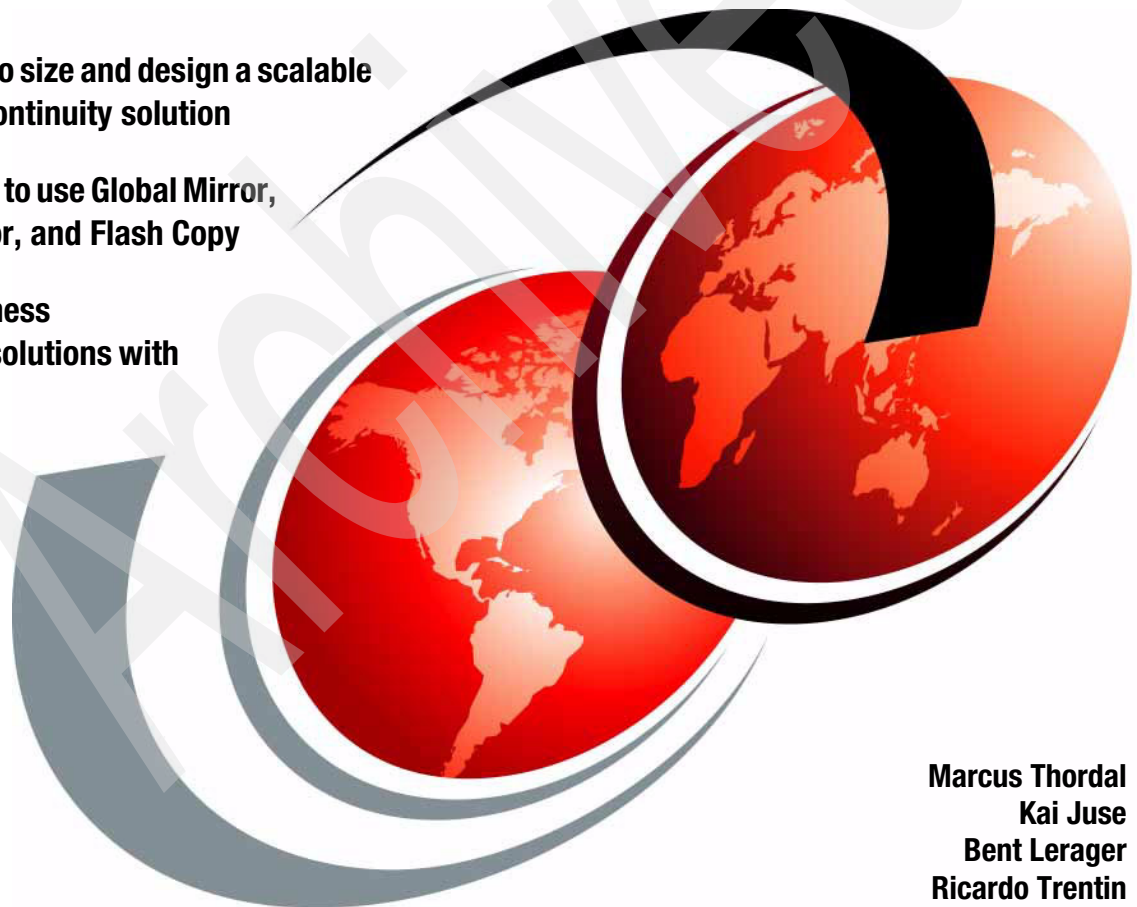


Using the SVC for Business Continuity

Learn how to size and design a scalable Business Continuity solution

Learn when to use Global Mirror, Metro Mirror, and Flash Copy

Learn Business Continuity solutions with



Marcus Thordal
Kai Juse
Bent Lerager
Ricardo Trentin



International Technical Support Organization

Using the SVC for Business Continuity

September 2007

Archived

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

First Edition (September 2007)

This edition applies to Version 4.2.0 of the IBM System Storage SAN Volume Controller.

© 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	ix
Trademarks	x
 Preface	 xi
The team that wrote this IBM Redbooks Publication	xi
Become a published author	xiii
Comments welcome	xiv
 Chapter 1. Introducing Business Continuity	 1
1.1 Business Continuity	2
1.2 IT Business Continuity	2
1.2.1 Continuous Operation	4
1.2.2 High availability	4
1.2.3 Disaster recovery	5
1.2.4 Relationship between high availability and disaster recovery	5
1.3 Recovery objectives	6
1.3.1 Recovery Point Objective	6
1.3.2 Recovery Time Objective	6
1.3.3 Hardware data integrity versus transaction integrity	6
1.4 The seven tiers of disaster recovery	9
1.4.1 Tier 0: No off-site data	10
1.4.2 Tier 1: Data backup with no hot site	11
1.4.3 Tier 2: Data backup with a hot site	11
1.4.4 Tier 3: Electronic vaulting	11
1.4.5 Tier 4: Point-in-time copies	11
1.4.6 Tier 5: Transaction integrity	12
1.4.7 Tier 6: Zero or little data loss	12
1.4.8 Tier 7: Highly automated, business integrated solution	12
1.5 Selecting and building a recovery solution	13
1.5.1 From business process to recovery objectives	13
1.5.2 Matching recovery objectives with Business Continuity technologies	14
1.5.3 SVC features for Continuous Operation	14
1.5.4 SVC features for high availability	15
1.5.5 Disaster recovery designs with SVC	16
 Chapter 2. Designing Business Continuity solutions with the SVC	 17
2.1 Design	18
2.2 Concepts	19
2.2.1 High availability	19

2.2.2	Continuous operation	19
2.2.3	Continuous availability	20
2.2.4	Business requirements	20
2.2.5	Application requirements	22
2.3	Factors to consider for BC with SVC	23
2.3.1	Host attachment	23
2.3.2	Storage array controllers	23
2.3.3	SAN extension	24
2.3.4	SVC Copy Services	24
2.3.5	Performance	26
2.3.6	Automation	27
2.4	SAN extension	27
2.4.1	SAN, MAN, and WAN	28
2.4.2	Replication topology	29
2.4.3	Technology options for SAN extension	29
2.5	Distance factors for intersite communication	34
2.5.1	Link speed, bandwidth, and latency	34
2.5.2	Link quality	35
2.5.3	Hops	36
2.5.4	Buffer credits	37
2.5.5	TCP window size	38
2.5.6	TCP selective acknowledgment	39
2.5.7	Path MTU and path MTU discovery	40
2.5.8	Compression	41
2.5.9	Write Acceleration	42
2.5.10	Encryption	44
2.5.11	Extended Fabrics and Routed Fabrics	44
2.6	Bandwidth sizing	46
2.6.1	Measuring workload I/O characteristics	47
2.6.2	Determining the bandwidth	47
2.6.3	Rule of thumb	48
Chapter 3. SAN Volume Controller Mirroring Solutions		51
3.1	Function	52
3.2	Concepts of operation	53
3.2.1	SVC objects	53
3.2.2	Summary of data consistency	55
3.2.3	Modalities of Remote Copy	57
3.2.4	Summary of relationships and Consistency Group states	59
3.2.5	Remote Copy implementation differences	61
3.3	Remote Copy internals	62
3.3.1	Intercluster communication	62
3.3.2	Remote Copy layer	70

3.3.3 Read stability and write ordering	75
3.3.4 Global Mirror time restrictions	77
3.4 Remote mirroring compatibility	79
3.4.1 Operating systems	79
3.4.2 HBAs	79
3.4.3 RAID controllers	80
3.4.4 SAN switches	80
3.4.5 Intersite communication	80
3.5 Remote mirroring limitations	80
3.5.1 Relationship per VDisk	81
3.5.2 Number of relationships	81
3.5.3 Number of Consistency Groups	81
3.5.4 Number of relationships per Consistency Group	81
3.5.5 Data mirroring capacity	81
3.5.6 Distance limitation	81
3.5.7 Background copy	82
3.6 Remote copy requirements	82
3.6.1 Licensing	82
3.6.2 Fabric requirements	83
3.6.3 Intersite communication and storage controller requirements	84
3.6.4 SVC configuration requirements	84
3.7 Scenario implementation	85
3.7.1 Fabric configuration	86
3.7.2 Intersite Communication configuration	86
3.7.3 Cluster partnership	86
3.7.4 Storage controllers LUN creation	87
3.7.5 MDisks creation	89
3.7.6 MDisks Group Creation	92
3.7.7 VDisks creation	95
3.7.8 Creating Consistency Groups	102
3.7.9 Relationship creation	105
3.7.10 Starting a relationship or Consistency Group	112
3.7.11 Stopping a relationship or Consistency Group	113
3.7.12 Switching copy direction	114
3.7.13 Monitoring background copy progress and state	115
3.8 Performance guidelines	118

Chapter 4. Performance considerations in SVC Business Continuity solutions	119
4.1 Considerations when using SVC in a Business Continuity solution	120
4.2 Introduction to TPC performance reporting for SVC	121
4.2.1 Adding SVC to TPC	122
4.2.2 Displaying SVC asset information	123

