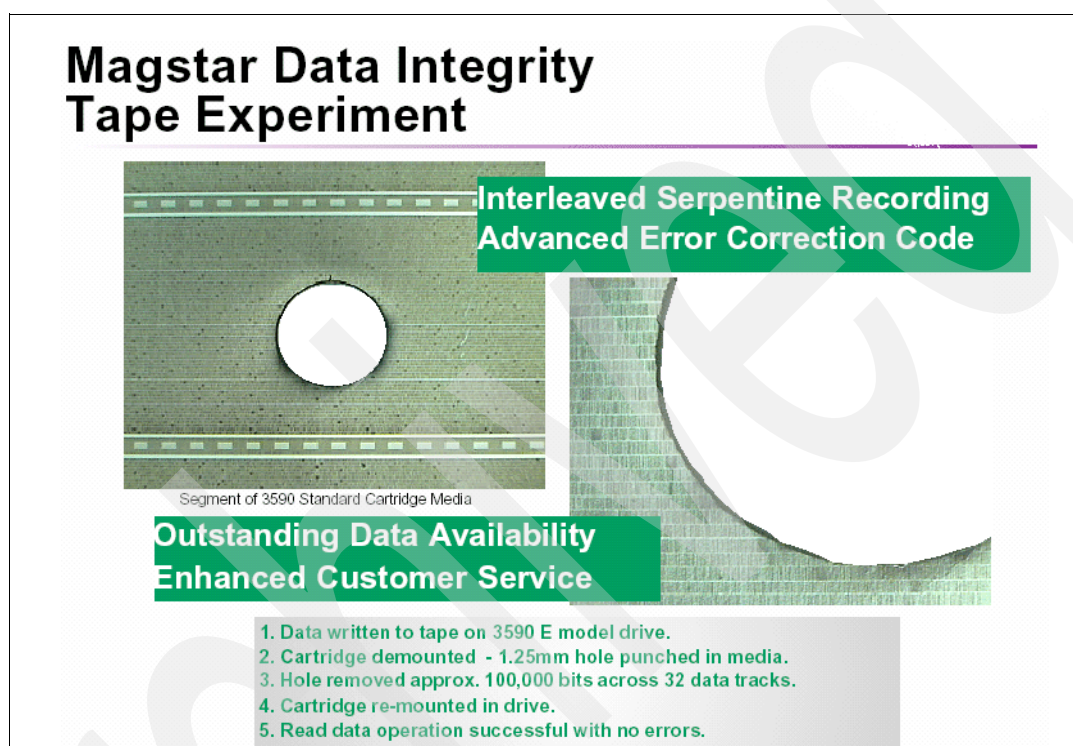


Figure 13-2 3590 reliability test

13.6 Recent tape technology advancements

In this section we cover the highlights of recent tape technology advancements. Some detailed discussion is covered in later sections, especially on the Virtual Tape Libraries (TS7510 and TS7740). With more focus on data security, we also have an overview section on the encryption tape feature on TS1120.

13.6.1 Tape drive enhancements

The following tape drive enhancements are provided:

- ▶ WORM tape, for regulatory requirements for data retention, is available on:
 - LTO3
 - TS1120
 - DR550 (see 4.4, "IBM Data Retention 550" on page 206)
- ▶ Tape encryption for improved data security is available on:
 - TS1120 (see 13.9, "IBM System Storage TS1120 tape drive and tape encryption" on page 479)

13.6.2 Tape automation enhancement

Virtual Tape Library enhancements for both open systems and mainframe facilitate the robust and efficient use of recent tape technology:

- ▶ TS7510 - Open Systems Virtual Tape (See 13.8, “IBM Virtualization Engine TS7510 overview” on page 473)
- ▶ TS7740 Grid - Mainframe Virtual Tape (See 13.7, “TS7740 Virtualization Engine overview” on page 462)

13.7 TS7740 Virtualization Engine overview

The IBM System Storage Virtualization Engine TS7700 represents the fourth generation of IBM Tape Virtualization for mainframe systems and replaces the highly successful IBM TotalStorage Virtual Tape Server (VTS).

The TS7700 Virtualization Engine is designed to provide improved performance and capacity to help lower the total cost of ownership for tape processing. It introduces a modular, scalable, high-performing architecture for mainframe tape virtualization. It integrates the advanced performance, capacity, and data integrity design of the IBM TS1120 Tape Drives, with high-performance disk and an inbuilt IBM System p server to form a storage hierarchy managed by robust storage management firmware with extensive self management capabilities.

The TS7700 Virtualization Engine utilizes outboard policy management to manage physical volume pools, cache management, to control selective dual copy, dual copy across a grid network, and copy mode control.

The TS7700 offers a new standards-based management interface and enhanced statistical reporting, compared to the VTS.

The TS7700 Virtualization Engine integrates the following components into the virtual tape solution:

- ▶ One *IBM Virtualization Engine TS7740 Server Model V06* (3957 Model V06)
- ▶ One *IBM Virtualization Engine TS7740 Cache Controller Model CC6* (3956 Model CC6)
- ▶ Three *IBM Virtualization Engine TS7740 Cache Drawers Model CX6* (3956 Model CX6)

Here are some important characteristics of these components:

- ▶ The TS7740 Server provides host connection of up to four FICON channels, and connections to the tape library and tape drives for back-end tape processing.
- ▶ A TS7700 with Grid Communication features can be interconnected with another TS7700 to provide peer-to-peer copy capability between Virtualization Engines for tape using IP network connections.
- ▶ The TS7740 Cache, comprised of the TS7740 Model CC6 and the TS7740 Model CX6 provides over 6 TB of tape volume cache capacity before compression.
- ▶ Each TS7700 supports up to a maximum of 128 3490E virtual tape drives and up to 500,000 logical volumes, each with a maximum capacity of 1.2 GB (assuming 3:1 compression) to 12 GB (assuming 3:1 compression and using the 400 to 4000 MB volume sizes).

Figure 13-3 shows the main components of the TS7740 Virtualization Engine.

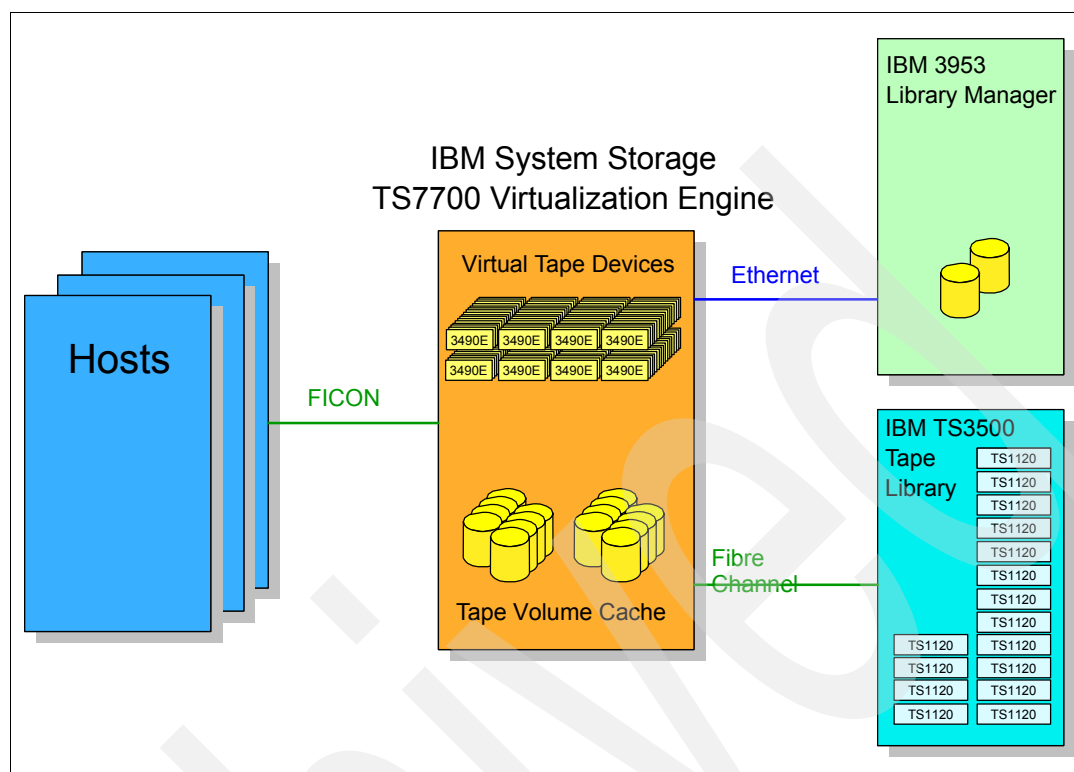


Figure 13-3 Main components of the TS7700 Virtualization Engine

Note: For details on IBM TS7740 Virtualization Engine, see the IBM Redbook, *IBM System Storage Virtualization Engine TS7700: Tape Virtualization for System z Servers*, SG24-7312.

13.7.1 TS7740 Virtualization Engine and business continuity

The TS7740 Virtualization Engine can take part in all three Business Continuity Segments.

Backup / restore

The TS7740 Virtualization Engine (Figure 13-4) provides virtualized tapes and tape drives for tape applications. This ensures full utilization of the high performance and high capacity of the TS1120 tape drives.

Multiple backup jobs can be streaming to the TS7740 tape cache then automatically destaged to the backend tape media utilizing the 100MB throughput of TS1120 drives.

In the restore process, multiple tape restores can be performed either from tape cache or via the seamless tape media staging to tape cache. The staging utilizes the superior tape performance of TS1120 including fast throughput and patented fast data access technology.

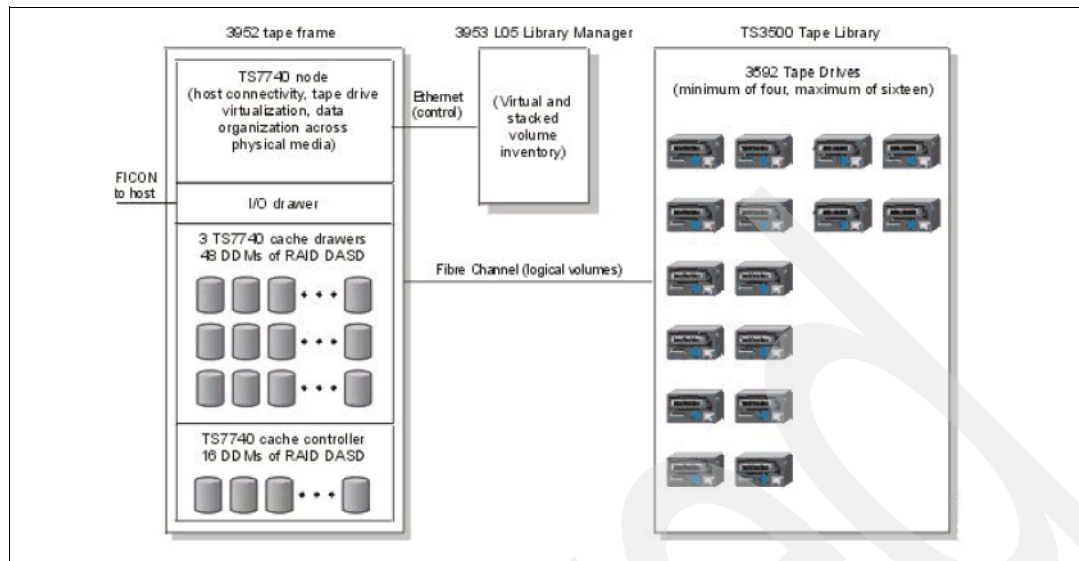


Figure 13-4 TS7740 cluster

Rapid Data Recovery

The TS7740 Virtualization Engine can be configured to have specific categories of application to be preferenced for retention in the cache pool for fast Disk to Disk restore. This complements the FlashCopy backup capability to improve the RTO in case of disaster.

The TS7740 Grid configuration (Figure 13-5) is a high availability tape automation configuration which further enables tape mirroring between two TS7740 Virtualization Engines. Two TS7700 Virtualization Engines can be interconnected via 1 Gb Ethernet to form a *Dual Cluster Grid*. Logical volume attributes and data are replicated across the Clusters in a Grid. Any data replicated between the Clusters is accessible through any other Cluster in a Grid configuration. By setting up policies on the TS7700s you define where and when you want to have a secondary copy of your data. You can also specify for certain kinds of data, for example test data, that no secondary copy is required.

A Dual Cluster Grid presents itself to the attached hosts as one large library with 256 virtual tape devices. The copying of the volumes in a Grid configuration is handled by the Clusters and is completely transparent to the host. Each TS7700 Virtualization Engine in a Grid manages its own set of physical volumes and maintains the relationship between logical volumes and the physical volumes on which they reside.

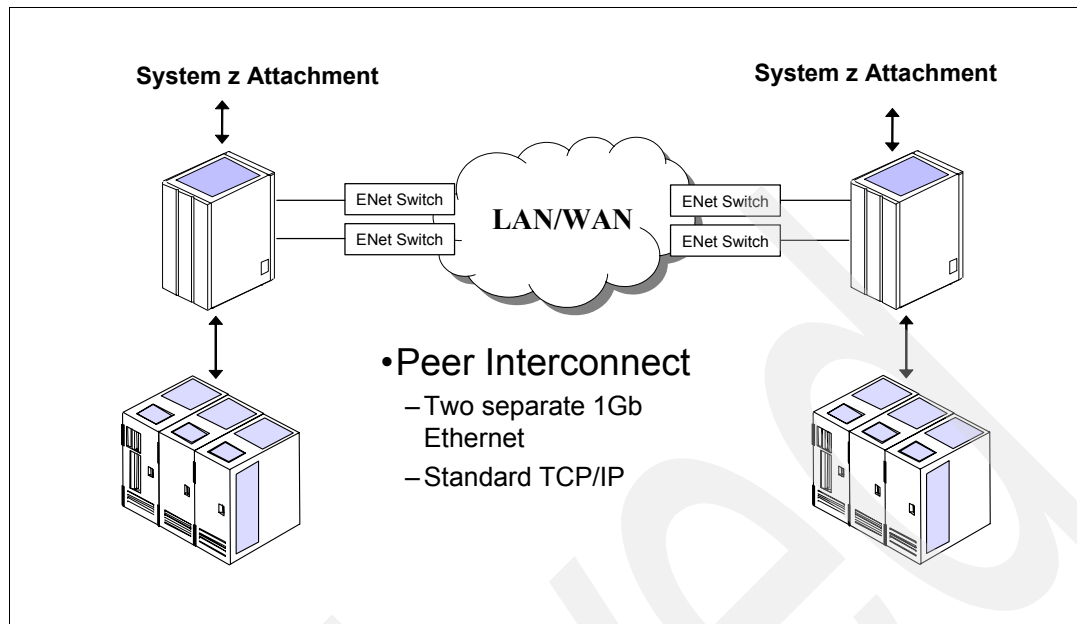


Figure 13-5 TS7700 Dual Cluster Grid Configuration

Continuous Availability

A Dual Cluster Grid can also take part in GDPS site level recovery together with GDPS/PPRC. This enables seamless automated cross site business recovery while retaining tape operation with the adequate tape data integrity base on the GDPS implementation design. Figure 13-9 on page 470 shows TS7700 Grid -Tier 7 solution with GDPS/PPRC in a System z environment.

Note: For details on IBM TS7700 Grid with a Geographical Dispersed Parallel Sysplex (GDPS), see chapter 9.4 - Geographically Dispersed Parallel Sysplex (GDPS) in the IBM Redbook, *IBM System Storage Virtualization Engine TS7700: Tape Virtualization for System z Servers*, SG24-7312.

Benefits of tape virtualization

Here are some of the main benefits you can expect from tape virtualization:

- ▶ High Availability and Disaster Recovery configurations
- ▶ Fast access to data through caching on disk
- ▶ Utilization of current tape drive, tape media, and tape automation technology
- ▶ Capability of filling high capacity media 100%
- ▶ Large number of tape drives available for concurrent use
- ▶ No additional software required
- ▶ Reduced Total Cost of Ownership (TCO)

,Figure 13-6 shows the TS7700 Virtualization Engine.



Figure 13-6 TS7740 Virtualization Engine Configuration

13.7.2 TS7740 operation overview

A virtual tape server presents emulated tape drives to the host and stores tape data on emulated tape volumes in a disk based cache rather than on physical tape media. The TS7740 emulates the function and operation of IBM 3490 Enhanced Capacity (3490E) tape drives and uses RAID5 disk to store volumes written by the host, see Figure 13-7. The disk space provided is called *Tape Volume Cache (TVC)*.

Emulated tape drives are also called *virtual drives*. To the host, virtual 3490E tape drives look exactly the same as physical 3490E tape drives - the emulation is transparent to the host and to the applications. The host always writes to and reads from virtual tape drives, it never accesses the physical tape drives in the back-end. In fact, it doesn't have to know that these tape drives exist.

As a consequence of this, even an application that only supports 3490E tape technology, can use the TS7700 Virtualization Engine without any changes to the application and still benefit from high capacity/high performance tape drives in the back-end. Tape virtualization provides a large number of virtual devices, 128 per TS7700. When jobs are contending for native attached physical drives, tape virtualization can solve this issue by providing a large number of virtual devices for concurrent use.

As the host exclusively accesses the virtual tape drives, all data must be written to or read from emulated volumes in the disk based Tape Volume Cache (TVC). We call tape volumes residing in the TVC, *virtual volumes*.

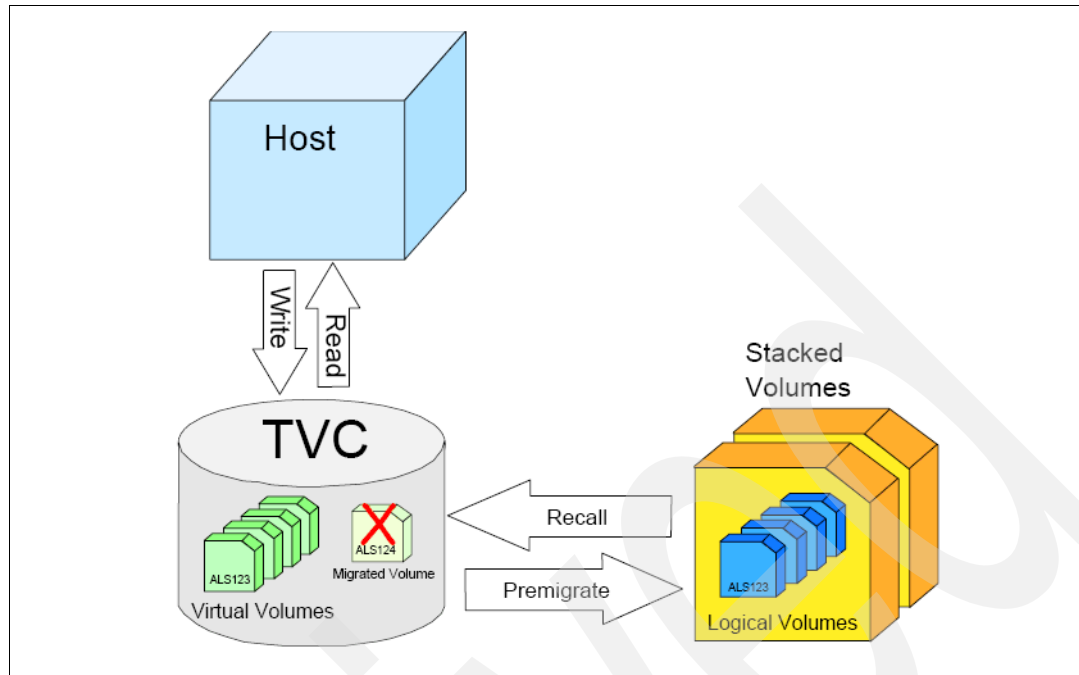


Figure 13-7 TS7740 Tape Virtualization - Tape Volume Cache Processing

When the host requests a volume that is still in cache, the volume is virtually mounted - no physical mount is required. Once the virtual mount is complete the host can access the data at disk speed. Mounting of scratch tapes also is virtual and does not require a physical mount.

Although you define maximum sizes for your volumes, a virtual volume takes up just the space in cache that the data on the volume actually requires. Later, when a virtual volume is copied from disk to tape, it also requires only the amount of tape capacity occupied by the data. Tape virtualization makes very efficient use of disk and tape capacity.

The TS7740 Node manages the physical tape drives, or physical volumes, in the tape library and controls the movement of data between physical and virtual volumes.

Data that is written from the host into the tape volume cache, is scheduled to be copied to tape at a later point in time. The process of copying data still existing in cache to tape is called *premigration*. When a volume has been copied from cache to tape, the volume on the tape is called a *logical volume*. A physical volume can contain a large number of logical volumes. The process of putting several logical volumes on one physical tape is called *stacking*, and a physical tape containing logical volumes is therefore referred to as a *stacked volume*.

As many applications cannot fill the high capacity media of today's tape technology, you can end up with a large number of under-utilized cartridges, wasting a lot of physical media space and requiring an excessive number of cartridge slots in the tape library. Tape virtualization reduces the space required by volumes and fully utilizes the capacity of current tape technology.

When space is required in the TVC for new data, volumes that already have been copied to tape are removed from the cache. The user has the choice, using a DFSMS Management Class, of specifying whether the logical volume should be removed from cache based on LRU or should be preferenced for quick removal based on size. The process of copying volumes from cache to tape and afterwards deleting them is called *migration*. Accordingly, volumes that have been deleted in the cache, and now only exist on tape are called *migrated volumes*.

When the host has to access a volume that has previously been migrated, the volume has to be copied back from tape into the TVC, as the host has no direct access to the physical tapes. When the complete volume has been copied back into the cache, the host is able to access the data on this volume. The process of copying data back from tape to the Tape Volume Cache is called *recall*.

13.7.3 TS7740 Grid overview

Two TS7700 Clusters can be interconnected to provide a disaster recovery/high availability solution. The Grid enablement feature must be installed on both TS7700 Virtualization Engines and an Ethernet connection is required between the engines.

Logical volume attributes and data are replicated across the Clusters in a Grid to ensure the continuation of production work, should a single Cluster become unavailable. Any data replicated between the Clusters is accessible through any of the other Clusters in a Grid configuration. A Grid configuration looks like a single storage subsystem to the hosts attached to the Clusters, and is referred to as a composite library with underlying distributed libraries similar to the prior generation's Peer-to-Peer Virtual Tape Server. Multiple TS7700 Virtualization Engine Grids can be attached to host systems and operate independent of one another.

Currently, a *Multi Cluster Grid* comprises two Clusters, and is therefore also called a *Dual Cluster Grid*. The architecture is designed to allow more than two Clusters in a Multi Cluster Grid configuration in the future.

Multi Cluster Grid related terms

In this section we explain terms that are specific to a Multi Cluster Grid configuration.

Composite library

The composite library is the logical image of the Grid, which is presented to the host. As opposed to the IBM Virtual Tape Server, a *Single Cluster* TS7700 Virtualization Engine also has a Composite LIBRARY-ID defined. From an architectural perspective, a stand-alone TS7700 is considered as a Grid consisting of just one cluster. We refer to such a Cluster as a Single Cluster Grid.

Both with a TS7700 Single Cluster Grid configuration and a TS7700 Dual Cluster Grid configuration a composite library is presented to the host. In the case of a stand-alone TS7700, the host sees a logical tape library with eight 3490E tape control units, each of them with sixteen IBM 3490E tape drives, attached through two or four FICON channel attachments. In the case of a Dual Cluster Grid the host sees a logical tape library with sixteen 3490E tape control units, each of them with sixteen IBM 3490E tape drives, attached through four or eight FICON channel attachments.

Distributed library

A distributed library is a Library Manager partition in a TS3500/3953 library associated with a TS7700 Virtualization Engine. In a Grid configuration, which includes two TS7700 Virtualization Engines each of which has the Grid Enablement Feature installed, two distributed libraries are required. The host has sufficient knowledge about the distributed libraries to allow appropriate console message handling of messages from the Library Manager of a distributed library. On the host, the distributed library is only defined to SMS. It is defined using the existing ISMF panels and has no tape devices defined. The tape devices are defined for the composite library only.

Figure 13-8 shows an example of the Composite Library and Distributed Library relationship in a TS7740 Dual Cluster Grid Configuration.

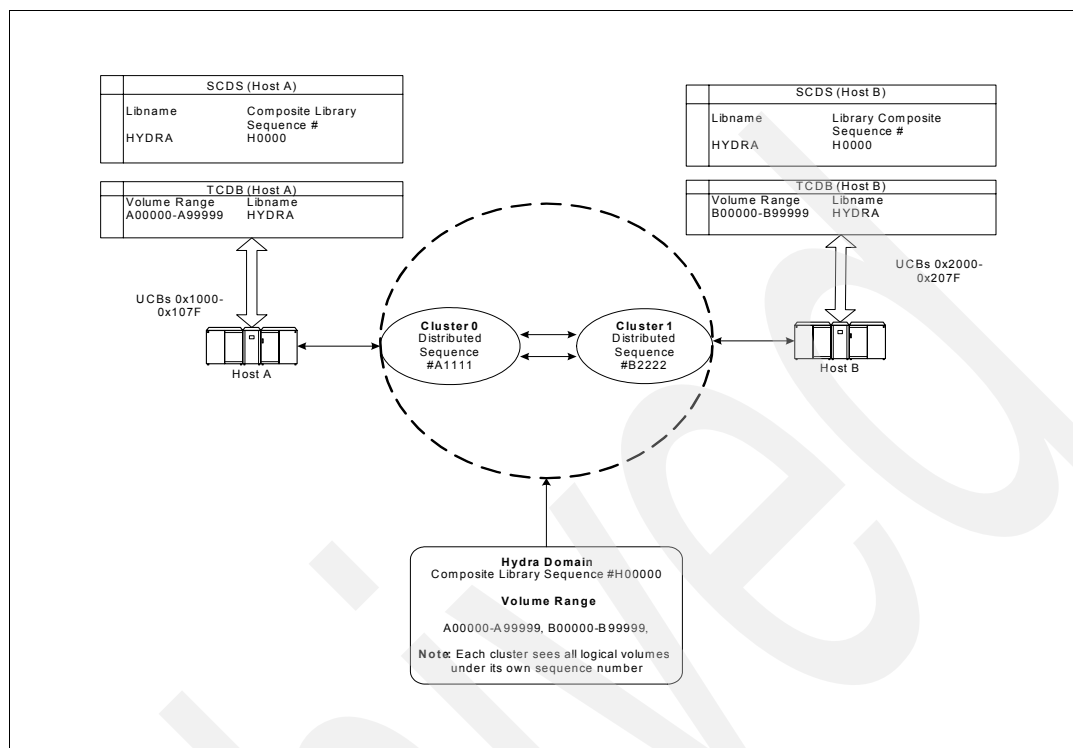


Figure 13-8 Example of TS7740 Dual Cluster Grid Configuration

13.7.4 TS7740 and GDPS (Tier 7) Implementation

The IBM System z multi-site application availability solution, Geographically Dispersed Parallel Sysplex (GDPS), integrates Parallel Sysplex technology and remote copy technology to enhance application availability and improve disaster recovery.

The GDPS topology is a Parallel Sysplex cluster spread across two sites, with all critical data mirrored between the sites. GDPS provides the capability to manage the remote copy configuration and storage system(s), automates Parallel Sysplex operational tasks, and automates failure recovery from a single point of control, thereby improving application availability.

GDPS is a multi-site management facility incorporating a combination of system code and automation that utilizes the capabilities of Parallel Sysplex technology, storage subsystem mirroring, and databases to manage processors, storage, and network resources. It is designed to minimize and potentially eliminate the impact of a disaster or planned site outage. The GDPS provides the ability to perform a controlled site switch for both planned and unplanned site outages, with no data loss, maintaining full data integrity across multiple volumes and storage subsystems, and the ability to perform a normal DBMS restart (not DBMS recovery) at the opposite site.

Figure 13-9 shows a TS7700 Grid in a Tier 7 solution with GDPS/PPRC in a System z environment.

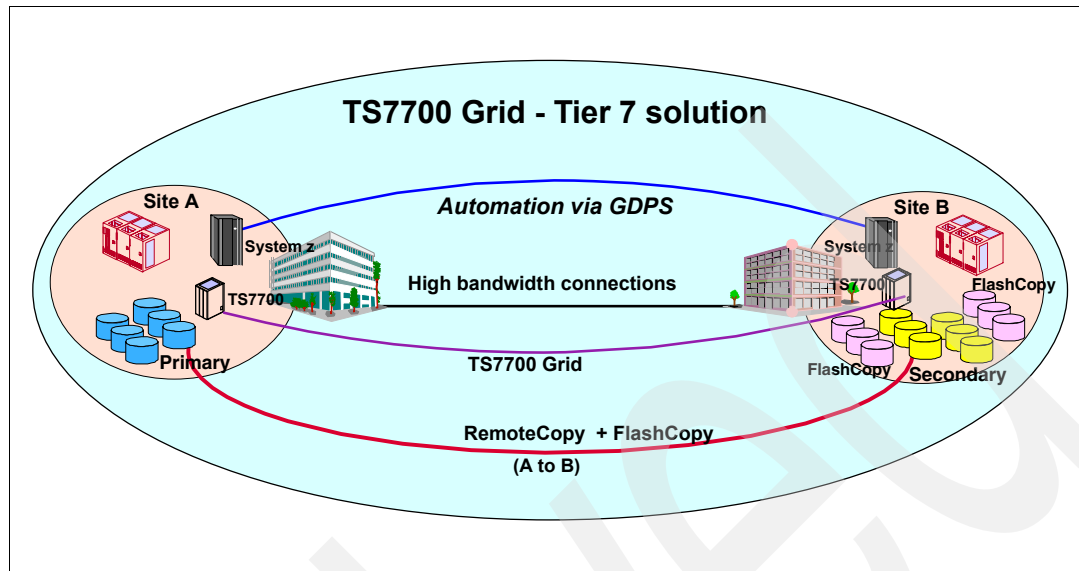


Figure 13-9 A TS7700 Grid -Tier 7 solution with GDPS/PPRC in a System z environment

GDPS functions

GDPS provides the following functions:

- ▶ Remote Copy Management Facility (RCMF), which automates management of the remote copy infrastructure
- ▶ Planned reconfiguration support, which automates operational tasks from one single point of control.
- ▶ Unplanned reconfiguration support, which recovers from a z/OS, processor, storage subsystem, or site failure

Remote copy management facility

RCMF was designed to simplify the storage administrator's remote copy management functions by managing the remote copy configuration rather than individual remote copy pairs. This includes the initialization and monitoring of the copy volume pairs based upon policy and performing routine operations on installed storage systems.

Planned reconfigurations

GDPS planned reconfiguration support automates procedures performed by an operations center. These include standard actions to:

- ▶ Quiesce a system's workload and remove the system from the Parallel Sysplex cluster.
- ▶ IPL a system
- ▶ Quiesce a system's workload, remove the system from the Parallel Sysplex cluster, and re-IPL the system. The standard actions can be initiated against a single system or group of systems. Additionally, user-defined actions are supported (for example, a planned site switch in which the workload is switched from processors in site A to processors in site B).

Unplanned reconfigurations

GDPS was originally designed to minimize and potentially eliminate the amount of data loss and the duration of the recovery window in the event of a site failure; however, it also minimizes the impact and potentially mask an z/OS system or processor failure based upon GDPS policy. GDPS uses PPRC or XRC to help minimize or eliminate data loss. Parallel Sysplex cluster functions along with automation are used to detect z/OS system, processor, or site failures and to initiate recovery processing to help minimize the duration of the recovery window.

If a z/OS system fails, the failed system is automatically removed from the Parallel Sysplex cluster, re-IPLed in the same location, and the workload restarted. If a processor fails, the failed system(s) are removed from the Parallel Sysplex cluster, re-IPLed on another processor, and the workload restarted.

With PPRC, there is limited or no data loss, based upon policy, since all critical data is being synchronously mirrored from site A to site B in the event of a site failure. There is limited data loss if the production systems continue to make updates to the primary copy of data after remote copy processing is suspended (any updates after a freeze are not reflected in the secondary copy of data) and there is a subsequent disaster that destroys some or all of the primary copy of data. There is no data loss if the production systems do not make any updates to the primary PPRC volumes after PPRC processing is suspended.

Depending on the type of application and recovery options selected by the enterprise, multiple freeze options are supported by GDPS (the freeze is always performed to allow the restart of the software subsystems).

GDPS considerations on a TS7700 Grid configuration

The default behavior of the TS7700 Virtualization Engine is to follow the Management Class definitions as well as considerations to provide the best overall job performance. In a Geographically Dispersed Parallel Sysplex (GDPS), all I/O must be local (primary) to the mount vNode. Host channels from the GDPS hosts to each remote TS7700 Cluster typically are also installed. During normal operation the remote virtual devices are set offline in each GDPS host. See Figure 13-10.

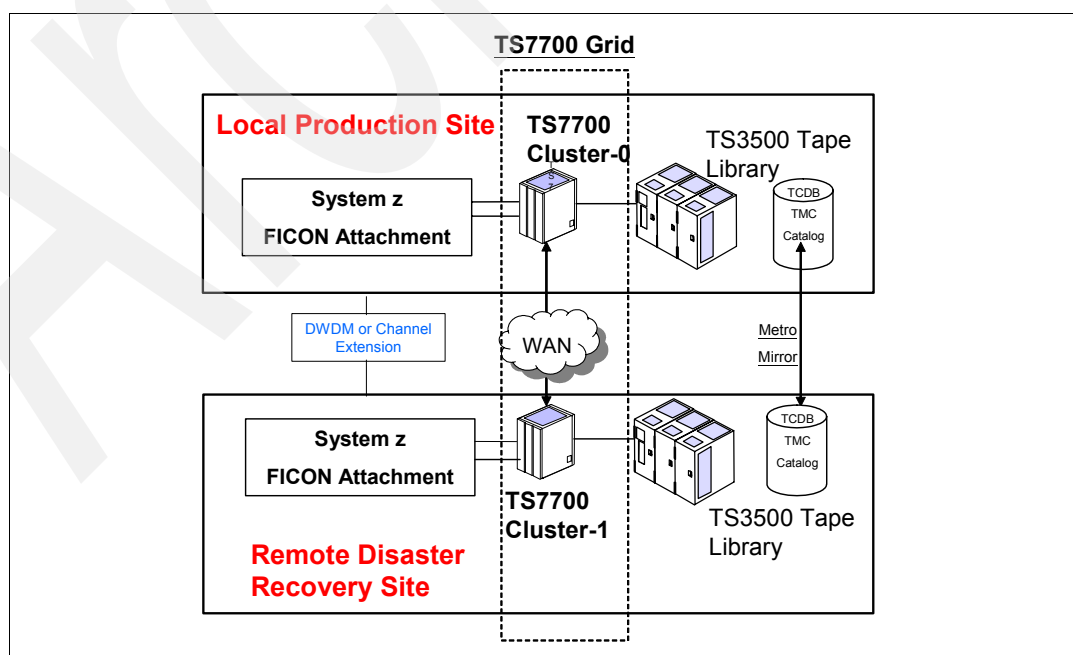


Figure 13-10 TS7740 Grid with GDPS Configuration

A dual Cluster Grid TS7700 under GDPS control Figure 13-10 is required to set Consistency Point to **RUN/RUN** (both Clusters) and Copy Policy Override (both Clusters as well) to:

► **Prefer local cache for fast ready mount requests:**

This override selects the TVC local to the mount vNode Cluster as the I/O TVC as long as it is available and a copy consistency point other than No Copy is specified for that Cluster in the Management Class specified with the mount. The Cluster does not have to have a valid copy of the data for it to be selected for the I/O TVC.

► **Prefer local cache for non-fast ready mount requests:**

This override selects the TVC local to the mount vNode Cluster as the I/O TVC as long as it is available and the Cluster has a valid copy of the data, even if the data is only resident on a physical tape. Having an available, valid copy of the data overrides all other selection criteria. If the local Cluster does not have a valid copy of the data, then the default selection criteria applies.

► **Force volumes mounted on this cluster to be copied to the local cache:**

This override has two effects, depending on the type of mount requested. For a non-fast ready mount, a copy is performed to the local TVC as part of the mount processing. For a fast ready mount, it has the effect of 'OR-ring' the specified Management Class with a copy consistency point of Rewind/Unload for the Cluster. The override does not change the definition of the Management Class, it serves only to influence the selection of the I/O TVC or force a local copy.

In case of failure in a TS7700 Grid environment together with GDPS, let us consider the following three scenarios:

1. GDPS switches the primary host to the Remote Location; the TS7700 Grid is still fully functional:
 - No manual intervention required
 - Logical volume ownership transfer is done automatically during each mount via Grid
2. A disaster happens at the primary Site; GDPS host and TS7700 Cluster are down or inactive:
 - Automatic ownership takeover of volumes, which are then accessed from the remote host, is not possible.
 - Manual intervention is required to invoke a manual ownership takeover.
3. Only TS7700 Cluster in GDPS primary site is down. In this case, two manual interventions are required:
 - Vary online remote TS7700 Cluster devices from primary GDPS host
 - Since the downed Cluster cannot take automatically ownership of volumes, which are then accessed from the remote host, a manual intervention is required to do this.

Note: For details on the IBM TS7700 Grid with a Geographical Dispersed Parallel Sysplex (GDPS), see section 9.4, “Geographically Dispersed Parallel Sysplex (GDPS)” in the IBM Redbook, *IBM System Storage Virtualization Engine TS7700: Tape Virtualization for System z Servers*, SG24-7312.

13.8 IBM Virtualization Engine TS7510 overview

The storage industry refers to the IBM Virtualization Engine TS7510 as a *virtual tape library*.

A virtual tape library provides high performance backup and restore by using disk arrays and virtualization software. The TS7510 includes an IBM System x 346 server running Linux attached to an IBM System Storage DS4000 disk system, running tape virtualization software, as shown in Figure 13-11.

13.8.1 TS7510 and Business Continuity

The TS7510 Virtualization Engine can take part in the following two Business Continuity Segments.

Backup/restore

The TS7510 Virtualization Engine provides virtualized tapes and tape drives for tape applications. The above ensure full utilization of the high performance and high capacity tape technology like TS1120 or LTO3.

Multiple backup jobs can be streamed to the TS7510 tape cache, which is a seamless Disk-to-Disk (D2D) copy. The Tape data on TS7510 can be retained on Tape Cache for fast restore.

The Virtual Tape can also be configured to destage to back-end tape of TS1120 or LTO3. This destaging enables tape cache to sustain the daily tape workload of D2D operation.

Rapid data recovery

The TS7510 Virtualization Engine can be configured to have specific categories of application to be retained in the cache pool for fast Disk to Disk restore. This complements the FlashCopy backup capability to improve the RTO in case of disaster.

13.8.2 TS7510 configuration overviews

Figure 13-11 shows the configuration of the TS7510.