4.2.3 SVC performance statistics and presentation	126
4.3 Design considerations and planning	130
4.3.1 Key metrics	131
4.3.2 Design guidelines	132
4.3.3 TPC-based SVC performance reporting for Remote Copy sizing . .	134
4.3.4 Subsystem-based performance reporting before introducing copy services.	135
4.3.5 Host-based performance reporting before introducing copy services. .	136
4.4 Monitoring SVC copy service performance and identifying issues	136
4.4.1 The secondary subsystem of a Remote Copy	136
4.4.2 FlashCopy	138
4.4.3 FlashCopy of Remote Copy secondary VDisks	140
4.4.4 Identifying Global Mirror colliding write operations	142
4.4.5 Identifying Global Mirror overload	144
4.4.6 Using TPC alerts	147
Chapter 5. Automation in a Business Continuity solution with SVC . . .	155
5.1 How and why to automate	156
5.1.1 Business Continuity tiers	157
5.1.2 Considerations for Point in Time copy	159
5.1.3 Level of automation in Business Continuity environment	160
5.2 TotalStorage Productivity Center	161
5.2.1 TotalStorage Productivity Center for Data	161
5.2.2 TotalStorage Productivity Center for Disk	162
5.2.3 TotalStorage Productivity Center for Fabric	162
5.2.4 TotalStorage Productivity Center for Replication	163
5.3 Automatic Configuration, single point-of-management	163
5.3.1 Automatic expansion of volumes and file systems	164
5.4 TotalStorage Productivity Center for Replication	165
5.4.1 TPC for Replication	165
5.4.2 TPC for Replication Two Site Business Continuity	166
5.4.3 TPC for Replication installation	171
5.4.4 Overview of TPC-R	177
5.4.5 Adding the SVC to TPC for Replication	181
5.4.6 Define the SVC Metro Mirror session	184
5.4.7 Define the SVC FlashCopy session	190
5.4.8 Command Line Interface	193
5.5 Volume Shadow Copy Service	194
5.5.1 Considerations for basic SVC and VSS	195
5.6 Logical Volume Manager versus Global Mirror, Metro Mirror, and automation	195
5.7 Scripting	196

5.7.1 Scripting tools for SAN Volume Controller	197
Chapter 6. Recovery solutions	199
6.1 What is Disaster Recovery	200
6.2 Backup/restore	200
6.2.1 Tape or disk backup	201
6.3 Point-in-time copy	202
6.3.1 FlashCopy	203
6.4 Microsoft Volume Shadow Copy Service	205
6.4.1 How it works	206
6.4.2 Installing the IBM TotalStorage hardware provider	207
6.4.3 Verifying the VSS configuration	213
6.4.4 Using VSS with NTBackup	216
6.5 Metro Mirror and Global Mirror	221
6.6 LVM versus Metro Mirror, Global Mirror, and DR	222
Appendix A. TPC performance metrics for SVC	225
TPC metrics at SVC node, IOGroup, and cluster level	226
TPC metrics at the SVC VDisk level	227
TPC metrics at the SVC MDisk level	228
TPC metrics at the SVC port level	228
Relationship between TPC metrics and SVC counters	229
Related publications	233
IBM Redbooks	233
How to get IBM Redbooks	233
Help from IBM	233
Index	235

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®
DB2®
DS4000™
DS6000™
DS8000™
Enterprise Storage Server®
FlashCopy®
GDPS®

HyperSwap™
HACMP™
IBM®
Redbooks®
Redbooks (logo) ®
REXX™
System z™
System Storage™

Tivoli Enterprise™
Tivoli Enterprise Console®
Tivoli®
TotalStorage®
z/OS®

The following terms are trademarks of other companies:

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

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

Snapshot, and the Network Appliance logo are trademarks or registered trademarks of Network Appliance, Inc. in the U.S. and other countries.

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

Active Directory, Microsoft, SQL Server, Windows Server, Windows, and the Windows logo are trademarks of Microsoft Corporation 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® Redbooks® publication gives a broad understanding of how you can use the IBM System Storage™ SAN Volume Controller (SVC) as the foundation for IT Business Continuity in your enterprise.

This book will help you select the appropriate techniques and technology including how to size and design your IT Business Continuity Solution, while utilizing the SVC Advanced Copy Services and obtaining a highly flexible and scalable storage infrastructure.

The team that wrote this IBM Redbooks Publication

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

Marcus Thordal is a Senior IT Specialist in IBM Global Services, Denmark and has been with IBM since 1998. He is a BCFP, MCSE, and IBM Certified IT Specialist, and his areas of expertise include Open Systems storage solutions, in particular design and implementation of Storage Virtualization, Information Lifecycle Management and Disaster Recovery solutions. He holds a Bachelor of Science in Engineering from The Technical University of Denmark. He coauthored eight previous IBM Redbooks on IBM Total Storage products and solutions.

Kai Juse is a System Engineer working for Systemvertrieb Alexander (SVA) in Germany. His areas of expertise include storage area networks, disk storage systems, storage virtualization solutions, and storage software in Open Systems environments. He joined SVA in 2003 after he finished his Computer Science degree from Technische Universität Darmstadt, Germany.

Bent Lerager is a Certified IT Specialist with 15 years of experience with IBM products and is currently working in IBM Denmark. He is involved throughout the pre-sales process from the design of end-to-end, high-end storage solutions to competitive positioning and all the way through to implementation and problem determination. He is certified in high-end storage products, storage virtualization, and SAN.

Ricardo Trentin is an IT Specialist who works for IBM Global Technological Services in Brazil where he gives L3 support for Storage and Unix. He has more than five years of experience in implementing, designing, and supporting storage

solutions in Multi vendor Environments. He carries major UNIX® flavor certifications as HP Tru64, HP-UX, AIX®, and Solaris™ besides 10 years of experience in complex environments.



Figure 0-1 The team: Bent, Ricardo, Kai, and Marcus.

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

Tom Cady
Charlotte Brooks
Emma Jacobs
Mary Lovelace
Leslie Parham
Deanna Polm
Sangam Racherla
Jon Tate
Sokkieng Wang
International Technical Support Organization, San Jose Center

Dave Sinclair
Rob Nicholson
Carlos Fuente

Alex Howell
Barry Whyte
IBM Hursley

Werner Bauer
IBM Mainz

Brian F Sherman
IBM Markham

Dorothy Faurot
Jeffrey Larson
Lu Nguyen
John Power
Chris Saul
John Sing
IBM San Jose

Silviano Gaona
Brian Steffler
Brocade Communications Systems

John McKibben
Darshak Patel
Cisco Systems

Jeff Gatz
McDATA Corporation

Tom and Jenny Chang
Garden Inn Hotel, Los Gatos, California

Become a published author

Join us for a two-to-six week residency program! Help write an IBM Redbooks publication 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 customer satisfaction. As a bonus, you will develop a network of contacts in IBM development labs, and increase your productivity and marketability.

To obtain more about the residency program, browse the residency index, and apply online at the following Web site:

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 Redbooks form found at:

ibm.com/redbooks

- ▶ Send your comments in an email to:

redbooks@us.ibm.com

- ▶ Mail your comments to:

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



Introducing Business Continuity

This chapter introduces the evolving concept of *Business Continuity* and its related terms. We discuss the motivation and goals for employing Business Continuity designs and solutions in an enterprise environment and the role *IBM System Storage SAN Volume Controller* can play in those solutions.

1.1 Business Continuity

Today's highly competitive business marketplace allows enterprises little room for error when it comes to availability, continuous operation, and recovery from unplanned disruptions. Few events impact a company in the same way as a computer system outage does, even if it is only for a matter of minutes and then finding the incident reported in the media.

Today your customers, employees, and suppliers expect to be able to perform business with you around the clock and from all corners of the globe. Outages can damage your reputation and brand image.

We define Business Continuity as the ability to rapidly adapt or respond to any internal or external opportunity, demand, disruption, or threat, and continue business operation without significant impact. It is an enterprise-wide effort toward readiness to adapt and respond to risks, as well as opportunities, in order to maintain continuous business operations, to be a more trusted partner, and to enable growth. The scope of Business Continuity spans all levels of the enterprise and is not restricted to IT or the data center. Severe disruption can arise not just from the unavailability or corruption of business data, but also from disruption to other parts of the enterprise.

Business Continuity goes beyond disaster recovery. It is a management supported effort and is a process that relies on people, facilities, infrastructure, and applications with the goal to sustain operations at all times and under any circumstances. A *Business Continuity Plan* has a focus that is beyond the data center. It includes a crisis management plan, business impact analysis, human resource management, business recovery procedures, and documentation.

You can find a more detailed overview of Business Continuity and a more comprehensive variety of technologies outside the SVC scope of this IBM Redbooks publication in *IBM System Storage Business Continuity Solutions Overview*, SG24-6684 *IBM System Storage Business Continuity: Part 1 Planning Guide*, SG24-6547. These books on Business Continuity expand the scope to the thought process, methodology topics, and technology concepts, including IBM System z and Open System servers, as well as other IBM System Storage offerings.

1.2 IT Business Continuity

IT Business Continuity refers to the aspects of Business Continuity that relate to the overall IT infrastructure of an enterprise. This infrastructure includes a wide

range of elements, including software, hardware, and networks. Following are a few examples:

- ▶ Business applications
- ▶ Infrastructure applications, such as data backup software, monitoring solutions, and so on
- ▶ Servers
- ▶ Data communication and storage networks
- ▶ Storage systems

Because this IBM Redbooks publication considers Business Continuity in this IT-centric context, we now use the term Business Continuity to refer to IT Business Continuity.

We segment Business Continuity into the following:

- ▶ Continuous operations (CO)
- ▶ High availability (HA)
- ▶ Disaster recovery (DR)

Specific Business Continuity solutions can address all or some of these aspects of Business Continuity to a varying degree.

Figure 1-1 shows the relationship among the three Business Continuity aspects.