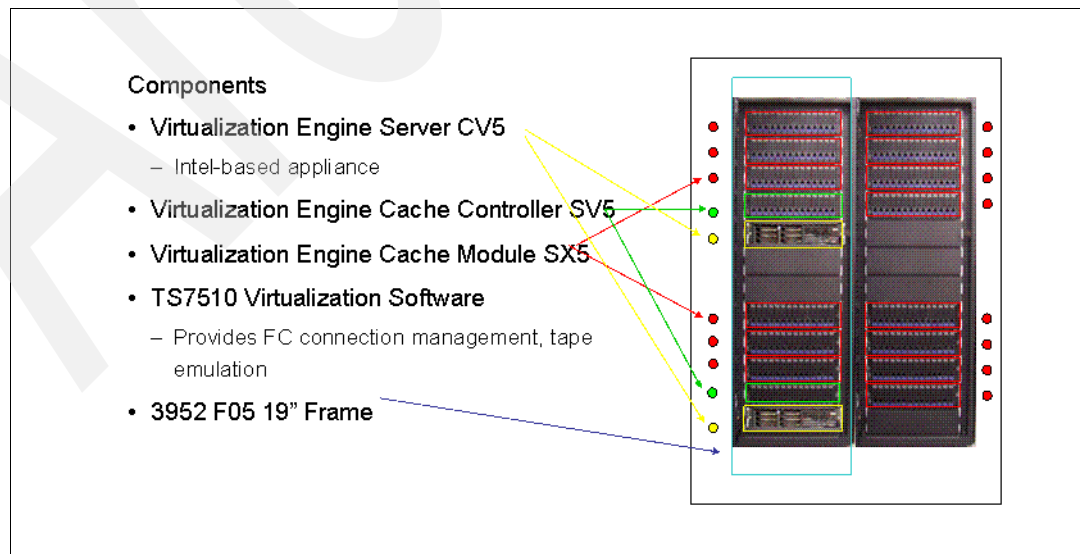


Figure 13-11 TS7510 Virtual Tape Library solution components

A virtual tape library is a unique blend of several storage tiers. The life cycle of data from its creation at the server level migrates by backup software to a virtual tape library. The virtual tape library is a combination of high performance SAN-attached disk and high performance servers running Linux emulating a tape storage device, to optimize the data placement. For example, the data can remain on the virtual tape library indefinitely, as long as there is enough space, or it can be migrated to tape for off-site storage, archive, or both. A virtual tape library, the TS7510, is shown in Figure 13-12.

Who requires a virtual tape library?

The answer is: any business that has a requirement for high performance backup and restore. Instant access backups and restores are capabilities that only the disk storage tier can provide. Not all backup software has the ability to back up to disk directly, which is why the virtual tape library is necessary. All backup software products were designed to back up to tape drives and libraries, and as mentioned earlier, a virtual tape library emulates tape. Environments with many small or large LAN free clients can go directly to a virtual tape library to complete their backups quickly.

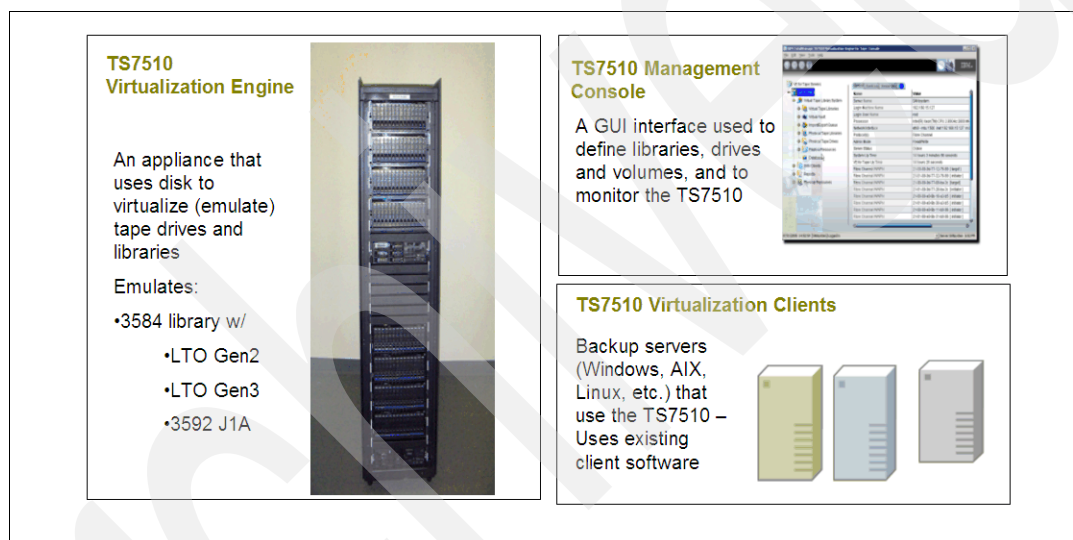


Figure 13-12 TS7510 Virtual Tape Library functional overview

Virtual tape libraries fill a void in the backup infrastructure for data that has to be restored at a moment's notice. Service-level agreements can now be designed to take advantage of data staged on disk. Backup software can be configured to back up data to a virtual tape library, and then create a virtual tape-to-tape copy for off-site deployment. No longer is it necessary to call the tapes back from off-site, unless data is required from years past. The IBM Virtualization Engine TS7510 has been designed specifically to provide this solution.

Here are the IBM Virtualization Engine TS7510 components and model numbers

- ▶ IBM Virtualization Engine TS7510 (3954 Model CV5)
- ▶ IBM Virtualization Engine TS7510 Cache Controller (3955 Model SV5)
- ▶ IBM Virtualization Engine TS7510 Cache Module (3955 Model SX5)
- ▶ IBM System Storage 3952 Tape Frame (3952 Model F05)
- ▶ IBM Virtualization Engine TS7510 Software Version 1 Release 1 (5639-CC7)

The TS7510 Virtual Tape Library features:

- ▶ Up to 600 MBps and 46 TB native capacity
- ▶ Up to 128 virtual tape libraries, 1024 virtual drives and 8192 virtual volumes supports
- ▶ Supports IBM 3494 and 3584 tape libraries
- ▶ Supports attachment to current models of IBM tape drives
- ▶ Optional features include network replication, compression and encryption
- ▶ Attaches to IBM System p, System z (Linux) and System x, selected HP and Sun Microsystems™ servers, and servers running supported Microsoft Windows and Linux operating systems

The TS7510 Virtualization Engine configuration table is shown in Figure 13-13.

	Single TS7510 Virtualization Engine							
Virtual Drives	512							
Min / Inc / Max Virtual Libraries	64							
Min / Inc / Max Virtual Volumes	4096							
Performance (~ max)	500							
Raw capacity	7.0	10.5	14.0	17.5	21.0	24.5	28.0	
Max Useable Capacity	5	8	11	14	17	20	23	
Number of Disk Drawers	2	3	4	5	6	7	8	

	Dual TS7510 Virtualization Engine							
Virtual Drives	1024							
Min / Inc / Max Virtual Libraries	128							
Min / Inc / Max Virtual Volumes	8192							
Performance (~ max)	600							
Raw capacity	31.5	35.0	38.5	42.0	45.5	49.0	52.5	56.0
Max Useable Capacity	25.5	28	31	34	37	40	43	46
Number of Disk Drawers	9	10	11	12	13	14	15	16

Figure 13-13 TS7510 Virtualization Engine Configuration Table

Implementing TS7510 Virtual Tape Library is not a complex task. Figure 13-14 provides an example of a high level installation and implementation of TS7510 which is ready to use with Tivoli Storage Manager, as though a native TS3500 (3584) Library and 3592 J1A Drives were installed.

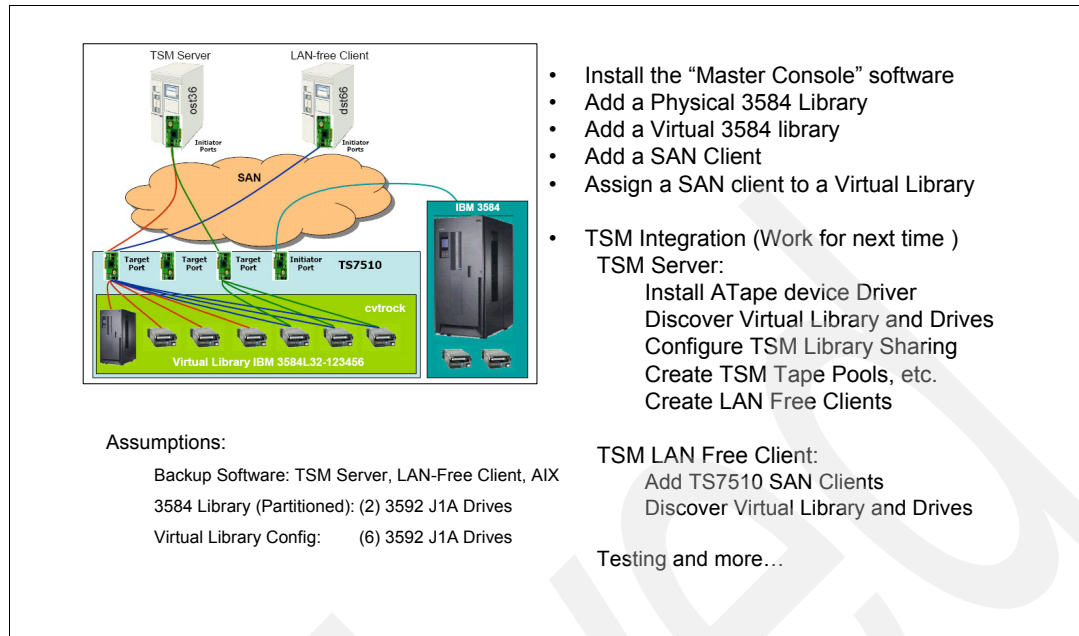


Figure 13-14 Installation and implementation solution example of TS7510 Virtual Tape Library

13.8.3 IBM Virtualization Engine TS7510 software

The IBM Virtualization Engine TS7510 application software provides tape library and tape drive emulation including virtual volumes. It includes the following features and functions, which we explain next:

- ▶ Backup and network compression
- ▶ Import and export including automatic archive
- ▶ Network replication including encryption

Network replication

Network replication provides a method to recover from complete data loss by sending copies of data off site. There are three methods of Network Replication: Remote Copy, Replication, and Auto Replication. To use the Network Replication function, you require two IBM Virtualization Engine TS7510s:

- ▶ The primary TS7510 that serves virtual tape drives to your backup servers
- ▶ A disaster recovery/remote Virtualization Engine TS7510

Remote Copy

Remote Copy is a manually triggered, one-time replication of a local virtual tape. Upon completion of the Remote Copy, the tape resides on the primary TS7510 and in the Replication Vault of the remote TS7510.

When using Remote Copy, the copied tape can reside either in one of the virtual tape libraries, in a virtual tape drive, or in the virtual vault. The Remote Copy option preserves the barcode from the Virtualization Engine which the remote copy initiated.

Figure 13-15 illustrates the Remote Copy movement. The primary Virtualization Engine is on the left, and the remote backup is on the right.

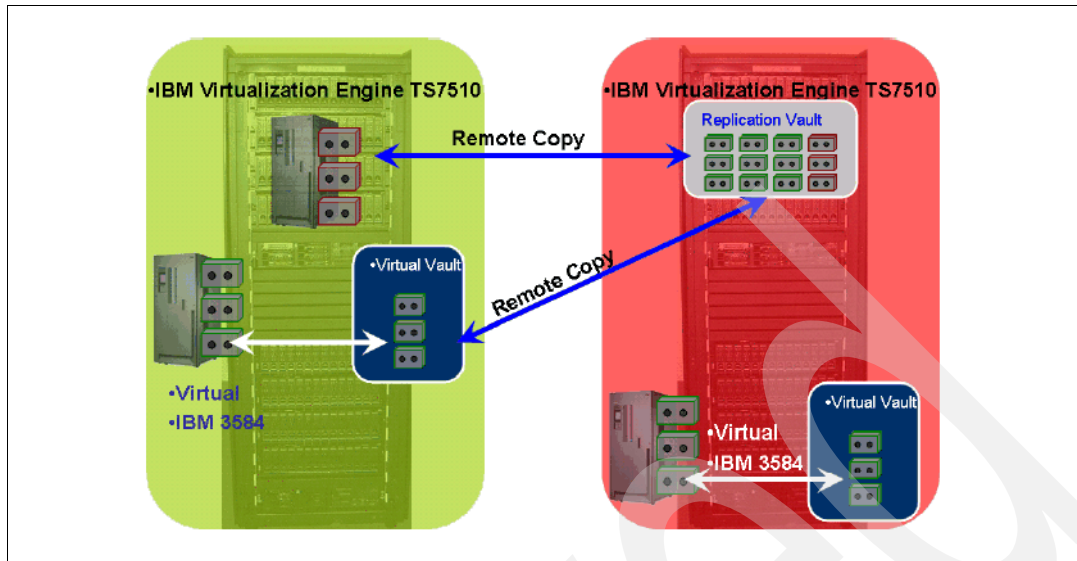


Figure 13-15 Remote Copy Data Movement

Replication

The Replication process is either triggered by a scheduled event or when the virtual volume reaches a certain predetermined size. When Replication is configured, a primary virtual volume is created and linked to the virtual replica on the remote Virtualization Engine. A replica tape is always linked to the original virtual tape. It cannot be used by any virtual library or for import/export by the remote Virtualization Engine until this linked relationship is broken. This condition is also known as *promoting a replica*. Its only purpose is to maintain an in-sync copy of the primary tape.

The replica tape simply gets incremental changes from the source tape, ensuring the two tapes are always in-sync at the end of a replication session. This is why it is a *dedicated relationship*. Since the incremental changes are trackable (because we know no one else is writing to or accessing the replica), there is never a requirement to replicate or remote copy the entire tape at each replication interval.

Data traveling across the replication path can be compressed, encrypted, or both. Additional license codes are required to activate these features, which we explain later. If the replica is promoted, it is placed in the virtual vault on the remote Virtualization Engine, with the same barcode label as the source virtual tape. It can then be used like any other virtual tape.

Figure 13-16 illustrates replication movement. The left TS7510 is the primary engine, and the right TS7510 is the backup. Data replicates from the primary to the backup utilizing the replication process. When the primary engine fails in order to use a replica that is on the backup engine, the virtual replica sitting in the replication vault is promoted to a virtual volume and moved to the virtual vault. It is either placed in a virtual library on the backup or copied back to the primary.

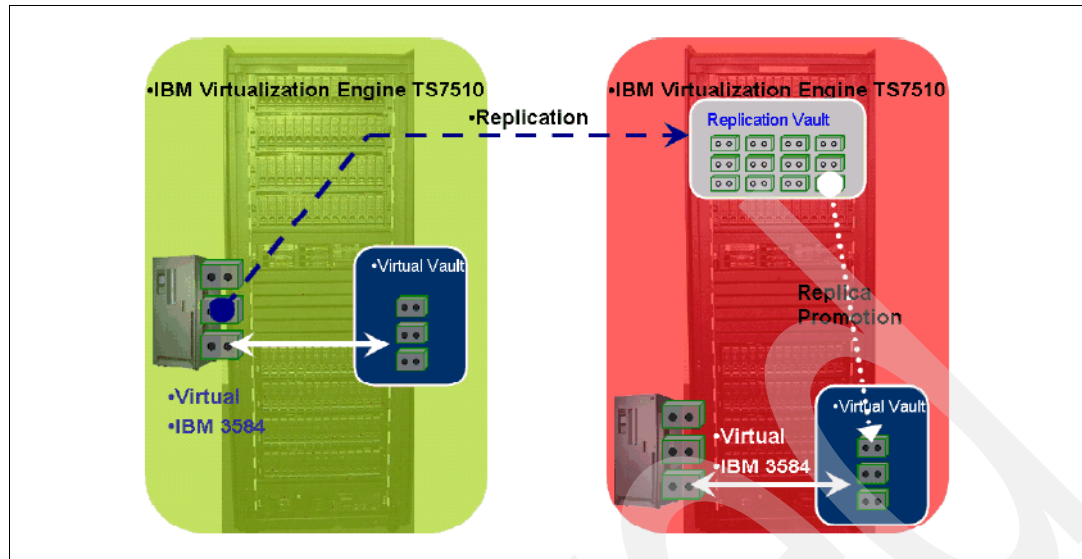


Figure 13-16 Replication data movement

Auto Replication

Auto Replication provides a one-time copy or move of a virtual tape to a remote Virtualization Engine as soon as the backup software has sent an **eject** command. Figure 13-17 illustrates the Auto Replication process. The left side shows the primary engine, and the right side shows the backup engine. The primary initiates the Auto Replication function. Also a one-time copy or move after the **eject** command is sent to the backup Virtualization Engine. The virtual volume is then placed in the replication vault.

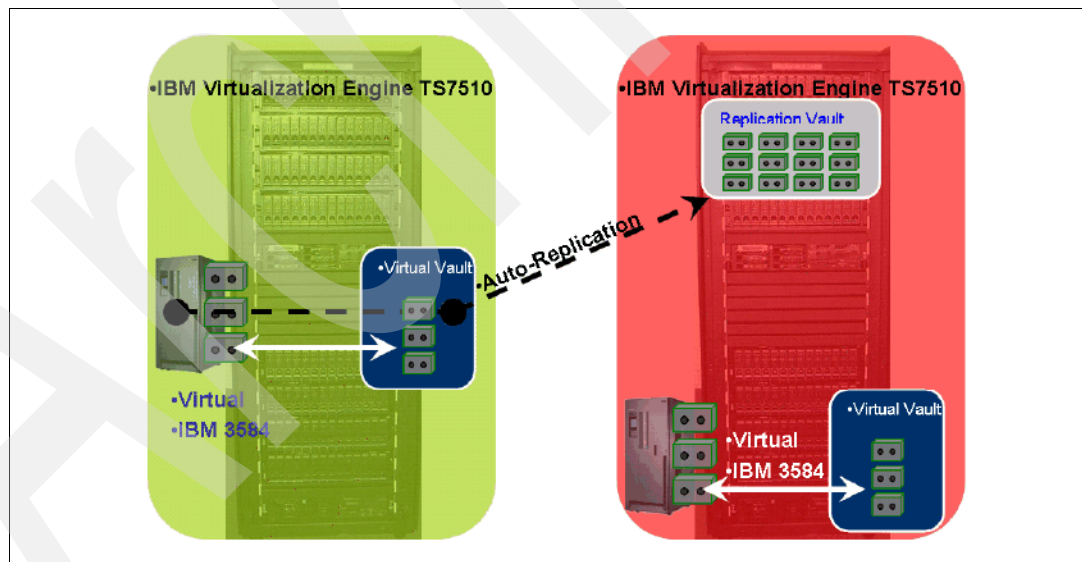


Figure 13-17 Auto Replication data movement

Network Encryption

With Network Encryption, feature code 7422, encrypts data while it is sent between two IBM Virtualization Engine TS7510s. The data is encrypted across the line on the fly and decrypted when it reaches the Remote Virtualization Engine.

Important: FC7421 (Network Replication) is a prerequisite for FC7422 (Network Encryption) and FC7423 (Network Compression). If two Model CV5 servers are configured in a high availability configuration, these software codes must exist on both models.

Network Compression

Network Compression, FC7423 compresses the data stream across the network to enhance the performance of all replication functions.

Note: For more details on IBM TS7510 “Configuration and Sizing”, see Chapter 3 of *IBM Virtualization Engine TS7510: Tape Virtualization for Open Systems Servers*, SG24-7189.