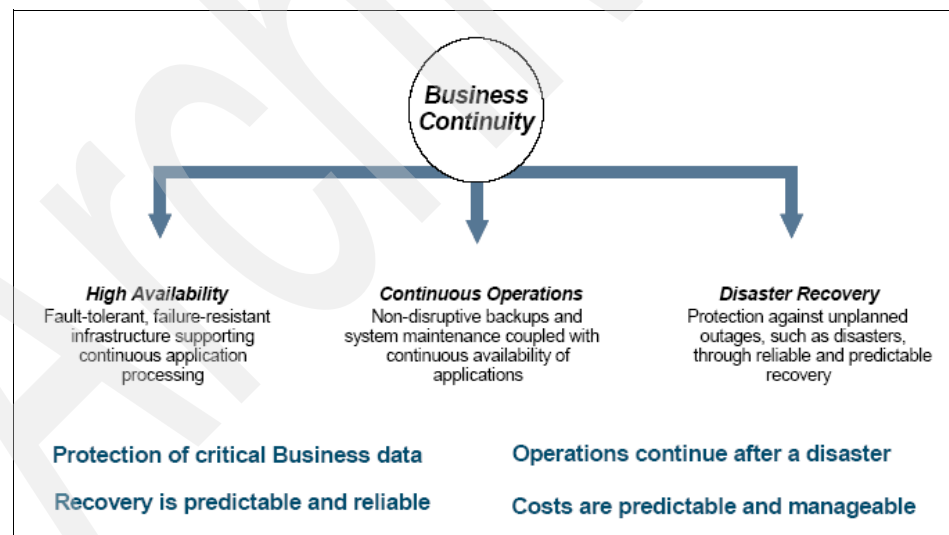


Figure 1-1 Three aspects of Business Continuity

1.2.1 Continuous Operation

The continuous operations aspect of Business Continuity is concerned with the ability of a Business Continuity solution to minimize outages during normal operation of the IT infrastructure. Normal operation is used here to contrast everyday activities such as back-ups, system maintenance, upgrades or reconfigurations, with failure situations that are addressed by other Business Continuity solutions.

IT systems that are tailored to the Continuous Operation concept allow an enterprise to operate and develop their IT infrastructure with minimal disruption. Some examples can be non disruptive dynamic reconfiguration of storage volume allocations and sizes, and concurrent software or hardware component upgrade ability. These features advance the Continuous Operation value of the systems themselves. In a comprehensive Business Continuity solution, you can also find another class of Continuous Operation features. These features and functions provided by one system are used to reduce disruptions in the operation of other systems. For example, a point-in-time copy functionality of a storage system can help reduce the operational impact of performing constant backups by greatly reducing the time of application suspension. It might also allow a more affordable creation of a volume copy for a testing environment in shorter cycles and with less disruption to the original systems.

Continuous Operation capabilities are a property of an individual system as well as a part of the Business Continuity solution as a whole.

1.2.2 High availability

The goal of using *highly available* systems is to obtain continuous availability of computing services even in failure situations. High availability is the property of a system designed so that users do not experience any impact due to an outage by employing hardware components, software, and operational procedures that will mask outages from the users.

High availability usually requires that the system recovers from an outage so quickly that the user does not perceive it as an outage. It also frequently requires the use of redundant components so that an alternate component can be used in case of a permanent component failure, or while a component is in maintenance.

Many major components of today's computer systems are fault tolerant to some degree, which means that they tolerate some faults through the use of the following:

- ▶ Redundant sub-components
- ▶ Error checking and correction for data
- ▶ Retry capabilities for basic operations

- ▶ Alternate paths for I/O requests
- ▶ Transparently mirrored data on separate storage devices

Fault tolerant systems are often designed so that they do not stop working if they have an internal component failure or if one of their devices suffers a failure.

A simple example of a redundant sub-component is dual power supplies. The failure of one of the redundant power supplies does not disrupt system operation and has no user impact on the computing services. Additionally, the nondisruptive repair of the failed power supply that is carried out later also has no impact on the user.

1.2.3 Disaster recovery

Disaster recovery (DR) is the process of reacting to a disaster by being able to provide computing services from another location. In most cases, the counter measures you employ to recover from a disaster are entirely different from the solution you use to achieve high availability. In a disaster situation, users are aware that an outage occurred at the central computer facility, and the duration of the outage is dependent on the recovery solution. Usually we measure this duration in two ways: the time until computing services are once again available to the user and the period of time prior to the disaster event for which data is lost. We discuss these aspects further in section 1.3, “Recovery objectives” on page 6.

1.2.4 Relationship between high availability and disaster recovery

All the components that make up high availability in a computer system are usually situated in the same building. Therefore, the building itself can represent a single point of failure. Despite the high availability that you designed into the system, a disaster compromising that building can cause you to lose significant parts of, or all of, your computing services.

When you are making preparations to react to a disaster, you can decide on the size of the recovery solution by performing a business impact analysis of your business. The solution you apply might be more of a recovery solution, where it is acceptable to that particular business that the user experiences the effects of the outage, rather than a high-availability solution that makes any outage transparent to the user. The costs and implementations may be very different in each case.

1.3 Recovery objectives

The goal of any disaster recovery solution planning is to protect the most business critical processes from outages, and minimize unplanned downtime. Keep in mind that all planning for any type of DR solution is always subject to balancing the solution's downtime goals with its costs. The following sections define terms and a structure for DR solutions to describe and quantify their characteristics.

In a disaster situation, users are normally aware that an outage occurred. Two of the dominating properties of a proposed DR solution are based on the user visible duration of the computing services impact. The properties are as follows:

- ▶ Recovery Point Objective
- ▶ Recovery Time Objective

1.3.1 Recovery Point Objective

The *Recovery Point Objective* (RPO) represents the extent of data loss you are willing to accept due to a disaster. It is measured as the duration of time prior to the disaster event for which you must re-execute your work (or accept its loss) after your system is recovered.

The RPO is a requirement on the currency of data available for recovery. For example, a business that believes it could acceptably afford to recreate or lose the data processed in the last five minutes preceding a disaster event has a RPO of five minutes. A business relying on daily backups stored off-site has to plan for a 24 hour RPO for a site disaster.

1.3.2 Recovery Time Objective

The *Recovery Time Objective* (RTO) is the duration of time following a disaster for which you are willing to accept the loss of computing services. The period starts from the moment of the disaster until the moment when systems are recovered. You can consider the RTO as a measure of how long a business can afford to have systems and applications down after a disaster. For example, a business that believes that it could afford to be without systems for eight hours has an RTO of eight hours.

1.3.3 Hardware data integrity versus transaction integrity

The ultimate objective of a recovery solution is the resumption of the most valued business processes that were disrupted by the disaster event. The IT recovery subset of this objective is concerned with the recovery of necessary computing

services for resumption of work at the application or user level. This means that the integrity of application level transactions needs to be restored. It is useful in discussing recovery solutions to distinguish this level of recovery (*transaction integrity*) from the level at which consistent data is recovered and data integrity is assured (*hardware data integrity*).

The difference between these levels becomes apparent when we consider the example of a transaction processing database system affected by a disaster. If we assume that all the data of the database system was synchronously mirrored to a recovery storage system, then we let a recovery server have access to the copied data. The database system that is started on the recovery server after the disaster event has full access to consistent data that was current at the moment of the disaster. Full hardware data integrity is achieved at that point.

However, the database system might now need to apply the information in the database log to ensure database transaction integrity. You may need to roll forward some work logged for committed transactions, and roll back some other work related to transactions that did not reach a committed state. At an application level the work that went into uncommitted database transactions is lost. Application or business transactions that depend on these activities need to be recovered.

The general IT recovery process following a disaster event is divided into two major phases as shown in Figure 1-2.

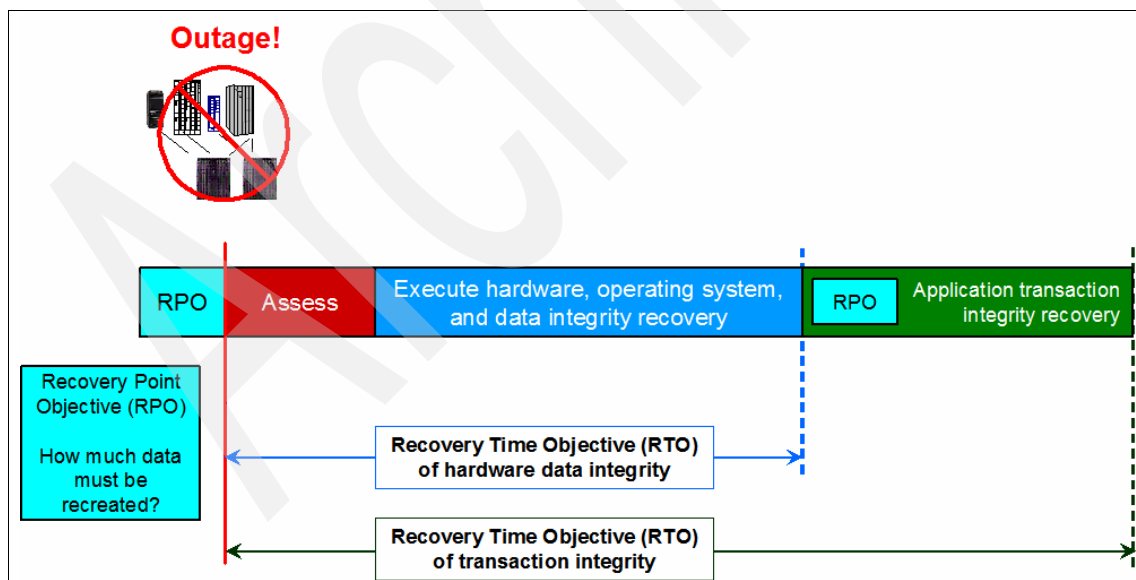


Figure 1-2 Timeline of an IT disaster recovery

The first phase begins with assessing the impact of the outage event and reaching a management decision as to whether you need to invoke the disaster recovery plan. This phase uses methods and technologies such as recovery standby systems deployed at a secondary site, off-site data backups, data replication, and remote copy or point-in-time copy functionality. The goal of this phase is to recreate a computing environment with consistent data no less current than the RPO, and in a period of time that is not longer than the RTO. The aim for the duration of this phase is the RTO for hardware data integrity.