High Availability Failover

TS7510 also provide the High Availability Failover option when a primary TS7510 takes over the identity of the secondary (HA) TS7510, because of a hardware or software failure that affects the proper operation of the Virtualization Engine. This option requires you to purchase the high availability configuration of the IBM Virtualization Engine TS7510.

High Availability Failover provides high availability for the TS7510 storage network, and protection from potential problems such as, for example, storage device path failure, or Virtualization Engine TS7510 server failure.

Note: For details on IBM TS7510 “High Availability Failover”, see section 3.1 on IBM Redbook, *IBM Virtualization Engine TS7510: Tape Virtualization for Open Systems Servers*, SG24-7189.

13.9 IBM System Storage TS1120 tape drive and tape encryption

Recent legislation has established circumstances where security breaches involving consumer information might trigger a requirement to notify consumers. According to privacyrights.org, a non profit organization, 32 states in the US have enacted this legislation, and over 90 million American consumers have been notified that their information has been potentially compromised. Similar actions are occurring around the world. This has resulted in a great deal of interest in methods of protecting data, one of the methods is encryption.

The IBM System Storage TS1120 Tape Drive (3592- E05), Figure 13-18, is designed for high-performance tape applications, including:

- ▶ High-speed data-save operations where backup windows are critical, and large amounts of data are archived to tape.
- ▶ Protection of tape data by high speed encryption for security or regulatory requirements.
- ▶ Large-scale automated tape environments where performance and reliability are required.
- ▶ Large-scale mass data archive applications where massive amounts of data have to be quickly saved to tape for storage and later recall (examples include the seismic industry, data warehousing, and record management applications).



Figure 13-18 TS1120 Tape Drive with Tape Encryption Capability

The TS1120 Tape Drive is the second generation of the 3592 tape product family - the first was the 3592-J1A. The TS1120 Tape Drive features:

- ▶ Larger capacity compared to the 3592 Model J1A, to help reduce the number of cartridges required and floor space required
- ▶ Smaller form factor compared to the 3590 Tape Drive, which can help reduce the automation foot print
- ▶ Dual-ported native switched fabric interface
- ▶ High reliability, to help improve operations
- ▶ Additional performance and access improvements over the 3592 J1A
- ▶ **Hardware tape encryption** to protect tape data integration with the new Encryption Key Manager component

TS1120 Tape Drive specifications

These are the specifications for the TS1120:

- ▶ 104 MBps performance (up to 260 MBps at 3:1 compression)
- ▶ 100 / 500 / 700 GB native capacity (up to 300 GB / 1.5 TB / 2.1 TB at 3:1 compression)
- ▶ Re-Writable and Write Once Read Many (WORM) cartridges
- ▶ Supports data encryption and key management
- ▶ Attaches to:
 - All IBM servers (IBM System z via TS1120 Controller)
 - Selected HP and Sun Microsystems servers
 - Selected versions of Microsoft Windows
 - Selected Linux editions
- ▶ Supported in:
 - IBM 3494 and TS3500 tape libraries
 - IBM 3592 C20 silo compatible frame
 - IBM 7014 Rack

13.9.1 TS1120 tape encryption and business continuity

Tape encryption, which is included at no charge with the TS1120, plays an important role in Business Continuity scenarios where backup and archive data are mostly stored to tape.

Tape encryption:

- ▶ Protects tape data in transit from the primary data center to a secondary data center or Business Continuity site

- Protects tape data generated by mainframe as well as open systems, using the same management infrastructure
- Protects tape data in transit to a trusted partner, but allows access once the data has arrived

TS1120 tape encryption operation overview

The TS1120 Tape Drive supports encryption of data on a tape cartridge. New T1120 Tape Drives are encryption capable and a chargeable upgrade is available for previously installed drives. The encryption capability is implemented through tape drive hardware as well as microcode additions and changes. All 3592 media, including WORM cartridges, can be encrypted. In addition, a new Encryption Key Manager program supports encryption. See Figure 13-19.

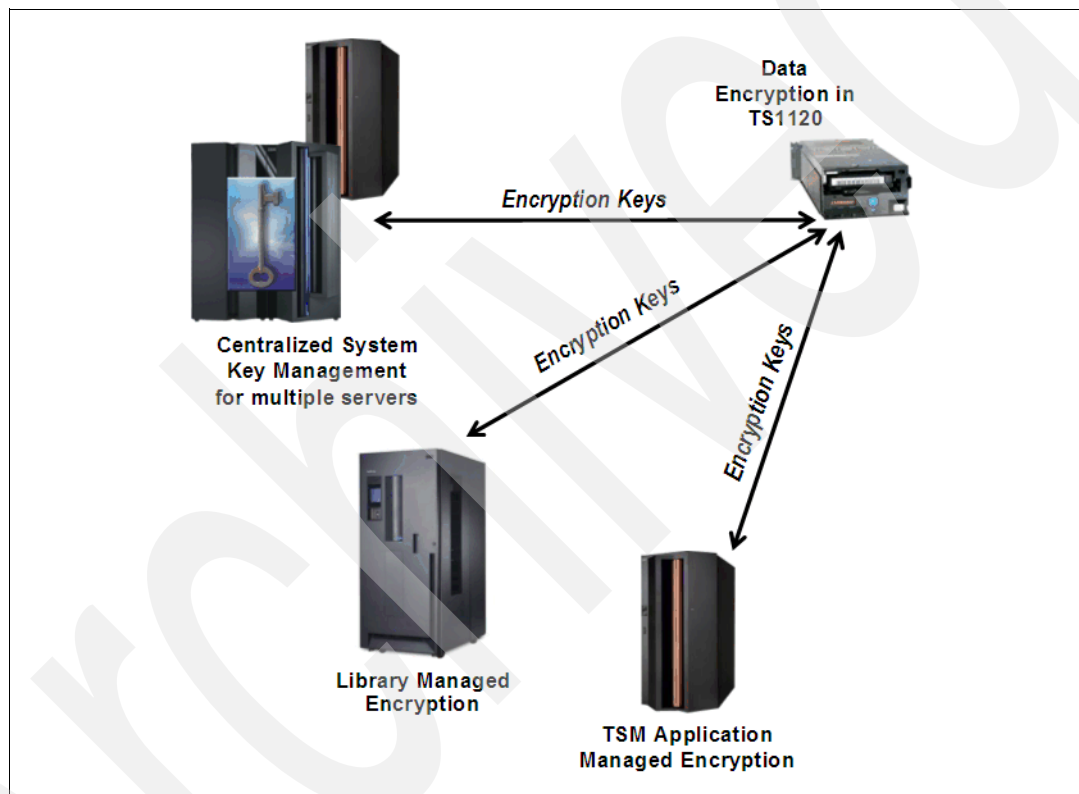


Figure 13-19 TS1120 Tape Encryption Support

The Encryption Key Manager program uses standard key repositories on supported platforms. This software has to be installed on a supported server and interfaces with the tape drive to support encryption in a System or Library Managed Encryption implementation.

Note: When encryption is enabled, the access time to data on the tape drive does increase. Also, the tape drive unload time increases. This is due to the time required to retrieve, read, and write the encryption key.

Encryption is available for System z and open systems environments. Three different methods of encryption are supported (Figure 13-20).

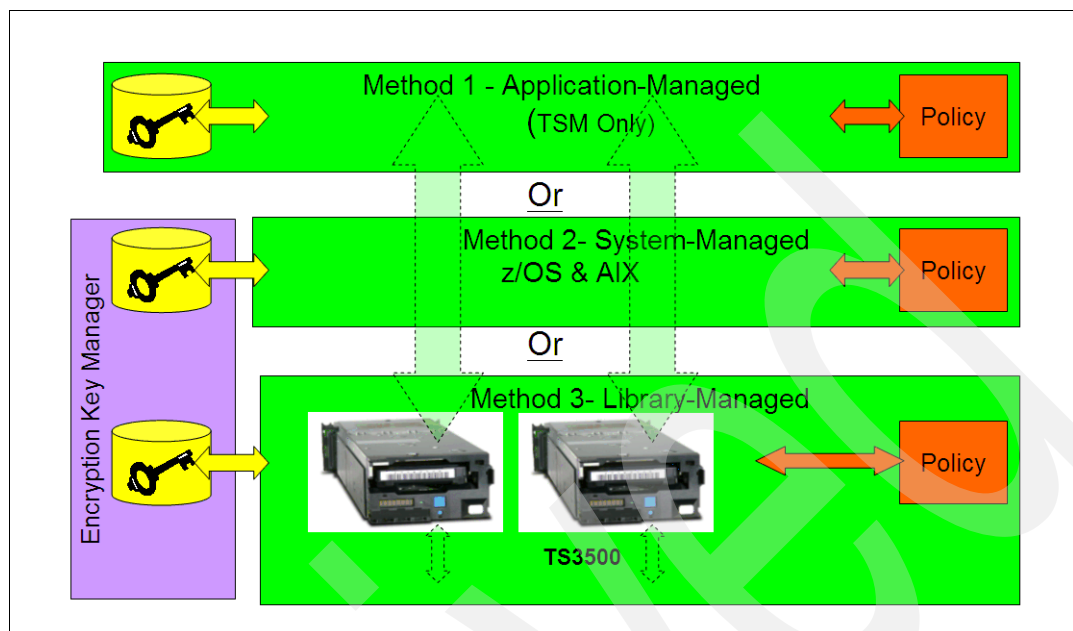


Figure 13-20 Three different methods of encryption

Application Managed

The TS1120 Tape Drive supports Application Managed Encryption for open systems environments. The application controls the encryption process, and generate and provides keys to the TS1120 tape drive. Tivoli Storage Manager is being enhanced to support this capability.

Library Managed

Encryption by volume and drive policy is supported. The user sets up and controls the encryption through the library interface. At the time of writing, this is supported for the TS3500 (3584) tape library in open systems environments. The Encryption Key Manager program is required for this support.

System Managed

With Systems Managed Encryption, the encryption policy is passed between the server and the drives. This is the only encryption method supported for z/OS environments and requires the Encryption Key Manager program. DFSMS supports the Encryption Key Manager component. Systems managed encryption will also be available for AIX. This support will require a new AIX tape device driver, as well as the Encryption Key Manager program.

System Managed Encryption on open systems is “in-band” where tape drive requests to the Encryption Key Manager component travel over the Fibre Channel channels to the server hosting the EKM.

Systems Managed Encryption on z/OS has two different configuration options. The first is “in-band”, where tape drive requests to the Encryption Key Manager component travel over the ESCON/FICON channels to the server proxy that is TCP/IP connected to the Encryption Key Manager. The second is “out-of-band”, where the tape controller establishes the communication to the Encryption Key Manager server over TCP/IP connections between the tape controller and the Encryption Key Manager server.

A router is required for out-of-band Encryption Key Management support and Encryption Key Manager connectivity for the TS1120 Tape Controller Model C06, the 3592 J70 Tape Controller, the 3952 F05 and 3953 F05 Tape Frames, the 3494 L10, L12, L14, L22 Tape Library Frame, and IBM TotalStorage 3590 C10 Subsystem.

TS1120 tape encryption implementation

Data encryption is not a new technology, however, its update has been limited by the difficulty on Encryption Key Management. IBM provides a key management solution incorporated with the tape technology to solve this long standing issue. Figure 13-21 describes the Encryption Key Generation and Communication on the TS1120 Tape Drives.

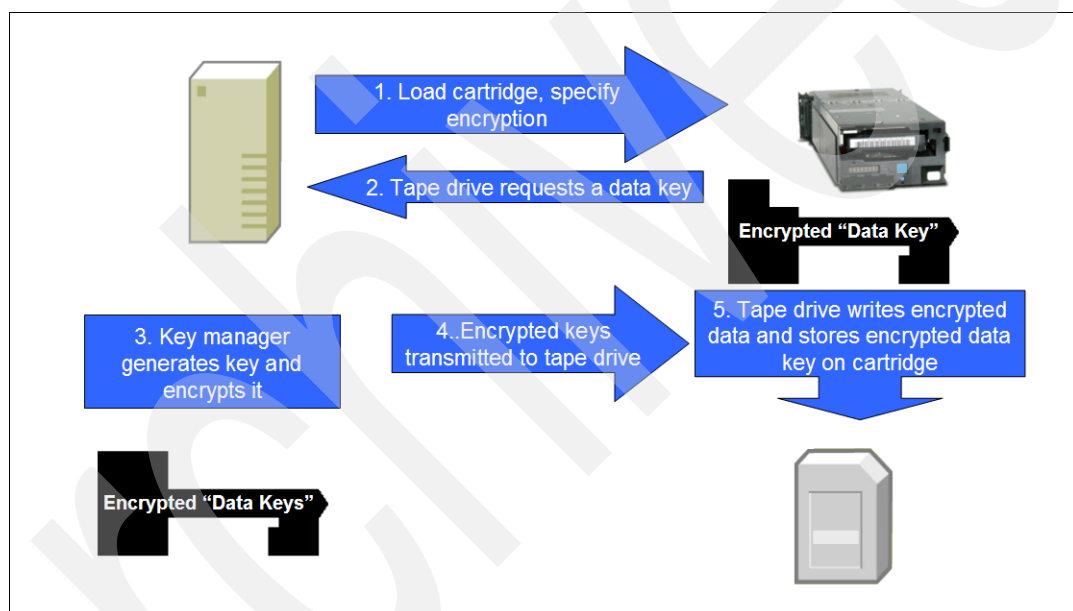


Figure 13-21 TS1120 tape encryption key generation and communication

The IBM TS1120 tape Encryption Key Management implementation is very flexible. The three encryption methods can be deployed in combination with each other (Figure 13-22).

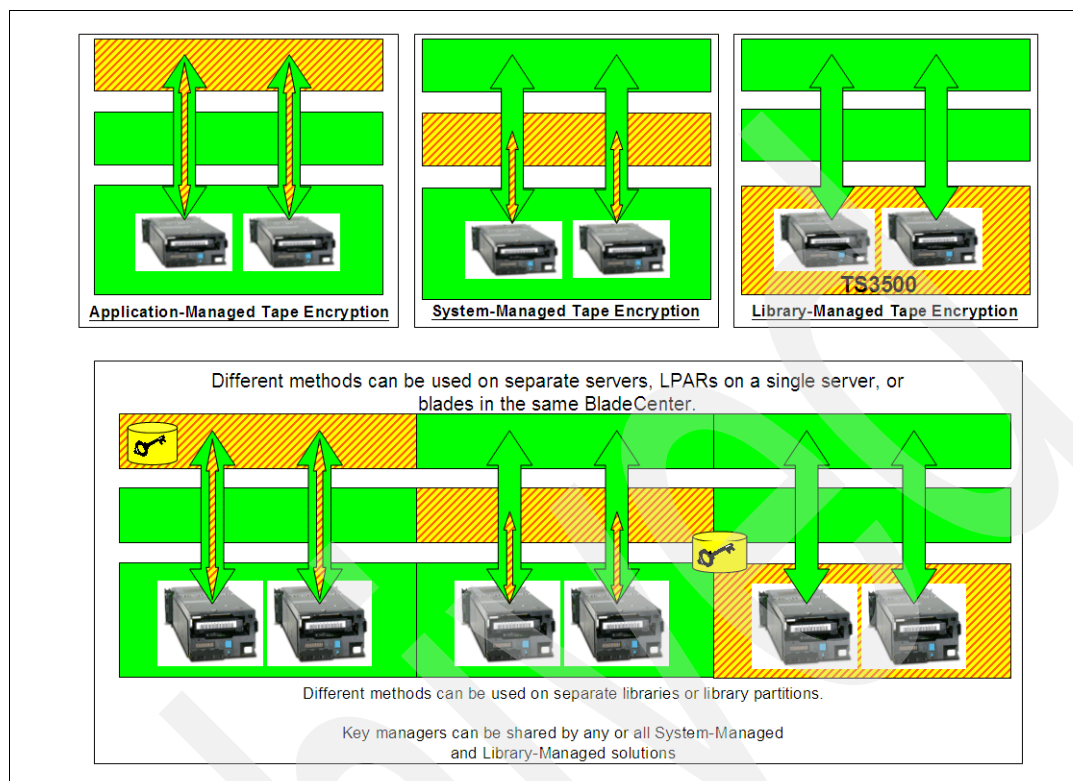


Figure 13-22 TS1120 encryption key management methods

For more details on the IBM TS1120 encryption support, see:

http://www-03.ibm.com/servers/storage/ewscast/data_encryption/

- ▶ Application Managed Encryption support information:
 - Supported Applications / ISVs: TSM
 - Supported OS: AIX, Windows, Linux, Solaris
 - Supported Storage: TS1120 in Open-attached 3584, 3494, C20 Silo, rack
 - Supported Key Managers: Provided by application
- ▶ System Managed Encryption support information:
 - Supported Applications: All applications which support zOS or the IBM AIX device driver (open systems ISV certification required)
 - Supported OS: zOS (via DFSMS), AIX (via IBM device driver) Atape 10.2.5.0
 - Supported Storage: TS1120 in 3584, 3494, C20 Silo, rack
 - Supported Key Managers: Encryption Key Manager
- ▶ Library Managed Encryption support information:
 - Supported Applications: All applications which support IBM storage listed below (open systems ISV certification required)
 - Supported OS: All open OS supported by the applications above
 - Supported Storage: 3592 in open-attached TS3500 (3584)
 - Supported Key Managers: Encryption Key Manager

Figure 13-23 summarizes the TS1120 Encryption Key support:

<u>Encryption Method</u>	<u>Policy Encrypt</u>	<u>Policy Key Label</u>	<u>Data Key Generation</u>
Application	TSM Devclass	NA	TSM
System Open	Atape Device Driver	Encryption Key Manager (EKM)	Encryption Key Manager (EKM)
System zOS	DFSMS Data Class or JCL DD	DFSMS Data Class, JCL DD or EKM	Encryption Key Manager (EKM)
Library	TS3500 (3584) Web Interface	TS3500 (3584) Web Interface or EKM	Encryption Key Manager (EKM)

Figure 13-23 TS1120 tape key encryption methods summary

13.10 Disaster recovery considerations for tape applications

In this section we discuss in general the possibilities for disaster recovery solutions with tape and without TS7700 Grid.

As discussed earlier, it is essential to define *at an application level* what conditions must be met for the restore of each application's data on tape in the event of a disaster. This means:

- ▶ What is my recovery point objective?
- ▶ What is my recovery time objective?
- ▶ What are the dependencies for the restore of the other application data?
- ▶ What is the amount of data to be restored from tape for this application?
- ▶ What are other dependencies beyond the tape restore (such as the time to recreate network infrastructure for this application)?
- ▶ Can the process run in parallel with others?

Summarizing the information for all applications, this provides us a picture of the quantity of data that must be restored within a given time. This leads to the next question: What tape infrastructure (including library, tape drives, connectivity) is required for the restore based on the backup/restore software (such as: how many restore processes can run in parallel)?

We described in Chapter "Tier levels of Business Continuity solutions" in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547, the different levels of disaster recovery solutions. If you read this, you can see that the scenarios from Tier 1 to Tier 4 are relatively technically easy and possible options, which are valid for mainframe and open systems environments.

If we discuss the higher Tiers 5-7 for data on tape we have to look at each tape application separately:

- **Backup:**

If the backup software writes its data parallel to two backup-pools, this would fit into a Tier 5. Since backup data is not actual active production data, achieving Tier 5 might not be useful.

- **Archive:**

Archive data can be active data. If the archive software could write data to two tape pools concurrently we might also reach Tier 5. In case of a disaster this running task would not have high priority. If the source data would be saved correctly in the disaster location, the archive jobs could then be restarted.