The second phase builds on the recovered computing resource and consistent data. It deals with advancing the recovery to a state of application level transaction integrity. Applications and database staff will work to back out of or finalize incomplete transactions to bring the database and applications into a consistent state for the resumption of work at a user level. The recovery point that is selected as achievable could have a significant impact on the actual time needed for this process.

In some cases the recovery staff can reach transactional integrity by performing a database restart operation with little additional effort and minimum elapsed time. In a recovery scenario that is based on a longer RPO, it takes considerably more time and effort in this phase to bring the database up to a state of transactional integrity. The potential need for a database recovery operation (involving image restoration and applying transaction logs) makes it harder to predict a reliable upper time limit on the duration of this phase.

The target duration of the time from the disaster event to completing the second phase is the RTO for transactional integrity.

In Figure 1-2 on page 7, note the difference in elapsed time between the RTO of hardware data integrity and the RTO of transactional integrity. When discussing the RTO, it is important to understand which of these two distinct variations is being referred to. Operations and application staff can have differing perceptions of RTO depending on whether the RTO is assumed to be at the hardware recovery level, or at the application recovery level. The fact that there are different RTOs at progressive levels of recovery is essential to understanding and planning for a DR solution.

Finally, observe how RPO is depicted in Figure 1-2 on page 7, where RPO is the amount of data recreation required prior to the point-of-outage and is shown as the time offset before the outage occurred.

Note: RPO data recreation happens in the transactional integrity recovery phase. RPO data recreation cannot be done in the hardware and operating system recovery phase because the server and storage components do not have any knowledge of the logical relationships between multiple applications and the database blocks of data.

1.4 The seven tiers of disaster recovery

In 1992, the SHARE user group in the United States, in combination with IBM, defined a set of DR tier levels. This was done to address the need to properly describe and quantify various different methodologies for successful mission-critical computer systems' DR implementations.

Accordingly, within the IT Business Continuity industry, the tier concept continues to be used, and it is very useful for describing the Business Continuity capabilities of today's technologies. Depending on the context they are used in, they need to be adapted for today's specific business continuity technologies and associated RTO/RPO. The seven tiers of business continuity solutions offer a simple methodology of how to define your current service level, the current risk, the target service level, and the target environment.

Figure 1-3 on page 10 shows the seven tiers of Business Continuity as they relate to typical recovery technologies. It arranges the tiers along a trade-off curve between the potential costs of a recovery solution and the expected RTO offered by that solution. Solutions designed for a higher level and shorter RTO tend to have higher costs, but also often offer greater value to the business operations.

Note: The technologies mentioned are not the only options at their respective level, and some technologies might cover more than one level, depending on the implementation.

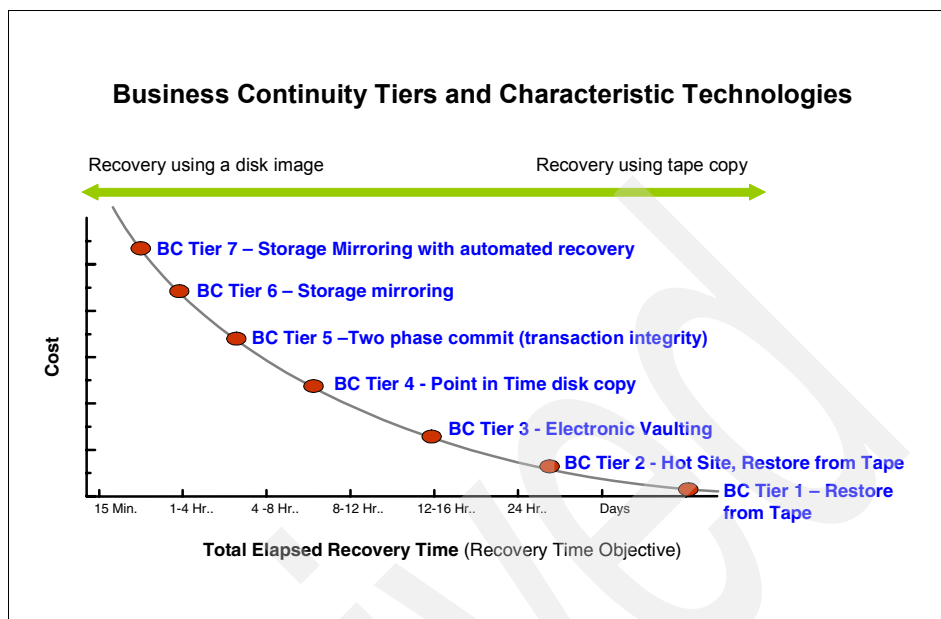


Figure 1-3 Seven tiers of Business Continuity with characteristic technologies

The following sections characterize the Business Continuity tiers by associating them with business characteristics with typical recovery technologies as discussed in *IBM System Storage Business Continuity: Part 1 Planning Guide*, SG24-6547.

1.4.1 Tier 0: No off-site data

A business with a Tier 0 business continuity solution has no Business Continuity plan.

- ▶ There is no saved information, no documentation, no back-up hardware, and no contingency plan.
- ▶ The length of recovery time in this instance is unpredictable. In fact, it might not recover at all.

This tier is defined for the purposes of differentiation and the evaluation of a business situation, rather than an adequate or desirable goal of a Business Continuity solution.

1.4.2 Tier 1: Data backup with no hot site

Business that uses Tier 1 Business Continuity solutions back up their data at an off-site facility. Depending on how often back-ups are made, they are prepared to accept several days to weeks of data loss, but their back-ups are securely stored off-site. However, this tier lacks the systems on which to restore data.

Examples for Tier 1 Business Continuity solutions/technologies are the Pickup Truck Access Method (PTAM), disk subsystem, or tape-based mirroring to locations without computing facilities, and using applications such as IBM Tivoli® Storage Manager.

1.4.3 Tier 2: Data backup with a hot site

A business that uses a Tier 2 business continuity solution performs regular back-ups to tape. This is combined with an off-site facility and infrastructure (known as a *hot site*) in which they can restore computing services from those tapes in the event of a disaster. This tier still results in the need to recreate several hours to days worth of data, but it is less unpredictable in recovery time.

Examples for Tier 2 Business Continuity solutions and technologies are PTAM with a hot-site available and using applications such as IBM Tivoli Storage Manager.

1.4.4 Tier 3: Electronic vaulting

Tier 3 solutions utilize components of Tier 2. Additionally, some mission critical data is electronically transferred to off-site storage. This electronically vaulted data is typically more current than data that is shipped through the PTAM. As a result there is less data recreation needed, or loss, after a disaster and thus allows an improved RPO.

Examples for Tier 3 Business Continuity solutions/technologies are electronic vaulting of data, and IBM Tivoli Storage Manager - Disaster Recovery Manager.

1.4.5 Tier 4: Point-in-time copies

Tier 4 solutions are used by businesses who require both greater data currency and faster recovery than users of lower tiers. Rather than relying largely on transporting tape, as is common on the lower tiers, Tier 4 solutions begin to incorporate more disk-based solutions. Several hours of data loss is still possible, but it is easier to make point-in-time (PIT) copies with a greater frequency than can be achieved using tape-based solutions.

Examples for Tier 4 Business Continuity solutions/technologies are batch/online database shadowing and journaling, Global Copy, FlashCopy®, FlashCopy Manager, Peer-to-Peer Virtual Tape Server, Metro/Global Mirror, IBM Tivoli Storage Manager - Disaster Recovery Manager, FlashCopy Backup/Restore for SAP® Databases, and DS4000™ Integrated Backup for Databases.

1.4.6 Tier 5: Transaction integrity

Tier 5 solutions are used by businesses with a requirement for consistency of data between production and recovery data centers. There is little or no data loss in these solutions; however, the presence of this functionality is entirely dependent on the applications in use.

Examples for Tier 5 Business Continuity solutions and technologies are software and two-phase commits, such as DB2® remote replication, Oracle® Data-Guard, and so on.

1.4.7 Tier 6: Zero or little data loss

Tier 6 business continuity solutions maintain the highest levels of data concurrency and a very low RPO. They are used by businesses with little or no tolerance for data loss and who need to restore data to applications rapidly. These solutions reduce or eliminate the dependence on the application to provide data consistency.

Examples for Tier 6 Business Continuity solutions/technologies are Metro Mirror, Global Mirror, z/OS® Global Mirror, GDPS® HyperSwap™ Manager, Peer-to-Peer VTS with synchronous write, and PPRC Migration Manager.

1.4.8 Tier 7: Highly automated, business integrated solution

Tier 7 solutions include all the major components being used for a Tier 6 solution with the additional integration of automation. This allows a Tier 7 solution to offer more consistency of data than Tier 6 solutions. Additionally, recovery of the applications is automated, allowing for restoration of systems and applications much faster and more reliably than is possible through manual business continuity procedures.

Examples for Tier 7 Business Continuity solutions and technologies are GDPS/PPRC with or without HyperSwap, GDPS/XRC, and AIX HACMP/XD with Metro Mirror.

1.5 Selecting and building a recovery solution

Finding and selecting the right combination of Business Continuity solutions is a challenging task. It aims to optimize the balance between the implementation, the operational, and the maintenance costs, and the benefits of higher availability, flexibility, and disaster resiliency of an enterprise. Solution selection is often driven by DR requirements and implementation cost analysis. The continuous operation and high availability aspects of Business Continuity must not be ignored when evaluating and selecting technologies for a comprehensive Business Continuity solution, since they might offer cost benefits for day-to-day, non-disaster operations of the IT environment.

When looking for a DR oriented Business Continuity solution, the following approach can help you better understand the business requirements, disaster impacts and their potential costs, and support you in making a balanced Business Continuity solution selection.

1.5.1 From business process to recovery objectives

When considering IT Business Continuity solutions, the requirements for continuous availability and recoverability arise at the business process level and need to be mapped down into the IT infrastructure. The only way to keep the implementation cost at an acceptable level is to identify the real recoverability and availability requirements of each business process, and then adopt a suitable Business Continuity solution that meets these requirements.

This can be achieved by performing a business impact analysis at a process-by-process level. The analysis maps the financial impact of a disaster or disruption to each business process, and helps you determine the maximum tolerable cost of the solution for recovering each business process. This analysis also leads to a refined estimate of RTO and RPO for each process.

After the possible impacts and recovery requirements (including an understanding of RTO and RPO) at the business process level are established, they are mapped to the applications that support each business process at the computing services level. The next step is to identify how important the application is to that business process. This helps in finding RTO and RPO values for the applications and sets the design space for the Business Continuity solution. The process of allocating RTOs and RPOs to applications can lead to a large number of different individual application requirements that need to be categorized and standardized. The seven tiers of Business Continuity offer a structure for categorizing the applications.

1.5.2 Matching recovery objectives with Business Continuity technologies

To finalize the Business Continuity technology selection process, the range of recovery objectives offered by different Business Continuity solutions and solution variants, and the range of application Business Continuity tier requirements need to be further narrowed down and matched onto a smaller set of (for example, three) *Business Continuity bands*. Fewer standardized solutions help in simplifying Business Continuity procedures and planning as well as the implementation and operation of the solutions. The Business Continuity band definition level brings the requirements broken down from the business levels together with the recovery capabilities support the selected Business Continuity solutions and their underlying technologies.

At the end of the process, the comprehensive Business Continuity solution might be a mixture of different solutions and technologies that provide you with the business recoverability you need at the minimum cost. Additional information about the approach presented here along with details on the structured methodology of selecting a Business Continuity solution can be found in *IBM System Storage Business Continuity: Part 1 Planning Guide*, SG24-6547. Discussions are further discussed in the following chapters of this book.

1.5.3 SVC features for Continuous Operation

The SVC is designed for the central and flexible management of virtualized storage drawn from underlying disk subsystems and enhanced with added functionality. As such, it offers many features that are useful elements to improve the continuous operation properties of the IT infrastructure that it is a part of. The following sections highlight a few examples where the use of the SVC can improve application availability, and also how it can help to ease management.

Integration of new storage subsystems

The growing demand for storage capacity alongside the best fit for performance and cost targets makes the upgrade and replacement of disk subsystems a regular task in an IT infrastructure. The migration procedures can have an impact on application availability due to any essential modifications to the underlying servers. These could, for example, include driver updates, multi-path software replacements and data copying, some of which may require application downtime or server reboots. One approach to improve the continuous operation characteristics of a storage environment is the use of the online data migration capability of SVC. This feature allows nondisruptive integration of new disk subsystems into the SVC environment and migration of virtual disks (VDisks) onto the new capacity without disrupting the attached server using the VDisks.

Dynamic VDisk features

The SVC allows you to grow the size of a VDisk without interrupting the LUN presentation. This feature can be used to quickly, dynamically, and non-disruptively grow file-systems with operating systems that support it. This is a particularly useful contribution to continuous operation in environments that do not have volume manager software, or where the number of volumes manageable by the operating system is limited. Together with the use of automation software, and the operating system environment, the VDisk growth feature helps you to adapt quickly to any changing demand for file system space, and helps to avoid disruptions due to full file systems.

Attention: The SVC also allows dynamic shrinking of VDisks. This operation reduces the VDisk size by freeing capacity from high logical block addresses. However, lack of supporting operating system features, a compatible volume manager, or a file system data allocation strategy on the volume, can cause data loss when shrinking a VDisk. Make sure you know the data allocation layout on the volume and have a recent and restorable backup before performing a VDisk shrink.

FlashCopy applications

FlashCopy makes a point-in-time copy of a source VDisk to a target VDisk. After the copy operation occurs, the target VDisk has the contents of the source VDisk as it existed at a single point-in-time, also known as a *T0 copy*. The target VDisk can then be used independently from the source for a number of purposes that could otherwise have required interruption to the application availability. A back-up system could use the point-in-time copy on the target disk to perform a consistent backup from stable data. Section 6.4, “Microsoft Volume Shadow Copy Service” on page 205, presents the integration of FlashCopy with IBM Tivoli Storage Manager and an automation interface (Microsoft® Volume Shadow Copy Services, VSS) to enable this scenario.

The use of FlashCopy can also help in application availability by supporting more frequent or repeatable application testing scenarios. A FlashCopy can be used for the quick preparation of a test environment from production data, as the FlashCopy can be taken from the production environment with little impact. Repeatable test runs can be performed by taking a FlashCopy prior to each run and performing the test on the copy.

1.5.4 SVC features for high availability

High availability describes the ability of a system to mask the consequences of a component failure to the system’s users. The SVC is designed to combine storage engines into a cluster designed to support high availability. The SVC

storage engines are built on extremely reliable hardware with redundant components. While the SVC cannot increase the HA characteristics of other systems in the IT infrastructure, its HA design helps to make SVC a strong link in the chain from the application down to the data. Through its design, the SVC can contribute to the overall HA characteristics of a comprehensive Business Continuity solution.

1.5.5 Disaster recovery designs with SVC

Disaster recovery is the process to recover data center operations at a different site after a disaster renders the main computing services facilities inoperable. The SVC provides the Metro Mirror and Global Mirror functions for maintaining synchronous or asynchronous consistent data copies at a secondary site to support Tier 6 and Tier 7, dual site DR solutions. Chapter 3, “SAN Volume Controller Mirroring Solutions” on page 51 provides a detailed look at the mode of operation and some uses of SVC Remote Copy functionality.

The SVC integrates automation technologies that perform Remote Copy operations on sets of VDisks. This can help you towards achieving continuous availability at the application level.



Designing Business Continuity solutions with the SVC

This chapter introduces the SAN Volume Controller Copy Services and shows how each one can address a different Business Continuity tier. It also discusses the factors to consider when implementing Business Continuity with the SVC.

For a complete discussion of the SVC, we recommend that you read the IBM Redbooks publication *IBM System Storage SAN Volume Controller*, SG24-64233.

2.1 Design

Has an automatic teller machine ever refused to give you cash because it was “temporarily out of service”? Have you ever stood in line at a department store register, while the line never moved, only to have the cashier explain, “Sorry, the computer is down”?

These scenarios may be as a result of typical computer outages. Sometimes outages are caused by faulty hardware or software. Although unplanned outages such as these are still fairly common with end user systems, they have become rare in recent years for sophisticated multiuser systems.

Other outages are caused by hardware, software, and maintenance work that needs to be performed at regular intervals. Planned outages are usually scheduled at a period of low activity, typically in the early hours of the morning. We are accustomed to the fact that many systems are not operational for a certain time period every night.

In today’s business environment, an increasing number of enterprises require that both types of computer outages, planned and unplanned, are eliminated or substantially reduced. An advanced computer system is not a single box, but a complex assembly of components such as processors, peripherals, networks, operating systems, and application software. Consequently, an outage-free system cannot be purchased off the shelf. It requires an appropriate system design, using elements such as redundant or fault tolerant components and the most appropriate systems management.

Designing and implementing a suitable continuous availability solution is not a simple task. It involves considerable effort and expense. A solution involves the following activities:

- ▶ It must be designed and built to match the business requirements.
- ▶ It may require additional computing and infrastructure resources.
- ▶ It certainly requires the development and testing of many new procedures. These new procedures and processes must be compatible with existing operations. Staff from different departments are involved, and they must work together when developing and implementing the solution.
- ▶ It involves a trade-off between cost, performance, recovery speed, and the scope of outages covered.

2.2 Concepts

Business Continuity is a complex subject because it involves not just the infrastructure technology, but also different disciplines such as human resources and facilities. The cornerstones of Business Continuity are as follows:

- ▶ High availability
- ▶ Continuous operations
- ▶ Continuous availability

2.2.1 High availability

High availability is the ability of a system to provide service to its users during defined service periods, at an acceptable or agreed level.

These service periods, as well as a definition of an “acceptable” service level, are either stated in a service level commitment by the service provider, or in a service level agreement between end users and the service provider.

Typically, a service level above 99.7 percent is accepted as high availability. High availability is maintained by avoiding or reducing any unplanned outages. The fact that the service is only provided during defined service periods (for instance, between 6:00 AM and 8:00 PM) makes it possible to perform most change and maintenance work outside those service hours.

2.2.2 Continuous operation

Continuous operation is the ability of a system to provide service to its users day and night without scheduled outages to perform system and data maintenance activities.

It is obviously difficult to perform change and maintenance work on a system that is supposed to be in continuous operation. However, without preventative maintenance, a system can be in continuous operation, but its availability may not be as high as it should be since it may suffer unscheduled outages more frequently.

Even if true continuous operation turns out to be impossible, to implement some applications a realistic goal might be to increase the defined service period, for example, from 14 hours per day to 18 hours per day. In this book, we discuss solution approaches that can contribute to extensions of the service period in the continuous availability context.

2.2.3 Continuous availability

Continuous availability is the ability of a system to provide both high availability and continuous operation at the same time. The system must be designed so that users experience scheduled nor unscheduled outages.