- **Hierarchical Storage Management:**

When a hierarchical storage management program migrates data from disk to tape, it is very important, that this data is also written concurrently to the disaster recovery site. DFSMSHsm, for example, on the mainframe supports the ability to write on two tapes in parallel. If the storage management software in use does not support concurrent dual copy of tape data, then other means should be considered before the migration/deletion of the source data starts (such as doing a backup before the migration task starts).

- **Batch:**

Batch processes have the advantage, that they are restartable at certain points-in-time. If you cannot tolerate a long recovery time, PtP VTS is an ideal solution. This also guarantees, that all tape data has been automatically copied to the disaster location. Nevertheless, if your service levels accept longer restart times you can tolerate a solution based on Tier 1-4.

- **System Recovery:**

On occasion single servers are lost due to hardware or software errors. Bare Machine Recovery of the system from tape is feasible in many cases. Refer to 12.7, "Bare Machine Recovery" on page 427.

- **Disaster Recovery:**

Although a site might have RTOs and RPOs that require some applications to use Tier 7 solutions, others that meet Tier 1-4 requirements might use tape for recovery. The policy-based copy management of the PTP VTS allows you to mix Tier 7 data replication requirements with those of lower tiers, providing a single cost-effective solution.

Summary

It is very important to include tape data into your disaster recovery plans. If you look at the points listed above, in most cases tape data fits very well in Tier 1-4 solutions. Virtual tape and specifically, the IBM TS7700 Grid solution fits the higher (5-7) tiers in an efficient manner. What should be considered very carefully is the backup software and the tape hardware to be used.

In Figure 13-24 you can find several solutions categorized between different values of RTO and RPO. Be aware, that software is not included. Especially the backup and recovery of databases is very sophisticated. Data loss can be minimized even if we have for example only a remote library scenario. The graphic does not reflect the complexity and the amount of data of a client's situation.

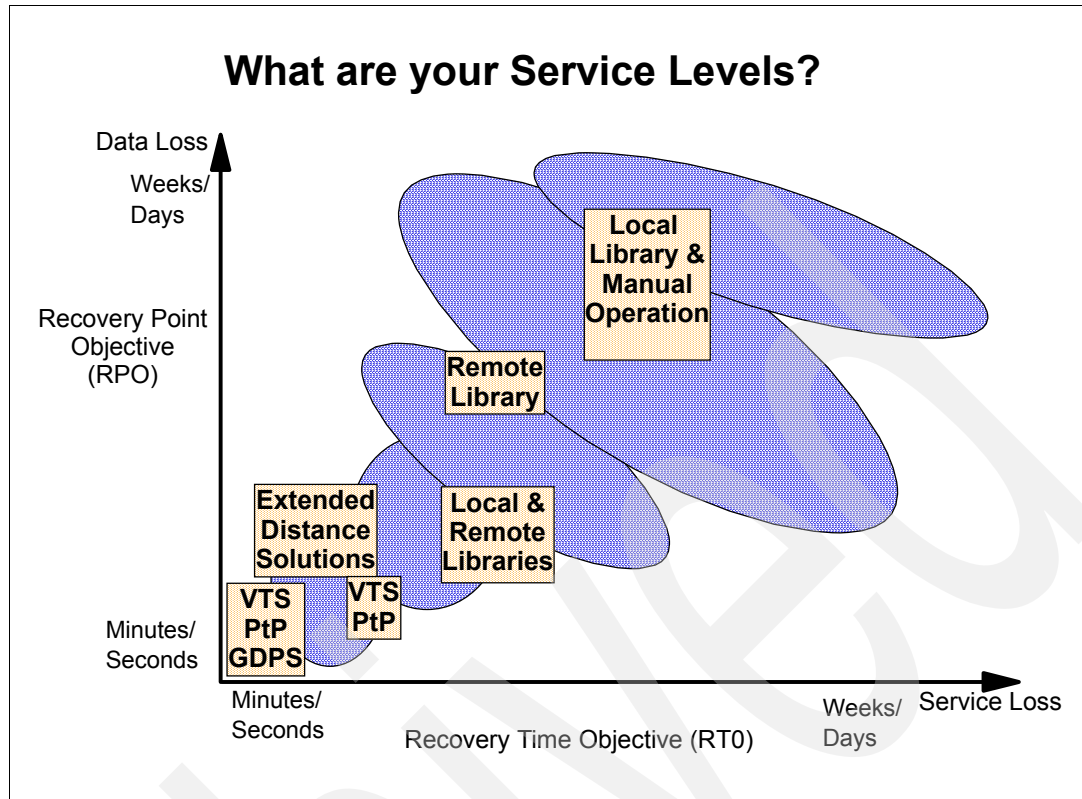


Figure 13-24 Tape solutions and disaster recovery service levels

Figure 13-25 is an example of electronic vaulting of data. Some backup software, such as Tivoli Storage Manager or DFSMSHsm, can write multiple copies of the same backup data at the same time. In this example the active Tivoli Storage Manager server creates off-site copies to a tape library at a remote site. The off-site vaulting location can have a standby server that can be used to restore the Tivoli Storage Manager database and bring the Tivoli Storage Manager server online quickly.

We discussed a tier 7 solution (TS7740 with GDPS) in 13.7.4, “TS7740 and GDPS (Tier 7) Implementation” on page 469.

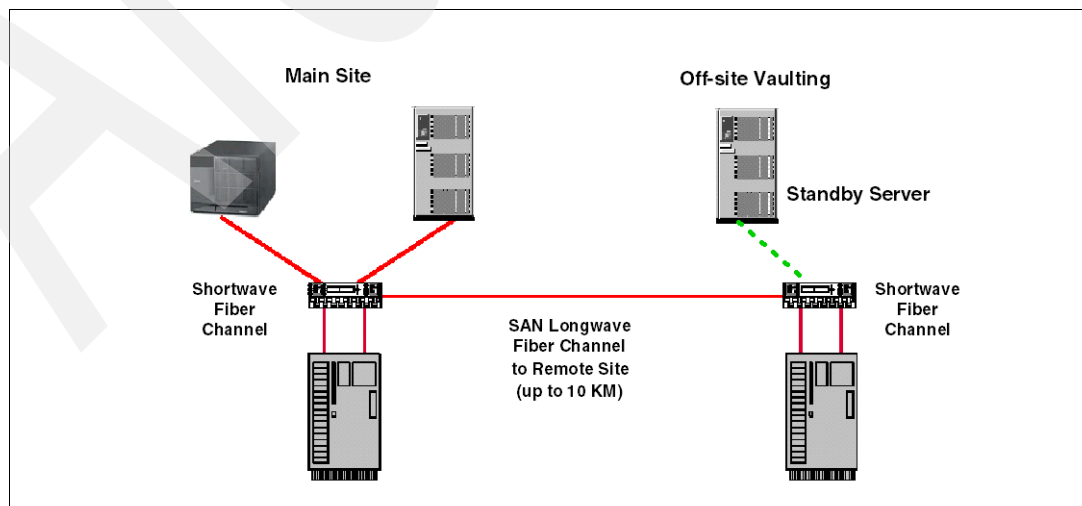


Figure 13-25 Remote tape vaulting

Archived



System Storage resources and support

This appendix provides you with descriptions of the various IBM resources that are in place to assist you and your organization in finding answers to your questions on the capabilities and features of IBM System Storage Solutions.

Where to start

The recommended place to start when searching for information about System Storage products is the external IBM System Storage products homepage available at:

<http://www.ibm.com/servers/storage>

We also suggest you browse the storage solutions section of the IBM System Storage products homepage at:

<http://www-03.ibm.com/servers/storage/solutions/>

The solutions section illustrates how multiple product and IBM Global Services components can be aggregated to deliver solutions. Both these sites contain a wealth of information about IBM System Storage products and solutions.

Advanced Technical Support: System Storage

Advanced Technical Support (ATS) provides pre-sales technical information and expertise to IBM sales teams, Business Partners, and Clients. As part of the pre-sales process, ATS System Storage can offer the following services.

- ▶ Presales technical information and expertise
- ▶ Bandwidth analysis
- ▶ Non-billable proof of concepts at regional lab facilities.

IBM employees can contact ATS through the internal Global Technical Sales Support Web site's Tech Xpress form. Business Partners can contact ATS through Partnerline or through their IBM representative.

Refer to the following IBM **internal** Web sites for additional information:

<http://w3-03.ibm.com/support/techxpress.html>

- ▶ For global technical sales support:

<http://w3-03.ibm.com/support/techdocs/atmastr.nsf/Web/Techdocs>

- ▶ For the technical information database and for technical questions and answers:

<http://w3-03.ibm.com/support/techdocs/atmastr.nsf/WebIndex/prs1223>

IBM System Storage Solution Centers

IBM System Storage Solution Centers can deliver one-stop shopping for storage hardware, software, and consulting services. The Solution Centers offer you both a local venue for a hands-on test drive of IBM storage solutions and a platform for proof of concept and benchmarking activities. IBM Business Partners will provide the expertise to help you select and implement the right solution to enable your business to succeed in today's dynamic marketplace.

There are, as of this writing, over 100 IBM System Storage Solution Centers worldwide. To schedule a visit and test drive your storage solution today, visit:

<http://www-1.ibm.com/servers/storage/solutions/tssc/>

The locator on this Web page provides information about the Solution Center program and how it could be the perfect answer to all your storage questions.

IBM System Storage Proven Solutions

The IBM System Storage Proven™ Solutions provide pre-tested configurations from a broad portfolio of products and solutions. This means that applications and hardware you choose from leading companies, combined with IBM state-of-the art technology, provide you with flexible, ready to run, easily installed solutions that address your software and infrastructure requirements.

Storage Proven means integrated solutions are more than just hardware and software; they are validated together for specific applications and configurations. IBM takes the guesswork out of putting a total solution together to help simplify your purchase decision, freeing you to focus on running your business.

IBM System Storage Proven Solutions are tested rigorously so that you are assured that work together reliably and efficiently.

System Storage Proven Solutions continue to add new products and partners to a growing list of compatible solutions for your business. They understand the value of the System Storage Proven program and the potential it brings for increasing demand for their products by working with the broad IBM storage portfolio. Together with IBM, they are dedicated to increasing client choices when selecting storage solutions.

Check this Web site for the most recent list of nominated products:

http://www.ibm.com/servers/storage/proven/all_solutions.html

IBM Global Services: Global Technology Services

IBM Global Technology Services offers several different solution offerings in this area of Disaster Recovery. See the following sections for their descriptions.

Solutions integration

IBM Global Services is able to work as a complete solutions integrator. As disaster recovery solutions depend on Servers, Storage, and Network Connectivity, IBM Global Services can be of great assistance in providing some or all parts of the solution and the expertise to bring everything together. Additionally, some offerings, such as GDPS and eRCMF, are available only as solutions from IBM Global Services

GDPS Technical Consulting Workshop

When the decision has been made to implement any of the GDPS Service Offerings, the first step is a three day Technical Consulting Workshop (TCW). During the TCW, IBM Global Services comes to your location and works with all parts of your business. The end result of this workshop will be:

- ▶ Business justification required to show value to decision makers
- ▶ A high level project plan for implementation
- ▶ a Statement of Work which will provide estimates of hours required and costs associated with the implementation.

The TCW is a fee-based service workshop, however, those costs will be deducted from the price of the GDPS implementation when it moves forward. IBM sales teams and Business Partners can arrange for a TCW through either their Geography's GDPS Marketing Team or by contacting their IBM Global Services representative.

Broad-ranged assessment

Tremendous growth in server system capacities and business demands for round-the-clock operations often challenge you to sustain a high level of system availability. An IBM services specialist performs the system assessment by working with your representatives from technical support, application development, and operations. In a workshop we review your specific availability requirements, assess current practices, identify new options and techniques, and recommend further enhancements to minimize planned and unplanned system outages

With this service, we assess your current practices for:

- ▶ Availability management
- ▶ Backup and recovery management
- ▶ Change management for hardware and software currency
- ▶ Problem management
- ▶ Security management
- ▶ Operations management
 - Storage management
 - Job scheduling
 - Event and error monitoring
- ▶ Performance and capacity management
- ▶ Support resources
 - Staffing and skills
 - System documentation
 - Technical support

The summary report we create as a result of the workshop gives your team a hardcopy review of our recommendations.

Our workshop and summary report helps define high-priority tasks. Knowing the high-priority actions to take helps you:

- ▶ Increase overall system availability
- ▶ Reduce backup and recovery times
- ▶ Improve change management for hardware and software currency
- ▶ Minimize risk caused by system changes
- ▶ Manage your problems more effectively
- ▶ Improve system security
- ▶ Improve operations
- ▶ Improve system performance
- ▶ Improve performance and capacity management
- ▶ Increase your support resource efficiency for staffing and skills, system documentation, and technical support

For more information visit the Web site at:

<http://www.ibm.com/services/pss/us>

Business Resiliency & Continuity Services (BRCS)

The IBM Business Resiliency & Continuity Services (BRCS) is an organization within IBM Global Services that is dedicated solely to the business continuity concerns of clients of IBM. BRCS professionals have the expertise and tools necessary to design the right Business Continuity Plan for your enterprise. Whether an individual component or an end-to-end solution, their services include:

- ▶ Assessment of continuous operation readiness for critical processes.
- ▶ Development of in-depth business continuity strategies that map to business and IT requirements.
- ▶ Solution design, encompassing proven continuous-availability techniques and risk management, as well as traditional disaster recovery disciplines, processes, and methodologies.
- ▶ Integration of Business Continuity with critical business applications and IT initiatives, including e-business, enterprise resource planning (ERP), availability management, asset management, and server consolidation.
- ▶ Documented plans for the entire enterprise or individual business unit that integrate the full range of business continuity and recovery strategies.
- ▶ Transformation of plans into detailed procedures and processes, including testing to help evaluate readiness.
- ▶ Strategy proposals to help prevent high-impact risks and emergency situation management preparation.
- ▶ Validation of existing recovery assumptions, including shortfalls. This validation also can include testing your plan to simulate real-life disaster declarations.

As a leading provider of business continuity and recovery solutions, IBM delivers distinct advantages. In addition to decades spent perfecting our own business continuity programs, we offer an unrivaled track record of helping companies anticipate, prevent, and recover from the disruptions that impact their business operations.

Our people understand the role technology plays in business and the impact a technology disruption can have. Every solution we develop takes into consideration both the immediate and long-term impacts a disruption can have on your business.

Whether a single facility disruption or a major disaster of regional or worldwide proportions, we have the knowledge, experience, and tools required to help get your business operational.

When it comes to business continuity and recovery solutions, IBM is a name you know, a company you can trust, people you can count on. Whatever the size of your company, scope of e-business, or type of IT platform, rely on the people of IBM Business and Continuity Recovery Services.

The following are just a few of the facts about IBM that we think you should be aware of as you consider your disaster recovery alternatives:

- ▶ Thousands of business continuity and recovery specialists worldwide
- ▶ Over 12,000 contracts
- ▶ Experience from over 400 recoveries
- ▶ More than 20,000 test events
- ▶ Over 100 recovery facilities worldwide
- ▶ ISO 9001 certified
- ▶ Highest client satisfaction rating in the industry

IBM Business Continuity and Recovery Services have received three top industry awards recently, including:

- ▶ Reader's Choice Award - *Today's Facility Manager Magazine*
- ▶ Hall of Fame Award - *Contingency Planning & Management Magazine*
- ▶ 1999 Solution Integrator Impact Customer Satisfaction Award - *Solution Integrator Magazine*

ClusterProven program

High availability has become a critical requirement for virtually every industry in the business world today. A computer system failure quickly results in the failure of a business to operate, and every minute of downtime means a minute of lost revenue, productivity, and profit.

For systems to achieve continuous operations, all of the system components (hardware, operating system, middleware, and applications) must be implemented in a fashion where failures are transparent to the user. In addition, mechanisms must be in place to allow system maintenance without disrupting the current workload.

The IBM ClusterProven® program sets specific criteria for validating end-to-end solutions that meet industry standards for high availability on every IBM platform. Developers who achieve ClusterProven validation earn the use of the ClusterProven mark, which may be used in marketing their solution to clients and receive co-marketing assistance from IBM.

Benefits of ClusterProven validation

The ClusterProven program eliminates a significant amount of guesswork for clients that are deploying a high-availability solution on an IBM platform. When clients choose the combination of IBM hardware and ClusterProven software, they gain the high-availability features of clustering in a pretested configuration that cuts implementation time.

In addition, ClusterProven validation is an excellent investment for developers in the years ahead, as close to 100 percent availability becomes an absolute requirement for e-business. ISVs who have achieved ClusterProven validation will continue to be technology front-runners, offering broader and deeper solutions than many of their competitors. ClusterProven validation is the final piece of the availability puzzle, keeping developers very competitive in the Internet economy.

ClusterProven solutions

Developers' solutions that have been validated as ClusterProven are noted as such in the Global Solutions Directory, which receives over 1.6 million page visits monthly. Applications in the GSD may be sorted by ClusterProven and then by solution name, company name, industry, or IBM for ease in locating a solution that fits the requirements of the user. See the list of IBM products that have achieved the ClusterProven validation at:

<http://www-1.ibm.com/servers/clusters>

Support considerations

Selecting the right server and SAN storage infrastructure for a specific project in an multi vendor, distributed system scenario designed for high availability and disaster tolerance has become a sizeable challenge. You have to get the appropriate support from all involved parties and vendors.

The most important questions that have to be answered are:

- ▶ Who is responsible for support on which component?
- ▶ Which level of support is guaranteed by whom?
- ▶ Will the support I receive for each component fit in my SLA (Service Level Agreement)?

System dependency matrix

In Figure A-1 the dependencies between the solution building blocks consisting of the Server Building Block, the SAN Building Block and the Storage System Building Block are displayed. We discuss their individual relationships and dependencies in detail.

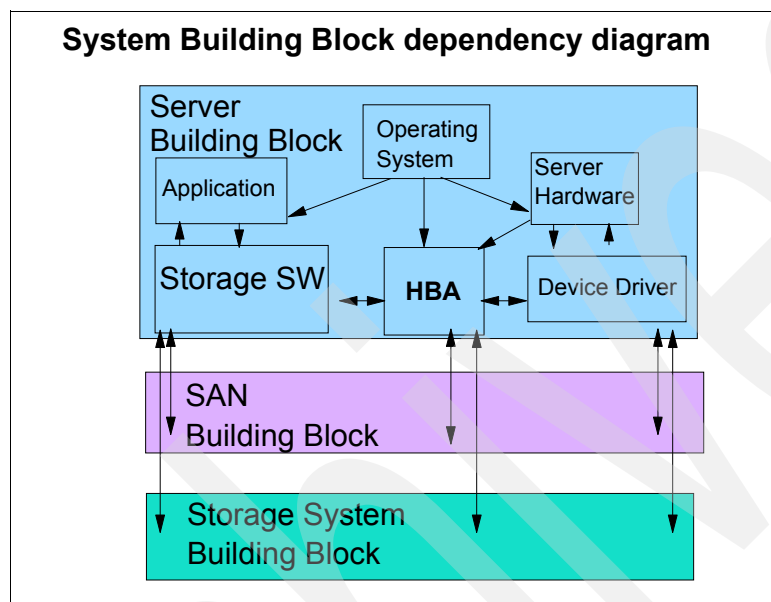


Figure A-1 System dependency matrix

To match the specific requirements projected for the solution and to determine if a proposed solution consisting of server, server options (such as adapters), operating system, SAN components and the storage subsystem is entirely supported, you must have a positive statement of support and compatibility from all of the involved vendors. This leads to the following questions that have to be answered for each of the building blocks:

1. Server Building Block:

The server hardware in combination with the Host Bus Adapters, the storage software, the operating system, the application and the projected storage solution must be compatible with each of the building blocks components. For example certain applications such as SAP also require hardware certification at the server level.

Questions regarding the Server Building Block:

- Is the operating system supported on the server hardware?
- Is the application supported on the hardware and the operating system?
- Is the Host Bus Adapter supported for the operating system in combination with the storage software (such as IBM Tivoli Storage Manager) and the Server hardware?
- Is the Server type itself supported by the Storage System?
- In the case of a High Availability solution (such as MS Cluster, or HACMP) in addition to the above, the combination of Server, HBA and Storage also has to be officially certified from the vendor of the operating system, to provide a certified and supported solution.

2. SAN Building Block:

The SAN building block refers to the components being used to establish the SAN infrastructure such as SAN Switches, Directors, Gateways and Storage Virtualization Controllers.

Questions regarding the SAN Building Block:

- Is the SAN building block component supported by the storage subsystem vendor?
- Is the storage subsystem supported by the SAN Building Block vendor?
- Are the Host Bus Adapter(s) used in the servers supported and the prerequisites such as firmware levels and microcode match the level of support for the SAN Building Block and the Storage Subsystem Building Block?

3. Storage System Building Block:

The storage system building block refers to the storage components in the system.

Questions regarding the Storage System Building Block:

- Are the specific device drivers for the Host Bus Adapters in the server supported?
- Is the required functionality (such as path failover, load balancing, tape support) provided?
- Is the Server and Storage hardware supported for specific storage software (such as, VERITAS Volume Manager, IBM Tivoli Storage Manager)?

Solution Assurance

By following the IBM Solution Assurance Process, experts will review your proposal and reduce your risk, while saving you time and problems. A pre-sale **Solution Assurance Review (SAR)** is a technical inspection of a completed solution design. Technical Subject Matter Experts (SMEs) who were not involved in the solution design participate to determine:

- ▶ Will it work?
- ▶ Is the implementation sound?
- ▶ Will it meet the client's requirements and expectations?

In a pre-install SAR, SMEs also evaluate the client's readiness to install, implement, and support the proposed solution. Where do you get support for the Solution Assurance process? Check on the IBM intranet (**internal**) SAR Homepage about the current Solution Assurance Process.

The IBM Solution Assurance Web site:

<http://w3.ibm.com/support/assure/assur30i.nsf/Web/SA>

If a solution does not pass the Solution Assurance Review process due to the probability of incompatibility or other obstacles, IBM can do testing and certification of unsupported configurations on an individual scenario level by raising an RPQ (**R**esult **P**rice **Q**uotation) for IBM System Storage Systems or a SPORE (**S**erver **P**roven **O**pportunity **R**esult for **E**valuation) process, if System x systems are components within the proposal. Your IBM representative will be the primary contact for requiring a SPORE or an RPQ for individual solutions.

Recommendations

Here are our recommendations:

1. **Get advice.** Before making a proposal to a client, make sure it is from independent experts, that are not prejudiced when it comes to platform, storage, and software vendor choices. Ensure your advisors have proven track records!
2. **Don't just throw money and technology at the problem.** Make sure you thoroughly understand what you want to achieve — not simply from a technical standpoint but also considering the business demands that are driving those technical changes.
3. **Calculate current Total Cost of Ownership (TCO).** A primary motivator for a change in storage strategy is to reduce costs. However, very few people establish what their current TCO is and therefore have no fiscal barometer to either justify new investment or measure the success of the new project.
4. **Identify the business goals.** Your project and your client's satisfaction will be measured by the level of success on reaching those goals.
5. **Avoid assumptions.** Everyone wants *bigger* and *faster*, but where is it best applied and how big does it have to be? By understanding exactly who is using what capacity, where, when, and how, you can design a scalable solution from an informed position based on fact rather than assumption.
6. **Conserve the appropriate architecture.** Many organizations will have significant investments in under utilized equipment and software, much of which can be redeployed. Identify the scope of the SAN and the server design. Server and Storage consolidation methods are a good proposal for a solution.
7. **Avoid complexity.** With so many vendors, products, and new technologies it is very tempting to over-engineer a new storage solution. Don't be tempted. If you build in complexity, then you build in cost and inflexibility — not things you want to burden your client's business within times of changing technology and economic unpredictably.
8. **Be a consultant for your client.** Understand really what are his requirements and which products are best to meet those requirements in a solution. What skills are available with your client's IT team? Set a measurable criterion for success!
9. **Establish a project team.** The team must be made accountable and responsible for the future of this project. Have a professional project team with a leader who takes over the responsibility of the project to have the solution properly implemented and deployed. IBM Global Services offers a wide variety of services that will cover all areas regarding the implementation of the solution.
10. **Help your client to define SLAs.** By using SLAs and TCO analysis, your client at the end has an instrument that allows him to track the cost of his solution to calculate his ROI.

Support links to check

In the following support links you will find the compatibility and support matrix for each vendor's products. It is important that when proposing a solution based on a variety of building blocks, that all items are cross certified and supported with each other.

This list of links is not complete, but refers to the IBM Server and System Storage proven program as well as to the compatibility links for various vendors.

IBM Server Proven System x

<http://www.ibm.com/servers/eserver/serverproven/compat/us>

IBM System Storage Solutions for System p

<http://www.ibm.com/servers/storage/product/p.html>

IBM System Storage Solutions for System i

<http://www.ibm.com/servers/storage/product/i.html>

IBM System Storage Solutions for System z

<http://www.ibm.com/servers/storage/product/z.html>

IBM System Storage Proven

<http://www.ibm.com/servers/storage/proven>

CISCO

<http://cisco.com/>

McDATA

<http://www.mcdata.com>

Brocade

<http://www.brocade.com>

Software certification dependencies:

Tivoli

<http://www.ibm.com/software/sysmgmt/products/support>

Microsoft

<http://www.microsoft.com/windows/catalog/server/>

VERITAS

http://seer.support.veritas.com/nav_bar/index.asp?content_sURL=/search_forms/techsearch.asp

SUSE Linux

<http://www.novell.com/partners/yes/>

Red Hat Linux

<http://hardware.redhat.com/hcl>

Oracle

<http://otn.oracle.com/support/metalink/index.html>

IBM System Storage Copy Services function comparison

As shown in Figure B-1, IBM System Storage offers a wide array of disk systems with varying capabilities. The purpose of this chart is to show a general, though not absolute, statement of positioning in terms of capabilities. In the rest of this appendix we attempt to further quantify the capabilities of each of our disk system offerings.

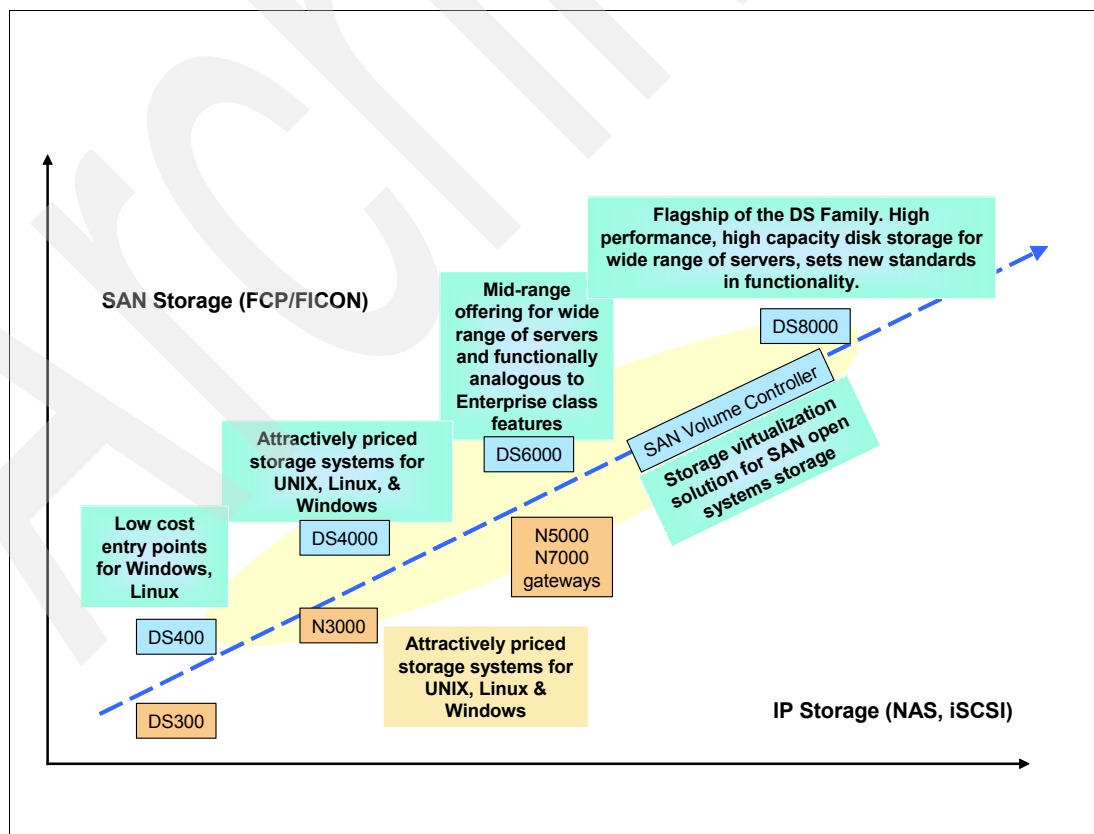


Figure B-1 General positioning of IBM System Storage

IBM System Storage point-in-time copy comparison

Figure B-2 provides a quick reference to the point-in-time copy (FlashCopy) functions available on the IBM System Storage SAN Volume Controller, DS4000, ESS - DS6000 - DS8000, and N series.

Each box has one of the following indicators:

- ▶ **Y** – function is currently generally available
- ▶ **N** – function is not available

The table is intended as a guideline only; consult the appropriate product chapters and product documentation for details of the individual product implementations.

As product enhancements are continually made, see your IBM representative for the latest information.

	Flashcopy Feature	SVC	DS4000	ESS	DS6000	DS8000	N series
Base	Open File level	N	N	N	N	N	Y
	z/OS Data Set	N	N	Y	Y	Y	N
Speed	Volume/LUN level	Y	Y	Y	Y	Y	Y
	Multiple concurrent copies	N	Y	Y	Y	Y	Y
Scalability	Copy source/target R/W avail immediately	Y	Y	Y	Y	Y	Y, Target=Read/Only
	Physical copy	Y	Y	Y	Y	Y	N
Usability	Logical copy (no copy)	Y	Y	Y	Y	Y	Y
	In-band Copy to remote site	N	N	Y ^D	Y ^D	Y ^D	N
TCO	Incremental copies	N	N	Y	Y	Y	N
	Across any-to-any storage arrays	Y ^A	N	N	N	N	Y ^C
	"Reverse" Copy	N	N	Y	Y	Y	Y ^B
	Consistency Groups	Y	N	Y	Y	Y	N
	Persistent copies	Y	N	Y	Y	Y	Y
	Transition nocopy => copy	Y	N	Y	Y	Y	N/a
	Logical copy (space efficient)	N	Y	N	N	N	Y

^A -- if attachment to SVC is supported ^B -- Using SnapRestore software ^C -- N series Gateways ^D -- System z only

Figure B-2 Point-in-time copy product function comparison

FlashCopy function definitions

File Level: Is the point-in-time copy performed at the file level?

z/OS Data Set: Is it possible to perform the point in time copy at the data set level in z/OS environments?

Volume / LUN Level: Is the point-in-time copy performed at the block / LUN / volume level?

Multiple Concurrent Copies: Is it possible to make multiple concurrent copies from the same source at the same point in time?

Copy Source/Target Available immediately: Will the target and source data be available immediately while the copy proceeds in the background or must the full copy be completed first.

Physical Copy: Can a physical copy be made of the entire source to the target?

Logical Copy (no background copy): Does the capability exist to not copy the complete source LUN / volume? For example, ESS/DS6000/DS8000 NOCOPY, or SAN Volume Controller 0 background copy rate parameter.

Inband Copy to Remote Site: Is it possible to issue PiT copy commands across a remote copy link to the disk system in the remote site?

Incremental Copies: Is there an option to update only changed data on an ongoing basis, or must each PiT be a completely new copy with all data being transferred?

Source / Target Read / Write Capable: Are both the source and target fully read / write capable?

Incremental Copies: Are changes to the source tracked so only changed data is copied to the target?

Any to Any across Storage arrays: Is it possible to make a Point In Time copy that resides on a separate physical disk system than its source?

“Reverse” Copies: Is it possible to copy from target to source, backing out any changes that have been made to the source since the last PiT?

Consistency Groups: Can associated volumes be treated as one or more groupings and have the point-in-time operation performed across the grouping at the same point-in-time? Consistency groups were described in detail in Chapter 6 “Planning for Business Continuity in a heterogeneous IT environment” in *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547.

Persistent Copy: Is it possible to create PiT copy relationships that continue to exist until explicitly withdrawn?

Logical Copy (space efficient): Does the capability exist to utilize less target space than occupied by the source?

Transition Nocopy → Copy: Does the capability exist to change no background copy to a full Copy without doing a new point-in-time copy operation?

Heterogeneous disk subsystems: Can the point-in-time copy have source and target on different vendor disk subsystems?

IBM System Storage disk mirroring comparison

The following table provides a quick reference to the disk mirroring (Synchronous and Asynchronous) functions available on the IBM System Storage SAN Volume Controller, DS4000, ESS - DS6000 - DS8000, and N series

Each box has one of the following indicators:

- ▶ **Y** – function is currently generally available
- ▶ **N** – function is not available

Figure B-3 is intended as a guideline only; consult the appropriate product chapters and product documentation for details of the individual product implementations.

As product enhancements are continually made, see your IBM representative for the latest information.

Mirroring Features		SVC	DS4000	ESS	DS6000	DS8000	N series
Base	Synchronous Mirror	Y	Y	Y	Y	Y	Y
	Synchronous Distance Supported	100km ^{A,B}	100km ^{A,B}	300km ^A	300km ^A	300km ^A	100km ^A
	Asynchronous Mirror	Y (4.1)	Y	Y	Y	Y	Y
	z/OS specific Asynchronous Mirror	N	N	Y	Y	Y	n/a
	3 site synchronous to asynchronous mirror	N	N	Y	N	Y	Y
Scalability	Failover / Failback support	Y	N	Y	Y	Y	Y
	Suspend / resume support	Y	Y	Y	Y	Y	Y
	Maximum Consistency Group size	1024	64 ^(c)	No limit	No limit	No limit	N/a
ability	Consistency Groups	Y	Y ^c	Y	Y	Y	Y
	Consistency Group Freeze support	Y	N	Y	Y	Y	n/a- uses diff.tech.
Usability	Across any-to-any storage arrays (Open)	Y	N	N	N	N	Y ^d
	Across any-to-any storage arrays (System z)	N	N	Y (with zGM)	N (no zGM)	Y (with zGM)	N
	Dynamically switch Synchronous to Asynchronous	N	Y	N	N	N	N
	Primary / Secondary same cluster	Y	N	Y	Y	Y	Y

^A -- with channel extenders; ^B -- longer distances via RPO; ^C -- for DS4000 Global Mirror only ^D -- Gateways

Figure B-3 Disk mirroring product function comparison

Disk mirroring function definitions

File Level: is the point-in-time copy performed at the file level?

Block / LUN Level: is the point-in-time copy performed at the block / LUN level / volume level?

Synchronous copy - host notified of Write Complete after Source and Target I/O have completed.

Distance Support: SAN Volume Controller (Metro Mirror) – 100 km Fibre Channel Protocol with channel extenders or IP/DWDM based solutions, unlimited distance support through Asynchronous. SVC Metro Mirror is further described in 11.2.3, “SVC copy functions” on page 369.

ESS / DS6000 / DS8000(Metro Mirror)– 303 km over Fibre Channel or IP/DWDM based network; unlimited distance supported for async mirroring. Metro Mirror is further described in 7.7, “Copy Services functions” on page 270.

DS4000(ERM Metro Mirror)— 10-50 km over FCP; channel extender IP/DWDM based solutions for longer distances, unlimited distance in asynchronous. ERM is further described in 8.7.5, “Enhanced Remote Mirroring and VolumeCopy Premium Feature” on page 310.

N series (Sync Mirror or SnapMirror)-- 300m over FCP, 100km over a switched FCP or IP network. Unlimited distances in asynchronous. Sync Mirror and SnapMirror are both described in 9.3, “N series software overview” on page 334.

Asynchronous copy with consistency: target I/O applied in with data integrity; data suitable for database restart at remote site.

zOS Specific Asynchronous copy: is there a z/OS specific form of mirroring?

Three site (synchronous/asynchronous) mirror: Is it possible to mirror to multiple disk systems, either through a cascade or by mirroring to multiple targets in order to protect data at more than a primary and secondary disk?

Failover / Failback Support: is the capability to quickly re-establish Target → Source relationships in event of a Primary site failure, copying only incremental changes back.

Consistency Groups: can associated volumes be treated as one or more groupings such that disk mirroring error triggers cause suspension and keep data integrity across an entire grouping of volumes?

Consistency Group Freeze Support: Can the volumes associated in a consistency group use internal or external control software to cause a halt to all mirroring in the recovery site in order to preserve all recovery data at one consistent point in time?

Suspend / Resume: can the remote copy environment be suspended in the event of a planned or unplanned outage and then resumed? This means only changes while the remote copy environment is suspended are copied to the targets versus the entire remote copy environment must be copied in its entirety.

Maximum Consistency Group size: How many volumes can be protected by a single consistency group?

Any to Any Storage Arrays: Can the disk mirroring technology write to unlike disk systems?

Dynamically Switch from Synchronous to Asynchronous: Is it possible to switch from synchronous to asynchronous if the requirement arises? For example, situations in which write activity peaks and synchronous cannot keep up with writing to both locations.

Primary Secondary Same Cluster: Can you do remote copy by mirroring from one set of disk to another within the same disk system?