This goal seems difficult to achieve, as hardware and software components are usually not entirely error free and maintenance free, and large computer systems undergo frequent component additions and changes. The solution is to employ hardware components, software, and operational procedures that mask outages from the user. This solution usually requires that recovery from an outage must be performed so quickly that the user does not perceive it as an outage. It also frequently requires the use of redundant components, so that an alternate component can be used in case of a permanent component failure or while a component is in maintenance.

2.2.4 Business requirements

Analyzing the business impact of unplanned outages should lead to specifying the desired improvements in terms of high availability and continuous availability.

The continuous availability specifications determined in the business requirements analysis are called *service level objectives*. They are used as input to the next step in our structured systems design approach.

Business impact due to an outage

An unplanned outage of computer systems results in some business processes coming to a standstill. From a business point-of-view, it may result in financial impact in several ways.

Lost work time of computer users

End users who need the systems for their work might be idle for as long as the systems are down. They might have to repeat certain parts of their work when the systems are up again. They even might require overtime to make up for the outage.

Lost customer transactions

If the computer serves any customer transactions online, the customer is not likely to wait for the system to come up again. The customers might try to repeat their online transactions at a later time, but then, they might not. Some transactions, and any associated turnover, are lost through an outage.

Lost computer capacity

The work that is interrupted through an outage needs to be repeated at a later point in time. Such repetition of workload requires extra system resources. Systems that experience frequent outages need some spare capacity to make up for these outages.

In large computer networks, the sum of these cost items can amount to a substantial figure. A business impact analysis, conducted by experienced consultants, can try to determine that cost.

Benefits of continuous operation

There are a number of potential financial gains for the enterprise that can be realized by extending the service periods. The business can benefit in several ways.

Increased turnover

Where there is a direct relationship between computer transactions and the business volume, as with an order entry system, extension of the service availability periods can result in increased turnover.

Improved service quality

Some businesses need to offer customer service (for instance, over the telephone) beyond the scheduled information systems' service periods. Sometimes the customer service agents have to rely on reduced or back-level information during these times. Providing these agents (and the customers) with access to up-to-date online information improves overall service quality, and leads to better customer acceptance.

Customer operated transaction services

Technological advances have inevitably shifted much of the transaction workload outside the normal business hours and that requires supporting computer systems to be operative at all times.

Global services spanning many time zones

Globally operating enterprises frequently offer a single computer service in many countries, spanning many time zones. The traditional concept of "business hours" is no longer applicable in these cases, as the computer service periods must be substantially expanded or even be available at all times.

Measuring Business Continuity

At this point, the required improvement of service in terms of high availability and continuous operation need to be documented, for example, by defining the following items for an online service:

- ▶ The hours of service
- ▶ Maximum/minimum response time for different applications
- ▶ Maximum number of outage per time interval
- ▶ Mean Time To Repair (MTTR) and recovery time
- ▶ Process or print turnaround times

These service level objectives serve as input to the next step (determining the data processing requirements) in our structured system design approach. The service level objectives have to be negotiated with the end users and eventually be turned into service level agreements.

2.2.5 Application requirements

The following sections describe the requirements that applications have.

Recovery Point Objective

The Recovery Point Objective (RPO) is the requirement for currency of data. Also expressed as the amount of data that could acceptably be recreated post disaster. For example, a business that believes it could acceptably afford to create or lose five minutes worth of data has an RPO of five minutes.

Recovery Time Objective

The Recovery Time Objective (RTO) is the requirement for restoration of systems and applications, expressed in terms of how long a business can afford to have systems and applications down after a disaster. For example, a business that believes that it could afford to be without systems for eight hours has a RTO of eight hours.

Replication

Synchronous or asynchronous replication methods are selected by considering between RPO, RTO, and any application tolerance to latency. Synchronous replication provides the fastest resumption of business operation, while asynchronous copy lacks some data that sometimes can be recreated by the application or may actually be lost.

Asynchronous replication has RPOs and RTOs larger than synchronous replication, but allows replication over longer distances. Synchronous remote copy adds latency directly to the application I/Os response time.

Time dependency

The acceptable disk latency that an application can cope with must be investigated before the Business Continuity design phase. This factor is important to pick the correct copy service, intersite technology, and even to determine the maximum distance between primary and secondary sites. There are ways to reduce latency, but when huge distances are necessary, even with those methods in place, applications suffer extra disk latency. This maximum time must be known before making any decisions. Asynchronous remote copy must be used when considerable distances (and therefore latency) are expected, and this additional latency cannot be tolerated by the applications.

Intersite link bandwidth

Remote data replication needs a certain amount of bandwidth to transmit every write operation from primary disks to secondary disks. Sufficient link bandwidth must be available to allow data flow without impacting application performance even on the busiest days. Use of redundant links is a good practice. Underestimating link bandwidth can cause performance problems and compromise RTO and RPO. Link bandwidth is discussed later.

2.3 Factors to consider for BC with SVC

The SVC acts as a cache for read/write SCSI operations between hosts and storage arrays controllers. Factors to consider are SAN, SAN extensions, host attachment, storage arrays, performance, and automation.

2.3.1 Host attachment

Host attachment, as any part of the whole solution, must allow business continuous availability. It means a single server must have at least two host bus adapters and a software layer that can use both paths to provide redundancy and load balance. The number of Host Bus Asaptors (HBAs) must be enough to take care of the system load as well. A server with two HBAs can be considered highly available, but if two HBAs cannot handle the system load it can bring a high response time into the environment, and from the users point-of-view the system as an entity is not available.

2.3.2 Storage array controllers

Storage array controllers must allow continuous availability in a method that is independent of the SVC. The SVC does not provide any kind of protection for disk failures. This is an exclusive function of an array controller. In the same way,

storage array controllers must be redundantly connected to the SAN and be capable of handling the application workload.

SVC Copy Services can add extra overhead to the storage array controllers and it must be carefully computed during the design phase. This is mainly when Remote Copy Services are implemented. Careful consideration is required when using asynchronous remote copy as the secondary storage, as it directly impacts the performance of the application performance.

2.3.3 SAN extension

SAN extensions connect the primary fabric where the source SVC cluster resides with a secondary fabric where the target SVC cluster is connected so that they can communicate with each other to implement an intercluster partnership.

Intersite communication is one of the most important topics in a Business Continuity solution when considering Disaster Recovery. Intersite communication choice is driven mainly by the business and application requirements. For example, if the distance between primary and secondary sites for an enterprise must be at least 100 km and can tolerate a latency of 5 ms, we will see later in this section that each 100 km introduces 1 ms of latency. Latency is an important consideration and must not be overlooked.

Another point to consider is the size of the link. The bandwidth must be carefully studied to allow data transfer without affecting performance at the primary site and without keeping remote copy from achieving its RPO.

In 2.5, “Distance factors for intersite communication” on page 34 we discuss this concept more.

2.3.4 SVC Copy Services

The SVC Copy Services are FlashCopy, Metro Mirror, and Global Mirror.

FlashCopy

FlashCopy is a point-in-time copy of a virtual disk on an SVC. The target volume is a completely new self-sufficient entity. The copy occurs between a source virtual disk and a target virtual disk. A FlashCopy can spread across different storage subsystems of even different vendors in the SVC back-end. Both copies may be updated independently. The target volume is immediately available and a background copy of the data starts (if not stopped by FlashCopy options). The functions can be invoked through the GUI, CLI, or scripts.

Important for disaster recovery scenarios is that FlashCopy supports Consistency Groups. This means that a group of virtual disks that belong to the same application may hold related data, and that for data consistency all of these virtual disks must be flashed together at a common point-in-time. FlashCopy can be used to design a Business Continuity solution from Tier 0 up to Tier 4. For a discussion of the Business Continuity tier models see section 1.4, “The seven tiers of disaster recovery” on page 9.

Metro Mirror

Metro Mirror is a constant mirror of a VDisk. It occurs between a source VDisk and a target VDisk. Metro Mirror operates in synchronous mode only. This means that the acknowledgement of the I/O is given to the host after the data is written to the secondary virtual disk. The source and target virtual disk can belong to the same I/O group, referred to as intracluster Metro Mirror, or to two different SVCs, also referred to as intercluster Metro Mirror.

The Metro Mirror functions can be invoked through the GUI, CLI, or scripts.

Note: SVC Metro Mirror supports Consistency Groups, which is a group of volumes that belong logically together and can be remotely mirrored.

Metro Mirror implements a synchronous copy between volumes from SVC clusters. Synchronous copy occurs when updates (writes) are written to both the local and remote site, and when both write operations are completed, the application is notified that the write has completed successfully. Synchronous replication is usually very sensitive to both bandwidth and latency. These factors may negatively influence I/O and application response times. Metro Mirror can be used in Tier 6 and Tier 7 of Business Continuity.

Metro Mirror is discussed in detail in Chapter 3, “SAN Volume Controller Mirroring Solutions” on page 51.

Global Mirror

Global Mirror occurs between a source VDisk and a target VDisk. Global Mirror is a constant mirror of a VDisk. It shares Metro Mirror commands and GUI interfaces. There is a time difference between the source and target VDisk. This time difference between source VDIs and target VDIs is the RPO. Global Mirror implements asynchronous copy between volumes from SVC clusters. Asynchronous happens when updates (writes) are written to the local site and then the application is notified that the write is completed successfully. The write operation is buffered locally and sent to the remote site at a later time, depending on the availability of the bandwidth. Asynchronous replication is impacted much less by latency. A reasonable bandwidth must be guaranteed so that the data

gets replicated in a timely manner to guarantee an agreed RPO. Global Mirror can be used in Tier 6 of Business Continuity.