Archived

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

IBM Redbooks

For information about ordering these publications, see “How to get IBM Redbooks” on page 506. Note that some of the documents referenced here may be available in softcopy only:

- ▶ *IBM System Storage Business Continuity Solutions Overview*, SG24-6684
- ▶ *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547
- ▶ *IBM System Storage Virtualization Engine TS7700: Tape Virtualization for System z Servers*, SG24-7312
- ▶ *IBM TotalStorage Productivity Center for Replication on Windows 2003*, SG24-7250
- ▶ *Using IBM Tivoli Storage Manager to Back Up Microsoft Exchange with VSS*, SG24-7373
- ▶ *IBM Tivoli Storage Manager for Advanced Copy Services*, SG24-7474
- ▶ *Disaster Recovery Using HAGEO and GeoRM*, SG24-2018
- ▶ *IBM TotalStorage Enterprise Storage Server Implementing ESS Copy Services with IBM eServer zSeries*, SG24-5680
- ▶ *IBM TotalStorage Enterprise Storage Server Implementing ESS Copy Services in Open Environments*, SG24-5757
- ▶ *Implementing High Availability Cluster Multi-Processing (HACMP) Cookbook*, SG24-6769
- ▶ *IBM System Storage DS6000 Series: Copy Services with IBM System z*, SG24-6782
- ▶ *IBM System Storage DS6000 Series: Copy Services in Open Environments*, SG24-6783
- ▶ *IBM System Storage DS6000 Series: Copy Services with IBM System z*, SG24-6782
- ▶ *IBM System Storage DS8000 Series: Copy Services with IBM System z*, SG24-6787
- ▶ *IBM System Storage DS8000 Series: Copy Services in Open Environments*, SG24-6788
- ▶ *IBM System i5, eServer i5, and iSeries Systems Builder IBM i5/OS Version 5 Release 4 - January 2006*, SG24-2155
- ▶ *Clustering and IASPs for Higher Availability on the IBM eServer iSeries Server*, SG24-5194
- ▶ *Introduction to Storage Area Networks*, SG24-5470
- ▶ *iSeries in Storage Area Networks A Guide to Implementing FC Disk and Tape with iSeries*, SG24-6220
- ▶ *IBM System Storage SAN Volume Controller*, SG24-6423
- ▶ *IBM System Storage DS8000 Series: Concepts and Architecture*, SG24-6786
- ▶ *IBM System Storage DS6000 Series: Concepts and Architecture*, SG24-6781
- ▶ *IBM eServer iSeries Independent ASPs: A Guide to Moving Applications to IASPs*, SG24-6802
- ▶ *Sysplex eBusiness Security z/OS V1R7 Update*, SG24-7150

- ▶ *Disaster Recovery with DB2 UDB for z/OS*, SG24-6370
- ▶ *IBM Tivoli Storage Management Concepts*, SG24-4877
- ▶ *IBM Tivoli Storage Manager Version 5.3 Technical Guide*, SG24-6638
- ▶ *IBM Tivoli Storage Manager in a Clustered Environment*, SG24-6679
- ▶ *Get More Out of Your SAN with IBM Tivoli Storage Manager*, SG24-6687
- ▶ *Disaster Recovery Strategies with Tivoli Storage Management*, SG24-6844
- ▶ *Z/OS V1R3 and V1R5 DFSMS Technical Guide*, SG24-6979
- ▶ *IBM System Storage DS4000 Series, Storage Manager and Copy Services*, SG24-7010
- ▶ *Understanding the IBM System Storage DR550*, SG24-7091
- ▶ *The IBM System Storage N Series*, SG24-7129
- ▶ *Using the IBM System Storage N Series with IBM Tivoli Storage Manager*, SG24-7243
- ▶ *IBM System Storage: Implementing an Open IBM SAN*, SG24-6116
- ▶ *SAN Multiprotocol Routing: An Introduction and Implementation*, SG24-7321

Other publications

These publications are also relevant as further information sources:

- ▶ *IBM TotalStorage Productivity Center for Replication: Installation and Configuration Guide* SC32-0102,
- ▶ *IBM TotalStorage Productivity Center for Replication: User's Guide*, SC32-0103
- ▶ *IBM TotalStorage Productivity Center for Replication: Command-Line Interface User's Guide* SC32-0104
- ▶ *IBM TotalStorage Productivity Center for Replication: Quick Start Guide* GC32-0105
- ▶ *IBM TotalStorage Productivity Center for Replication: Two Site BC Quick Start Guide* GC32-0106,
- ▶ *z/OS DFSMSHsm Storage Administration Guide*, SC35-0421

How to get IBM Redbooks

You can search for, view, or download Redbooks, Redpapers, Hints and Tips, draft publications and Additional materials, as well as order hardcopy Redbooks or CD-ROMs, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

Numerics

1750-EX1 258
300 GB DDMs 258
73 GB DDMs 258
9037 Sysplex Timer 13

A

ABARS 433–434
 and Mainstar 434
 manager 434
Accidental file deletion 423
adaptive sub-file backup 402
Advanced Copy Services 267
Advanced Technical Support see ATS
Aggregate Backup and Recovery Support see ABARS
AIX Micro-partitions and virtualization 91
application agent 356
Application Availability Analysis tool 80
Application Programming Interface (API) 131
application specific integration layer 222
ASP 152
asset management 392
asymmetric virtualization 361
asynchronous 139, 373
Asynchronous Peer-to-Peer Remote Copy 139
Asynchronous PPRC 269, 432
asynchronous remote 132, 372
asynchronous remote copy 132, 144, 376
Asynchronous Transfer Mode (ATM) 133, 136, 245
asynchronously 139
ATS 490
autoloaders 457
Automated Selection and Audit Process (ASAP) 434
Automated System Recovery (ASR) 427, 429
automatic resynchronization 312, 318
automation
 tape 456
auxiliary 140, 143
auxiliary VDisk 133, 143, 147, 372

B

background copy 146–147
backup
 tape strategies 453
backup and recovery software
 tape 179
backup methodologies
 tape 176
 tape differential backups 177
 tape full backups 177
 tape incremental backups 177
 tape off-site backup storage 178
 tape progressive backup 178

backup server 199, 416
backup sets 402
backup to both 198
backup to local 198
backup to Tivoli Storage Manager 198
Backup/Restore 4, 6–7
backups
 LAN-free 426
backup-while-open (BWO) 437
bandwidth 140, 322
Bare Machine Recovery (BMR) 393, 455
 central focus 427
base logical drive 299
 multiple FlashCopies 299
BIA 218
bitmaps 369
BladeCenter server 352
block aggregation 363
booting from the SAN 251
 serious considerations 252
boss node 365
BRCS 493
Business Continuity 1, 3, 108, 224
 iSeries 158
 planning for heterogeneous environment 18, 38, 59, 180, 501
 tiers 4
business requirement
 PPRC Migration Manager 170
Business Resiliency and Continuity Services see BRCS

C

cable
 type 316
cache 133, 372
Cache Fast Write 257
Calibrated Vectored Cooling technology 258
Capacity BackUp (CBU) 19
Capacity Backup (CBU) 28
capacity management 392
CartridgeGroup 385
CartridgeGroupApplication 385
CBU 28, 61
CEC 10
channel extender 133
Channel Extenders 323
CICS/VSAM Recovery see CICSVR
CICSVR 394, 437
CIFS 380
CIM 293
CIM see Common Information Model
CIM Agent 131, 363
CLO 21
ClusterProven program 494

- benefits 494
- Combining GLVM with HACMP/XD 90
- Common Endpoint 398
- Common User Access (CUA) 437
- Concurrent Copy (CC) 432
- configuration management 392
- connected 148
- consistency 144, 373–374
- Consistency Group 142, 146, 288, 503
 - commands 268
 - data consistency 288
 - what is it? 287
- Consistency Group (CG) 110–111, 114, 369, 501
- Consistency Group FlashCopy 275
- consistent 149–150
- Consistent Stopped state 147
- Consistent Synchronized state 147
- constrained link 140
- content management
 - application 209
- Continuous Availability for MaxDB and SAP liveCache
 - solution highlights 92
- Continuous Availability 1, 4, 10
- Continuous Availability for MaxDB and SAP liveCache 9
 - solution components 92
 - solution description 92
- Continuous Availability for maxDB and SAP liveCache
 - hot standby-greater detail 93
- Continuous Availability for open systems 9
- Continuous Availability for zSeries 9
- continuous availability for zSeries data
 - GDPS/PPRC HyperSwap Manager 109
- controller
 - ownership 298
- controller firmware 301
- copy service 139
- Copy Services
 - function comparison 499
 - interfaces 291
 - interoperability with ESS 293
- Copy Services for System i 9
- copy-on-write 299
- copy-on-write data 299
- core technologies
 - functionality 220
 - product examples 220
- core technologies layer 220
- Coupled System Data Mover support 30
- Coupling Facility Links 13
- Cristie Bare Machine Recovery 427
- Cross-Site LVM 88
- Cross-site LVM mirroring 88
- CTN 13–14
- CWDM 322

D

- dark fiber 136, 242
- data
 - retention-managed 206–207
- data availability 5

- data block 299
- data consistency 110, 139, 288, 373
 - Consistency Group 288
 - requirement for 18
- Data Facility Systems Managed Storage see DFSMS
- data life cycle 207
- data mining 207
- Data ONTAP 326, 334, 380
- Data Protection for Domino
 - components 412
- Data Protection for Exchange
 - VSS fast restore 199
 - VSS Instant Restore 199
- Data Protection for IBM DB2 UDB 407
- Data Protection for Informix 410
- Data Protection for Lotus Domino 411
- Data Protection for Microsoft Exchange Server 412
- Data Protection for Microsoft SQL Server 408
- Data Protection for mySAP 415
 - scope 415
- Data Protection for Oracle 409
 - LAN-free data transfer 410
 - Oracle RMAN 409
 - using for backup 409
- data rendering 207
- Data Retention 206–208
 - Compliance 208
 - Policy 208
 - System Storage Archive Manager 208
- Data Set FlashCopy 268, 274
- data sharing 231
- data shredding 207
- data/device/media migration 393
- database log 142
- DDR 211
- delta logging 305, 307, 312
- dependent writes 113, 141–142
- DFSMS 393
 - family of products 431
- DFSMS Optimizer HSM Monitor/Tuner 438
- DFSMS/VM 211
- DFSMS/zOS Network File System (NFS) 434
- DFSMSdfp 56, 431
 - Copy Services 432
- DFSMSdss (Data Set Services) 435
- DFSMSHsm 394
 - Disaster Recovery using ABARS 434
- DFSMSHsm Fast Replication 433
- DFSMSrmm (Removable Media Manager) 435
- DFSMSStvs Transactional VSAM Services 436
- direct attached storage (DAS) 352
- Directory Service 313
- dirty bit 143
- Disaster Recovery
 - for your Tivoli Storage Manager Server 404
 - key aspects 240
 - tape applications 485
- disaster recovery 206, 316
- disconnected 148
- disk drive 208

- disk mirroring
 - comparison 501
 - function definitions 502
- disk subsystem 109, 112
- distance limitations 132, 371
- distance limitations of Fibre Channel inter-switch links 315
- DriveGroup 385
- DriveGroupApplication 385
- DS CLI 292
- DS Open API 293
- DS Storage Manager 259, 292
- DS300 351–352
 - data protection 355
 - Disaster recovery considerations 357
 - FlashCopy 351, 355
 - iSCSI 354
 - overview 352
- DS300 FlashCopy
 - Tier 4 352
- DS400 351–352
 - data protection 355
 - Disaster Recovery considerations 357
 - FlashCopy 351, 355
 - overview 352
- DS400 FlashCopy
 - Tier 4 352
- DS4000
 - Asynchronous Remote Mirroring 298
 - Consistency Group (CG) 298
 - Dynamic Mode Switching 303, 307
 - dynamic mode switching 298
 - hardware overview 296
 - Storage Manager 298
 - storage terminology 298
 - Suspend / Resume 298
- DS4000 Family 295, 357
- DS4000 series 215, 296
- DS6000 254–255
 - Advanced Copy Services 267
 - DS6800 controller enclosure 256
 - expansion enclosure 258
 - hardware overview 256
 - highlights 259
 - naming convention 255
 - storage capacity 258
 - supported servers environment 259
 - switched FC-AL subsystem 257
- DS6800
 - processors 257
- DS6800 controller enclosure 256
- DS8000 254, 260
 - Advanced Copy Services 267
 - console 263
 - FlashCopy 270
 - hardware overview 261
 - highlights 264
 - internal fabric 262
 - LPARs 264
 - naming convention 260

- POWER5 processor technology 262
- S-HMC 263
- storage capacity 263
- supported environment 264
- DS8100 262
- DS8300 262
- DWDM 322
- Dynamic Mode Switching 303, 307
- dynamic transaction back (DTB) 437

E

- EKM 482
- electronic vaulting 188, 452
 - over a SAN 404
 - Tivoli Storage Manager 189
 - with virtual volumes 405
- Enterprise Removable Media Manager see eRMM
- Enterprise Removable Media Manager see eRMM
- Enterprise Resource Planning 414
- Enterprise Tape Library (ETL) Expert 394
- eRCMF 6, 107, 112, 222
 - ESS Copy Services 116
 - overview 120, 125
 - solution description 116, 120
 - solution highlights 126
- eRMM 382
- ESCON 293
- ESS 200, 254, 265
 - Advanced Copy Services 267
 - hardware overview 265
 - host adapters 266
 - naming conventions 265
 - processor technology 265
 - storage capacity 266
 - supported servers environment 267
- ETL Expert
 - performance management 444
- event/alert management 393
- events 147
- Exchange
 - integrity check 198
 - offloaded backup 198
 - VSS fast restore 199
 - VSS Instant Restore 199
- Exchange Server 412
- EXN1000 334
- EXP100 296
- EXP100 unit 208
- EXP700 296, 309
- EXP710 296, 309
- extended long busy (ELB) 114
- Extended Remote Copy 432

F

- fabric configuration 313
- failover 133, 372
- failover situation 132, 371
- Failover/Failback 503
- Fast Reverse Restore 60

- FC-AL 257
- FCP 380
- feature key 300
- Fibre Channel (FC) 133, 136, 227
- Fibre Channel disk drives 263
- Fibre Channel inter-switch link 315
- Fibre Channel switch
 - zoning 230
- FICON 227
- File corruption 423
- FlashCopy 268–270, 310, 500
 - backup/restore for SAP 7
 - benefits 272
 - Consistency Group 275
 - data set 274
 - DS300 351, 355
 - DS400 351, 355
 - establish on existing RMC primary 275–276
 - inband commands 277
 - Incremental 272
 - iSeries 158
 - mapping 129, 370
 - Multiple Relationship 274
 - no background copy 272
 - options 272
 - persistent 276
 - Refresh Target Volume 272
 - SAN Volume Controller 369
- FlashCopy features 370, 372, 374
- FlashCopy logical drive 299–300
- FlashCopy Management Command Line Tool 355
- FlashCopy Manager 107, 166
 - components 168
 - highlights 167
 - overview 167
 - solution description 167
 - use in DR 168
- FlashCopy repository 299
- FlashCopy repository logical drive 299–300
- FlexVol 397
- freeze 115, 503
 - command 115–116
- Freeze and Go 39
- Freeze and Stop 39
- FREEZE function 18
- full synchronization 307, 309, 318

G

- GDOC 67
 - consulting and planning 68, 70
 - details 69
 - implementation and deployment 69
 - overview 67
 - solution components 68
 - solution description 68
 - solution highlights 68
 - Tier 7 68
 - VERITAS software components 69
- GDPS 9, 15, 222
 - automation of zSeries Capacity Backup 61
 - control of Metro Mirror data consistency 37
 - exploitation of zSeries Capacity Backup 38
 - extended distance support between sites 54
 - FlashCopy support 64
 - how support of PtP VTS works 63
 - IBM Global Services 17, 36
 - integration 37
 - management of zSeries operating systems 23
 - multi-site workload 49
 - Open LUN management 46
 - prerequisites 65
 - PtP VTS support 62
 - single site workload 49
 - solution offerings defined 15
 - solution offerings introduction 14
 - summary 66
 - support of data consistency for tape and disk storage 37
 - systems 19
 - Technical Consulting Workshop 36, 491
 - three site solutions overview 34
 - value of GDPS automation 37
- GDPS HyperSwap Manager 6, 26, 172, 222
- GDPS z/OS Metro/Global Mirror 34
- GDPS/GDOC 69
- GDPS/Geographically Dispersed Open Clusters (GDOC) 67
- GDPS/Global Mirror 10, 15
- GDPS/GM 31, 37
- GDPS/Metro Global Mirror 35
- GDPS/Metro Global Mirror implementation uses a cascade of data which passes from Metro Mirror to Global Mirror 35
- GDPS/PPRC 10, 15–16, 111
 - attributes 19
 - FlashCopy support 36
 - HyperSwap 21
 - IBM Global Services 37
 - introduction 17
 - management for open systems LUNs 24
 - Multi-Platform Resiliency for zSeries 47
 - multi-platform resiliency for zSeries 24, 47
 - multi-site workload 52
 - overview 19
 - planned reconfiguration support 22
 - remote site outside the Parallel Sysplex 54
 - single site workload examples 49
 - summary 24
 - support for heterogeneous environments 23
 - support of Consistency Group FREEZE 38
 - Tier 7 Near Continuous Availability 25
 - topology 21
 - unplanned reconfiguration support 23
- GDPS/PPRC HyperSwap 21
- GDPS/PPRC HyperSwap Manager 10, 15, 37, 108–109
 - client requirements provided 27
 - description 109, 117
 - IBM Global Services 37
 - near continuous availability of data and D/R solution at metro distances 26

- near continuous availability of data within a single site 26
- overview 25
- Resiliency Family positioning 108, 117, 119
- solution highlights 27
- summary 27
- GDPS/PPRC multi-platform resiliency for System z 24
- GDPS/PPRC Multi-Platform Resiliency for zSeries 24
- GDPS/XRC 10, 16, 56, 58
 - coupled XRC System Data Mover support 30
 - details 56, 58
 - FlashCopy support 36
 - IBM Global Services 37
 - introduction 17
 - overview 28, 31
 - planned and unplanned reconfiguration support 29, 33
 - solution highlights 30, 33
 - summary 30, 33
 - topology 28, 31
- geographic networks
 - HACMP/XD
 - HAGEO 82
- Geographically Dispersed Parallel Sysplex 10
 - complete family 108
- Geographically Dispersed Parallel Sysplex (GDPS) 108, 222
- GeoRM 86
 - configurations 87
 - with HACMP 87
- Global 145
- Global Copy 268–269, 278, 286, 432, 503
 - PPRC Migration Manager 169
- Global Logical Volume Mirroring (GLVM) 89
- Global Mirror 15, 58, 114, 155, 220, 269, 278, 287, 432, 503
 - how works 280
- GLVM failover 90
- GM 139

H

- HACMP 71
 - Application Availability Analysis tool 80
 - application monitors 80
 - clients 79
 - elimination of single points of failure 79
 - Fallover and Fallback 73
 - GeoRM 87
 - High Availability Cluster Multi-Processing for AIX see HACMP
 - networks 79
 - nodes 78
 - physical cluster components 77
 - resource groups that include PPRC replicated resources 72
 - role 76
 - shared external disk devices 78
 - sites 73, 79
- HACMP/XD 9, 71–72, 81, 222
 - HAGEO

- criteria for the selection of this solution 85
- geographic mirrors 82
- geographic networks 82
- software components 84
- HAGEO cluster components 81
- HAGEO software 81
 - solution description 71
- HACMP/XD management of Metro Mirror pairs 74
- hardware infrastructure layer 219
- heterogeneous environment
 - planning for Business Continuity 18, 38, 59, 180, 501
- Hierarchical Storage Management (HSM) 417
- Hierarchical Storage Management see HSM
- High Speed Link (HSL) 156
- HMC 24
- host adapter
 - hardware configurations 392
- host adapters 263
- Host Bus Adapter (HBA) 251–252
- HSM FastAudit 441
- HSM FastAudit-MediaControls 441

I

- I/O activity 302
- I/O group 128, 365
- I/O request 226
- IASP 152
 - FlashCopy 154
 - Metro Mirror 156, 162
- IBM DB2 Content Manager 208
- IBM Geographic Remote Mirror for AIX 86
- IBM Geographic Remote Mirror for AIX see GeoRM
- IBM Global Services 218, 491
 - assessment 492
 - GDPS 17, 36, 222
 - GDPS Technical Consulting Workshop 36
 - GDPS/PPRC 37
 - GDPS/PPRC HyperSwap Manager 37
 - GDPS/XRC 37
 - RCMF 37
 - solutions integration 491
- IBM Open Enterprise System Virtualization 381
- IBM POWER5 262
- IBM Ramac Virtual Array (RVA) 360–361
- IBM System Storage DR550 206
 - main focus 209
- IBM System Storage DS4000 series 398
- IBM System Storage Enterprise Tape 460
- IBM System Storage Proven Solutions 491
- IBM System Storage Resiliency Portfolio
 - application specific integration layer 222
 - core technologies layer 220
 - hardware infrastructure layer 219
 - platform-specific integration layer 221
 - solution segments 7
 - strategic value 224
- IBM System Storage Solution Centers 490
- IBM TotalStorage DFSMSshm Monitor 438
- IBM TotalStorage Expert 443
- IBM TotalStorage Peer-to-Peer VTS Specialist 445

- IBM TotalStorage Productivity Center
 - overview 394
- IBM TotalStorage Productivity Center for Data 394
 - file sweep 396
 - NAS support 397
 - policy-based management 396
 - SAN Volume Controller reporting 396
 - subsystem reporting 396
 - V2.1 features 397
- IBM TotalStorage Productivity Center for Disk 394, 398
 - device management 399
 - performance monitoring and management 399
- IBM TotalStorage Productivity Center for Fabric 394
 - Common Endpoint 398
 - overview 397
 - V2.1 features 398
- IBM TotalStorage Productivity Center for Replication 394, 399
- IBM TotalStorage Tape Library Specialist 444
- IBM Virtual Tape Server (VTS) 360
- ICF/CF 11
- identical data 140
- image backups 402
- in-band 361
- inband
 - commands 268
- inband commands 277
- inconsistent 149
- Inconsistent Copying state 147
- Inconsistent Stopped state 147
- incremental export 406
- Incremental FlashCopy 36, 268, 272
- independent auxiliary storage pool (IASP) 152
- independent auxiliary storage pool see IASP
- Information Lifecycle Management (ILM) 207
- Informix 410
- Input/Output Adapter (IOA) 152
- Input/Output Processor (IOP) 152
- Integrated Catalog Facility (ICF) 441
- integrity 142
- integrity check 198
- intercluster 365
- Intercluster Metro Mirror 140, 371, 374
- inter-switch link 315
- intracluster 365
- Intracluster Metro Mirror 139, 371, 373
- IP storage
 - protocol summary 354
- iSCSI 228, 250, 353–354
 - gateway 228
 - initiators 250
- iSeries
 - Business Continuity 158
 - cross site mirroring (XSM) 165
 - FlashCopy 153, 158
 - Global Mirror 155
 - introduction 151
 - iSeries Copy Services toolkit 154
 - Metro Mirror 155, 162
- ISL hop count 140

K

- Key-Sequence Data Set (KSDS) 440
- K-System 18

L

- LAN-free backup 231, 242, 426
- latency 322
- LBA 143
- LIC 256
- Linear Tape Open see LTO
- link failure 318
- Load Source Unit (LSU) 154
- local disk for protection 422
- local mirroring
 - N series
 - local mirroring 345
- log 142
- Logical Block Address 143
- logical drive
 - primary 300
- logs 142
- long distance 322
 - considerations 245
- low tolerance to outage 4
- LPAR 264
- LPARs 262
- LTO
 - Accelis 459
 - Ultrium 459
- LUN 111
 - masking 230
- LVM
 - Single-Site Data Availability 88
- LVM mirroring
 - cross site
 - criteria for the selection 89
 - cross-site 88
- LVSA 402

M

- Mainstar
 - ABARS Manager 442
 - All/Star 443
 - disaster recovery utilities 441
- Mainstar ASAP 442
- Mainstar Backup and Recovery Manager 442
- Mainstar Catalog BaseLine 442
- Mainstar Catalog RecoveryPlus 440
- Mainstar FastAudit/390 441
- Mainstar Mirroring Solutions/Volume Conflict Rename (MS/VCR) 439
- Mainstar Software suite 393
- managed disk group (MDG) 366
- manual resynchronization 318
- mapping 369
- master 140, 143
- master VDisk 143, 147
- mDisk 365
- Media Manager 386

- Metro Cluster for N series 9
- Metro Mirror 15–16, 109, 136, 220, 268–269, 277, 286, 365, 371, 373, 375, 394, 502
 - distance limitations 136
 - IASP 162
 - intercluster 135
 - intracluster 135
 - iSeries 162
 - PPRC Migration Manager 169
 - SAN Volume Controller 133, 136
- Metro Mirror features 144
- Metro Mirror process 144
- Metro Mirror relationship 145
- Metro/Global Copy 269
- Micro-partitions 91
- Microsoft Exchange
 - N series backup 344
 - Single Mailbox Recovery 344
- Microsoft SQL Server
 - N series backup 344
- Microsoft Volume Shadow Copy Service 211
- mining 207
- mirror relationship 310
 - recreating 317
- mirror repository 298, 307, 319
- mirrored 133, 372
- mirrored copy 139
- mirroring 345
- MSCS 344
- Multiple Relationship FlashCopy 268, 274
- MWC 83

N

- N series 326
 - Data ONTAP 326, 334, 380
 - Exchange single mailbox recovery 344
 - EXN1000 334
 - expansion unit 334
 - FCP 380
 - mirroring 345
 - MS Exchange 344
 - MS SQL Server 344
 - N3700 327
 - operating system 334, 380
 - rapid backup and restore 344
 - SnapRestore 337
 - SnapShot 337
 - synchronous local mirroring 345
 - SyncMirror 345
 - WORM 326
- Name Service 313
- NAS 225, 228, 245
 - advantages 246
 - building your own 248
 - connectivity 246
 - enhanced backup 247
 - enhanced choice. 247
 - exploitation of existing infrastructure 246
 - gateway 228
 - heterogeneous file sharing 247

- improved manageability 247
- opposing factors 247
- overview 245
- resource pooling 246
- scalability 247
- simplified implementation 246
- storage characteristics 247
- NAS appliance 228
- NAS Gateway 249
- NASD 208
- Network Attached Storage see NAS
- network file system 422
- Network File System (NFS) 228
- networking terminology
 - tutorial 226
- NFS 380
- NOCOPY2COPY 36
- non-persistent shadow copy 198
- NVS 257

O

- offloaded backup 196, 198
- Open LUN management 46, 54, 111
 - GDPS/PPRC 111
- Oracle 409
 - RMAN 409
- out-of-band 361
- overwritten 369
- ownership 310

P

- Parallel Sysplex 10, 12
 - availability 111
- Partial Response Maximum Likelihood (PRML) 459
- Peer-to-Peer VTS 444
 - automatic copy function 455
- performance management 393
- Persistent FlashCopy 276
- Persistent Reservations 314
- persistent shadow copy 198
- planned outage 125
- platform-specific integration layer 221
 - AIX and pSeries 222
 - heterogeneous open systems servers 222
 - zSeries 222
- point-in-time 300, 302
- point-in-time copy 150
- point-in-time copy (PTC) 115, 128
 - comparison 500
- point-in-time copy see PTC
- point-in-time image 299
- policy decision 143
- PowerPC 257
- PPRC 16
 - commands 151
 - configuration limits 151
 - relationship 145
 - SAN Volume Controller 133
- PPRC Migration Manager 107, 166

- description 169
- diagnostic tools 170
- Global Copy 169
- load library 172
- modes of operation 170
- overview 167–168
- positioning with GDPS 172
- prerequisites 171
- Tier 6 169
- PPRC Migration Mirror
 - Metro Mirror 169
- PPRC-XD 268
- primary 133, 300, 372
- primary copy 143
- primary logical drive 298, 308
- primary site 110, 298
- production VDisk 143
- progressive backup 401
- pSeries 6
- PTAM 185
- PTC 270, 272
- PtP VTS 394
 - GDPS support 62–63

Q

- quorum disks 128

R

- R -System 18
- RAID-10 258, 264
- RAID-5 258, 264
- rapid backup and restore 344
- Rapid Data Recovery 1, 4, 6, 10, 107
- Rapid Data Recovery for UNIX and Windows
 - eRCMF 107
 - SAN Volume Controller 107
- Rapid Data Recovery for zSeries
 - product components 112
- RCMF
 - IBM Global Services 37
- RCMF/PPRC 16
 - highlights 28
- RCMF/XRC 16
 - overview 31
 - solution highlights 31
- real-time synchronized 132, 371
- recall 419
 - advanced transparent 420
- Records Manager 208
- Recovery Point Objective see RPO
- Redbooks Web site 506
 - Contact us xvi
- relationship 139–140, 373
- relationship state diagram 147
- remote machine protection 422
- Remote Mirror and Copy 269
- Remote Mirror and Copy see RMC
- remote physical volume 89
- remote site 110

- Remote Volume Mirror 302
- removable media management 393
- rendering 207
- repository logical drive 308
- Resiliency Family 215
- Resume Mirror 307
- resynchronization 312
- retention period 207
- retention policy 207
- retention-managed data 206–207
- Reversible FlashCopy 60
- RMC 277
 - comparison of functions 286
 - Global Copy 278
 - Global Mirror 278
 - Metro Mirror 277
- RMF/PPRC
 - overview 27
- role reversal 317–318
- Rolling Disaster 112
- RPO 131, 176, 287
- RPV client device driver 90
- RPV Server Kernel Extension 90
- R-System 59
- RTO 156, 176
- RVM
 - distance limitations 315
 - recreating a mirror relationship 317
 - switch zoning 313

S

- SAN 227–229
 - benefits 230
 - access 230
 - consolidation 230
 - booting from 251
 - design consideration 242
 - distance 241
 - dual 243
 - intermediate distances 244
 - long distances 244
 - recoverability 241
 - redundancy 240
- SAN File System 182
- SAN overview 229
- SAN portfolio
 - enterprise level 238
 - entry level 235
 - midrange 235
- SAN Volume Controller 107, 119, 127, 359, 371, 373, 375
 - basic benefits of copy functions 369
 - benefits 367
 - configuration node 128
 - Consistency Group (CG) 369
 - copy functions 369
 - Disaster Recovery considerations 375–376
 - FlashCopy 130–131, 369
 - FlashCopy services 128
 - Metro Mirror 133, 136, 138–139

- overview 366
- PPRC 133
- quorum disks 128
- UNIX 130, 136
- Windows 131, 138–139
- SAN Volume Controller (SVC) 399
- SAN Volume Controller FlashCopy
 - advantages 132
 - disadvantages 132
 - fundamental principles 128
 - practical uses 129
 - Tier 4 128
- SAN Volume Controller Metro Mirror
 - advantages 139
 - Consistency Group 114
 - disadvantages 139
- SAN zoning 313
- SATA 206, 208, 296
- ScratchPools 385
- scripting 143
- SDM 56
- SEC 208
- secondary 133, 372
- secondary copy 143
- secondary logical drive 308
- Secondary Site 298
- Serial Advanced Technology Attachment (SATA) 208
- Serial Advanced Technology Attachment see SATA
- Serial Storage Architecture (SSA) 265–266
- Server Time Protocol 13
- Server Time Protocol (STP) 54
- Server Timer Protocol 11, 21
- Server-free backup 242
- Server-less backup 242
- server-to-server hot-standby using incremental export 406
- Services and Skills 218
- shared data, shared access 10
- Shared processor pool 91
- S-HMC 291
- shredding 207
- Simple Network Management Protocol 143
- Simple Network Management Protocol (SNMP) 135, 398
- Single Mailbox Recovery 344
- single point 111
- small computer system interface (SCSI) 266
- Small Formfactor Plugables (SFP) 258
- SMI see Storage Management Initiative
- SnapRestore 337
- SnapShot 337
- snapshot
 - non-persistent 198
- SNIA
 - mission 362
 - shared storage model 362
 - SMI
- SNMP 143
- Solution Assurance 496
 - support links 497
- somewhat tolerant to outage 4
- source 143
- source logical drive 300–301
- source virtual disks 369
- SQL Server 408
- SRC 80
- SSA attached disk 266
- STANDARD LOGICAL Drive 299
- Standby Capacity on Demand (Standby CoD) 264
- state
 - connected 148
 - consistent 149–150
 - disconnected 148
 - inconsistent 149
 - overview 147
 - synchronized 150
- state fragments 149
- State overview 148
- states 147
- storage area network (SAN) 127, 227, 397
- storage area network see SAN
- storage device
 - centralized management 399
- Storage Management
 - key areas 392
 - software 215, 299
 - software solutions 393
- Storage Management Initiative 363
- Storage Management Subsystem (SMS) 431
- Storage Manager 298
- storage networking
 - benefits 226
 - connectivity 226
 - I/O protocol 226
 - main aspects 225
 - media 226
 - overview 226
- Storage Networking Industry Association (SNIA) 359, 398
- storage subsystem 301
 - standard logical drive 299
- storage virtualization 127
 - fabric level 361
 - levels 360
 - multiple level 364
 - overview 360
 - server level 361
 - storage subsystem level 360
 - volume managers 381
- STP 13–14
- Suspend Mirror 307
- SVC 365, 377
 - boss node 365
 - FlashCopy 139
 - glossary 365
 - I/O group 365
 - intercluster 365
 - intracluster 365
 - UNIX 139
- SVC PPRC functions 144
- switch 314

- switch zoning 313
- switched FC-AL subsystem 257
- Switched Fibre Channel Arbitrated Loop 263
- symmetric virtualization 361
- symmetrical multiprocessors (SMP) 265
- synchronization 146, 319
- synchronization priority 320
- synchronized 150
- synchronized before create 140
- synchronizing 140
- synchronous local mirroring 345
- Synchronous PPRC 109, 268, 432
- SyncMirror 345
- Sysplex Timer 13
- System Backup and Recovery 430
- System Management Facility (SMF) 440
- System Storage Rapid Data Recovery for UNIX and Windows
 - eRCMF 112
 - overview 120, 125
 - solution description 117
 - solution highlights 126
- System z 10
- System z Parallel Sysplex 10

T

- T0 369
- tape
 - applications 454
 - autoloaders 457
 - automation 456
 - backup and recovery software 179
 - backup methodologies 176
 - backup strategies 453
 - cartridge usage considerations 459
 - differential backups 177
 - drive decision criteria 457
 - drive technology 459
 - drives 457
 - full backups 177
 - incremental backups 177
 - libraries 457
 - off-site backup storage 178
 - progressive backup 178
 - rotation methods 178
- Tape Volume Cache 466
- tape-based storage 208
- target 143
- target logical drive 300–301
- target virtual disks 369
- Tier 1 182, 211–213
- Tier 2 182, 211
- Tier 3 182, 211
- Tier 4 182
 - DS300 FlashCopy 352
 - DS400 FlashCopy 352
 - SAN Volume Controller FlashCopy 128
- Tier 5 182
- Tier 6 182
 - PPRC Migration Manager 169

- Tier 7
 - GDOC 68
- Tier level 182
- tiers 4
 - blending 4–5
- time-zero 370
- Time-Zero copy 369
- Tivoli Continuous Data Protection for Files
 - Comparison with other backup solutions 422
 - local disk for protection 422
 - network file system 422
 - why is it required? 423
- Tivoli NetView 112, 118
- Tivoli Storage Manager 182, 393, 400
 - adaptive sub-file backup 402
 - backup 185
 - backup methods 401
 - backup sets 402
 - client 186
 - clustering 182, 192
 - Data Protection for mySAP 415
 - Disaster Recovery for the server 404
 - DRM 186
 - DRM process 188
 - electronic vaulting 182, 211
 - family 182
 - image backups 402
 - interface to Automated System Recovery 429
 - manual off-site vaulting 182, 211–213
 - off-site manual vaulting 187
 - progressive backup 401
 - reclamation 177, 463, 465, 472, 479
 - server 182
 - server-to-server hot-standby using incremental export 406
 - solutions 184
 - solutions overview 182
 - support for Automated System Recovery 429
- Tivoli Storage Manager (TSM) 182
- Tivoli Storage Manager Disaster Recovery Manager 403
- Tivoli Storage Manager for Application Servers 413
- Tivoli Storage Manager for Copy Services
 - backup to both 198
 - backup to local 198
 - integrity check 198
 - offloaded backup 198
- Tivoli Storage Manager for Data Retention 424
 - event-based management 425
 - functionality 424
- Tivoli Storage Manager for Databases 407, 454
 - Data Protection for IBM DB2 UDB 407
 - Data Protection for Microsoft SQL Server 408
 - Data Protection for Oracle 409
- Tivoli Storage Manager for Enterprise Resource Planning mySAP 414
- Tivoli Storage Manager for Mail 195, 411
 - Data Protection for Microsoft Exchange Server 412
- Tivoli Storage Manager for Space Management 417
 - advanced transparent recall 420
 - archive and retrieve 420

- automatic migration 418
- backup and restore 420
- platform support 420
- pre-migration 419
- recall 419
- selective 419
- selective migration 419
- threshold 418
- transparent 419
- Tivoli Storage Manager for Storage Area Networks
 - LAN-free backup and restore 426
- Tivoli Storage Manager for System Backup and Recovery 430
- Tivoli Storage Optimizer for z/OS 438
- Tivoli System Automation 112
- TotalStorage Copy Services
 - function comparison 499
- TotalStorage Management toolkit 436
- TotalStorage Productivity Center for Fabric 137
- TotalStorage Rapid Data Recovery 108, 117
- TotalStorage Rapid Data Recovery for UNIX and Windows
 - overview 125
- TS7700 Grid architecture 15
- TS7740 Grid Support 35
- TVC 466

U

- Ultrium 459
- unforeseen disasters 423
- Unplanned HyperSwap 23
- unplanned outage 125
- Unwanted file alteration 423
- use of Metro Mirror 143

V

- vDisk 128, 366
- VDS 197
- VERITAS 69
- VERITAS Cluster Server 68
- VERITAS CommandCentral Availability 70
- very tolerant to outage 4
- VIO 91
- virtual disks 128
- Virtual I/O Server 364
- Virtual IO 91
- volume managers 381
 - advantages 381
 - disadvantages 381
- VolumeCopy 300, 310
- VSS 195, 197, 211, 344, 412
 - backup to both 198
 - backup to local 198
 - backup to Tivoli Storage Manager 198
 - integrity check 198
 - offloaded backup 196, 198
- VSS fast restore 199
- VSS Instant Restore 199
- VTs 443

- health monitor 444

W

- WAFL 337
- WEBEM (Web Based Enterprise Management) 363
- Wide Area Network (WAN) 228
- WORM 326
- write caching 304–306
- Write Consistency Group 319
- Write Once Read Many (WORM) 208
- Write ordering 149
- write ordering 141
- writes 142
- WWPN 313

X

- XRC 16

Z

- z/OS Global Mirror 15–16, 269, 281, 287, 432
 - review 56
- z/OS Global Mirror(ZGM) 28
- z/OS Metro/Global Mirror 282–284
- z/VM 211
- Zero data loss 20
- zone 313
- zoning 230
- zSeries 6, 107

Archived



IBM System Storage Business Continuity: Part 2 Solutions Guide

(1.0" spine)

0.875" x 1.498"

460 <-> 788 pages



IBM System Storage Business Continuity: Part 2 Solutions Guide



**Apply Business
Continuity principles**

**Learn about IBM
System Storage
products for Business
Continuity**

**Design resilient
storage solutions**

This IBM Redbook is a companion to *IBM System Storage Business Continuity Guide: Part 1 Planning Guide*, SG24-6547. We assume that the reader of this book has understood the concepts of Business Continuity planning described in that book.

In this book we explore IBM System Storage solutions for Business Continuity, within the three segments of Continuous Availability, Rapid Recovery, and Backup and Restore. We position these solutions within the Business Continuity tiers.

We describe, in general, the solutions available in each segment, then present some more detail on many of the products. In each case, we point the reader to sources of more information.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-6548-00

ISBN 0738489727