Global Mirror is discussed in detail in Chapter 3, “SAN Volume Controller Mirroring Solutions” on page 51.

2.3.5 Performance

Before starting with SVC Remote Copy Services it is important to consider any overhead associated with their introduction. It is important that you know your current infrastructure fully.

Major attention must be paid to link distance and bandwidth, the current SVC clusters’ load, and the current storage array controllers load.

Bandwidth analysis and capacity planning for your links will help to define how many links you need and when you need to add more links in order to ensure the best possible performance and high availability. Bandwidth is discussed in detail in section 2.6.2, “Determining the bandwidth” on page 47.

As part of your implementation project, you may be able to identify and then distribute hot spots across your configuration, or take other actions to manage and balance the load.

You must consider the following:

- ▶ Is your bandwidth so little so that you may see an increase in the response time of your applications at times of high workload?
- ▶ Remember that the speed of light is less than 300,000 km/s, that is less than 300 km/ms on fiber. The data must go to the other site, and then an acknowledgement has to come back. Add any possible latency times of some active components on the way, and you approximately get one ms overhead per 100 km for write I/Os. Metro Mirror adds extra latency time due to the link distance to the time of write operation.
- ▶ Can your current SVC cluster or clusters handle extra load?
- ▶ Sometimes the problem is not related to SVC Remote Copy Services but rather hot spots on the disks subsystems. Be sure these problems are resolved.
- ▶ Is your secondary storage capable of handling the additional workload it receives—basically the same workload as the primary applications? Is it going to further slow down the response time?

A detailed discussion on SVC Remote Copy Services and performance is given in Chapter 4, “Performance considerations in SVC Business Continuity solutions” on page 119.

2.3.6 Automation

SVC Remote Copy Services are a hardware mirroring solution. A volume (or LUN) is paired with a volume (or LUN) in the remote disk subsystem (this is called a relationship). As the size of the environment grows, so does the complexity of managing it. You need a means for managing relationships, ensuring they are in a consistent, synchronized state, adding volume pairs as required, monitoring error conditions, and managing data consistency across SVC clusters.

When planning a SVC Remote Copy Service environment, consider the following topics:

- ▶ Design
- ▶ Maintenance
- ▶ Testing
- ▶ Support

Considering these topics prevents you from creating a situation where you implement an untested solution.

There are also solutions available for integrating SVC Remote Copy Services Mirror into a cluster, such as HACMP/XD for AIX (CAA), or IBM Total Storage Continuous Availability for Windows® (CAW). For more information see Chapter 5, “Automation in a Business Continuity solution with SVC” on page 155.

2.4 SAN extension

Intersite communication comprehends all aspects of interconnection between two nonadjacent sites for data copy purposes. Depending on the Business Continuity solution requirements, primary and secondary sites could be on different continents.

Intersite communication allows local and remote components to communicate as if they are local to each other.

In the particular case of the SVC, both SVC nodes in a cluster and partners use the Fibre Channel Protocol as a medium of packet transportation.

Inter-node communication packets carry messages to allow SVC cluster implementation. Inter-cluster communication packets carry messages to allow SVC inter-cluster implementations, as well as data from the primary SVC cluster to the secondary SVC cluster.

Inter-site communication is a complex matter and the objective of this topic is to discuss only the pieces of the of inter-site communication that affect the SVC and relate to the SAN infrastructure.

2.4.1 SAN, MAN, and WAN

In the topics that follow we discuss SAN, MAN, and WAN.

Storage Area Network

It is no secret that data is business. Because of the increased dependence on information, the storage needs of companies are growing exponentially. Research shows that storage needs for traditional brick-and-mortar companies are doubling every year and every 90 days for dot coms. This growth raises new concerns for the maintenance and protection of valuable data resource.

Historically in open system storage environments, physical interfaces to storage consists of parallel SCSI channels supporting a small number of SCSI devices. Storage Area Networks (SAN) use new technologies to connect a greater number of servers and devices. Deployment of SANs today is exploiting the storage focused capabilities of Fibre Channel. The Fibre Channel SAN consists of hardware components such as storage subsystems, storage devices, and servers that are attached to the SAN through interconnect entities—host-bus adapters, routers, hubs, and switches. Another hardware element being deployed in the SAN is commonly referred to as a *SAN appliance* or *SAN server*. These SAN appliances are computing elements attached directly to the SAN or installed in the storage data path. These SAN appliances are responsible for managing the Fibre Channel topology and additionally providing an abstraction of storage.

The heart of the storage management software is *virtualization*. The term virtualization, when it pertains to disk storage, refers to the representation of a storage unit or data to the operating system and application running on an application server. The storage unit or data presented is decoupled from the actual physical storage where the information may be contained. There is some method for providing the translation between the logical and physical storage. After the storage is abstracted, storage management task is performed with a common set of tools from a centralized point, which greatly reduces the cost of administration.

Metropolitan Area Network

A Metropolitan Area Network (MAN) is typically a city wide solution, connecting systems across 50 km up to 100 km. These networks typically carry traffic between office locations within a metropolitan area, hence the name. A MAN often carries many diverse networking protocols, and these protocols have their

own speed and performance characteristics. It can also be an intermediate media interfacing between a SAN and a Wide Area Network (WAN).

The MAN has two critical goals to achieve. It must meet the needs created by the dynamics of the ever increasing bandwidth requirements. It must also address the growing connectivity requirements and data access technologies that are resulting in demand for high-speed, customized data services.

Wide Area Network

Wide Area Network (WAN) is at the very core of a global network. Like the MAN, the WAN acts as a transport medium. This means they need to be resilient with a high level of capacity. These networks are often provisioned by SONET or SDH technology. Due to the high demand for bandwidth, these solutions are experiencing large demand on their fiber. WANs are likely to be public (in the case of the Internet) or private in the case of a company's self deployed dedicated WAN. Some WANs are based on an agreed bandwidth within a public network, and this leads to virtual private networks. Most large corporations deploy one of the last two options or a combination of the two.

2.4.2 Replication topology

Network topology is the way that sites are connected. Instead of two sites using a point-to-point topology, sites (and it can be more than two) are connected in different ways, such as a ring or mesh topology. The SVC supports a partnership with one unique SVC cluster. This means we can only think in point-to-point topology for SVC Remote Copy Services design, but we can still take advantage of the SAN when we need to.

2.4.3 Technology options for SAN extension

This section reviews the available technology for extending a SAN.

We roughly divide SAN extension technology into *Optical* and *IP*. In the Optical slice we have technologies such as dark fiber, coarse wavelength division multiplexing (CWDM), dense wavelength division multiplexing (DWDM), and SONET. These technologies transport FC frames from a primary to a secondary site. In the IP world, we have implementations such as FCIP and IFCP, where FC frames are encapsulated in IP datagrams before transportation. Table 2-1 summarizes some common technologies and their respective speed.

A deeper discussion about distance technologies can be found in *Introduction to SAN Distance Solutions*, SG24-6408.

Table 2-1 Interconnect technology summary

Technology	Speed
OC-255	13.21 Gbps
OC-192	10 Gbps
OC-96	4.976 Gbps
OC-48, STS-48	2.488 Gbps
OC-36	1.866 Gbps
OC-24	1.244 Gbps
OC-18	933.12 Mbps
OC-12, STS-12	622.08 Mbps
OC-9	466.56 Mbps
OC-3, STS-3	155.52 Mbps
CDDI, FDDI, Fast Ethernet, CAT 5	100 Mbps
OC-1, STS-1	51.84 Mbps
T-3, DS-3 North America	44.736 Mbps
E-3 Europe	34.368 Mbps
CAT 4	20 Mbps
Token Ring LANs	16 Mbps
Thin Ethernet, CAT 3, Cable Modem	10 Mbps
DWDM	1, 2, 10 Gbps
CWDM	1,2 Gbps
Dark fiber	-

Synchronous Optical NETWORK/SDH

A Synchronous Optical NETWORK or SONET (SDH in Europe) is a standard for transporting data across public telecommunication rings. Typically a business contracts with a telecommunication or Network Connectivity provider for a specific amount of bandwidth. The customer's building is then provided leads (the specific type of lead depends on how much bandwidth was contracted for) into the telecommunication company's network that is set up as a series of ring configurations. This allows the channel extension devices to connect the equipment in the data centers into the high speed network.

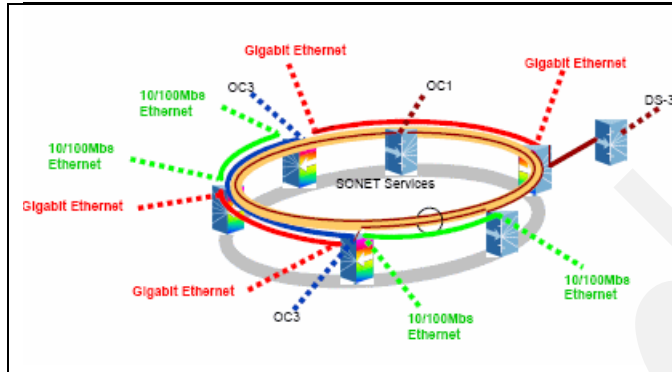


Figure 2-1 SONET representation

Because it is based on a ring or mesh topology, the SONET infrastructure is usually self-healing. If any point in the SONET ring is disabled, traffic is re-routed in the opposite direction. As a result this technology provides very high availability without necessitating a secondary route.

An example of a SONET ring is given in Figure 2-1. Here you see the leads entering the public network through leads OC1, OC3, and Gigabit Ethernet. These names refer to the amount of bandwidth available in that pipe, which is then mirrored in the amount of bandwidth that is provisioned for that pipe in the network. So, as an example, a business leasing a single OC3 through IBM Global Services must be provided with a link into its facility that is capable of handling 155 Megabits per second (Mbps). This pipe must then lead into the larger public network where that business receives 155 Mbps of the total shared bandwidth available.

Dark fiber

Dark fiber is fiber that is not shared, and thus not lit by other users. This is typically a privately owned fiber optic route that, unlike shared SONET rings, only has light passing through it when its owner connects devices. Because it is privately owned and there are no competing users, the business gets full use of the bandwidth provisioned at any time. The downside is that it tends to be an expensive solution and usually requires some form of multiplexing to control fiber transport costs as the business grows. Due to technical limitations, dark fiber can only be used for relatively short distances, especially when compared to SONET. Dark fiber is often used in Data Center Point-to-Point configurations.

Wave Division Multiplexing (WDM)

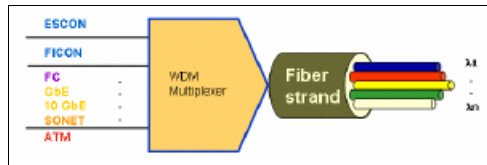


Figure 2-2 WDM

White light passing through a prism is split into a wider spectrum of colors, each with its own wavelength. In the Wave Division Multiplexing (WDM) world this is reversed and each device connected to the WDM is given a wave length of light (known as a *lambda*). This is similar to the different colors in the spectrum. The WDM takes these wavelengths and allows them to pass together across the fiber in the same area of the light spectrum, around 1550 nanometers. Figure 2-2 demonstrates the flow as the channels connect to the WDM and enter the fiber strand.

At the remote location the light passes through a second WDM. This device takes in the white light transmission of data and divides it back into individual wavelengths and connects to devices as though it were attached by its own fiber optic strand. Because this allows many devices to connect over a single pair of single-mode fiber strands, this represents a great improvement in efficiency over direct connections through individual fiber strands and represents a major cost saving given the expense involved in building up a fiber optic infrastructure.

Additionally, WDM technology represents a long term investment protection of the dark fiber infrastructure. Improvements in WDM technology improve the efficiency of the dark fiber infrastructure over time, thus reducing or removing the need to add additional fiber strands when more channels can be added to the existing WDM infrastructure.

There are two forms of WDM technology:

- ▶ Coarse Wavelength Division Multiplexing (CWDM)
- ▶ Dense Wavelength Division Multiplexing (DWDM)

While the basics are the same for both these forms, it is important to understand that each has options that affect the cost and scalability of solutions built upon them.

CWDM spreads out the lightwaves instead of trying to keep them closer together. The result is a smaller increase in the number of *lambdas* available through the fiber optic strands. Although this is not an advantage of CWDM technology when compared to DWDM, it tends to be significantly less expensive. As a result, it is an appropriate technology for businesses that use dark fiber but only have a

limited number of cross site links. CWDM is normally capable of sending eight channels of data across one pair of fiber strands. This, however, can be increased through sub-rate multiplexing.

The most common form of multiplexing is DWDM that tries to keep the lambdas as close together as possible. As a result, the number of channels that can pass through a pair of fiber optic strands is greatly increased. Current standards allow for up to 32 lambdas to pass through one pair of fiber strands, and this can be multiplied through technologies such as sub-rate multiplexing.

TCP/IP extension

This section describes IP based extension technologies.

Fibre Channel Over TCP/IP

Fibre Channel Over TCP/IP (FCIP) describes mechanisms that allow the interconnection of islands of Fibre Channel storage area networks over IP-based networks to form a unified storage area network in a single Fibre Channel fabric. FCIP relies on IP-based network services to provide the connectivity between the storage area network islands over local area networks, metropolitan area networks, or wide area networks.

Fibre Channel standards use distances between switch elements that are less than the distances available in an IP network. Since Fibre Channel and IP networking technologies are compatible, it is logical to turn to IP networking for extending the allowable distances between Fibre Channel switch elements.

The fundamental assumption made in the FCIP protocol is that the Fibre Channel traffic is carried over the IP network in such a manner that the Fibre Channel Fabric and all Fibre Channel devices on the Fabric are unaware of the presence of the IP network. This means that the Fibre Channel datagrams must be delivered in such a time frame as to comply with existing Fibre Channel specifications. The Fibre Channel traffic may span LANs, MANs, and WANs, so long as this fundamental assumption is adhered to.

For more information about FCIP, refer to the IETF standards for IP storage at the following Web location:

<http://www.ietf.org/rfc/rfc3821.txt>

Note: Brocade and Cisco rely on FCIP to extend Fabric communication over distances.

iFCP

iFCP is a gateway-to-gateway protocol that provides Fibre Channel fabric services to Fibre Channel devices over a TCP/IP network. iFCP uses TCP to

provide congestion control, error detection, and recovery. iFCP's primary objective is to allow interconnection and networking of existing Fibre Channel devices at wire speeds over an IP network.

The protocol implementation permits the attachment of Fibre Channel storage devices to an IP-based fabric by means of transparent gateways.

The protocol achieves this transparency by allowing normal Fibre Channel frame traffic to pass through the gateway directly, with provisions where necessary, for intercepting and emulating the fabric services required by a Fibre Channel device.

For more information about iFCP, refer to the IETF standards for IP storage at the following Web location:

<http://www.ietf.org/rfc/rfc4172.txt>

Note: McDATA equipment relies on the iFCP protocol to extend Fabric communication over distance.

2.5 Distance factors for intersite communication

In this section we introduce the key factors that distance brings to our attention when planning a solution where data is moved from one site to another.

2.5.1 Link speed, bandwidth, and latency

The *speed* of a communication link determines how much data can be transported and how long the transmission takes. The faster the link the more data can be transferred within a given amount of time.

Bandwidth is the network capacity to move data as measured in millions of bits per second (Mbps) or a billions of bits per second (Gbps).

In storage terms, bandwidth measures the amount of data that can be sent in a specified amount of time. Networking link bandwidth is usually measured in bits and multiples of bits.

Applications issue read and write requests to storage devices, and these requests are satisfied at a certain speed commonly called the *data rate*. Usually disk and tape device data rates are measured in bytes per unit of time and not in bits. One million bytes per second is expressed as 1MBps. Current technology storage device LUNs or volumes can manage sequential sustained data rates in the order of 10 MBps to 80-90 MBps. In other terms an application writes to disk

at 80 MBps. Assuming a conversion ratio of 1 MB to 10 Mbits (this is reasonable because it accounts for protocol overhead) we have a data rate of 800 Mbits. It is always useful to check and make sure that one correctly co-relates MBps to Mbps.

Latency is the time taken by data to move across a network from one location to another and is measured in milliseconds.

The longer the time, the greater the performance impact. Latency depends on the speed of light ($c = 3 \times 10^8 \text{m/s}$, vacuum = 3.3 microsec/km (microsec represents microseconds, one millionth of a second)). The bits of data travel at about two-thirds the speed of light in an optical fiber cable.

However, some latency is added when packets are processed by switches and routers and then forwarded to their destination. While the speed of light may seem infinitely fast, over continental and global distances latency becomes a noticeable factor. There is a direct relationship between distance and latency. Speed of light propagation dictates about one millisecond latency for every 100 miles. For some synchronous remote copy solutions, even a few milliseconds of additional delay may be unacceptable. Latency is a difficult challenge because, unlike bandwidth, spending more money for higher speeds reduces latency.

Tip: SCSI write over Fibre Channel requires two round trips per I/O operation, we have $2 \text{ (round trips)} \times 2 \text{ (operations)} \times 5 \text{ microsec/km} = 20 \text{ microsec/km}$. At 50 km we have an additional latency of $20 \text{ microsec/km} \times 50 \text{ km} = 1000 \text{ microsec} = 1 \text{ msec}$ (msec represents millisecond). Each SCSI I/O has one msec of additional service time. At 100 km it becomes two msec additional service time.

In this section we discuss techniques to improve service time such as write acceleration and compression.

2.5.2 Link quality

The optical properties of the fiber optic cable influence the distance that can be supported. There is a decrease in signal strength along a fiber optic cable. As the signal travels over the fiber, it is attenuated, and this is caused by both absorption and scattering, and is usually expressed in decibels per kilometer (dB/km). Some early deployed fiber is designed to support the telephone network, and this is sometimes insufficient for today's new multiplexed environments. If you are being supplied dark fiber by another party, you normally specify that they must not allow more than xdB loss in total.

The decibel (dB) is a convenient way of expressing an amount of signal loss or gain within a system or the amount of loss or gain caused by some component of a system. When signal power is lost, you never lose a fixed amount of power. The rate at which you lose power is not linear. Instead you lose a portion of power: one half, one quarter, and so on. This makes it difficult to add up the lost power along a signal's path through the network if measuring signal loss in watts.

For example, a signal loses half its power through a bad connection, then it loses another quarter of its power on a bent cable. You cannot add $1/2$ plus $1/4$ to find the total loss. You must *multiply* $1/2$ by $1/4$. This makes calculating large network dB loss both time-consuming and difficult.

Decibels, though, are logarithmic, allowing us to easily calculate the total loss/gain characteristics of a system just by adding them up. Keep in mind that they scale logarithmically. If your signal gains 3dB, the signal doubles in power. If your signal loses 3dB, the signal halves in power.

It is important to remember that the decibel is a ratio of signal powers. You must have a reference point. For example, you can say, "There is a 5dB drop over that connection." But you cannot say, "The signal is 5dB at the connection." A decibel is not a measure of signal strength; instead, it is a measure of signal power loss or gain.

A decibel milliwatt (dBm) is a measure of signal strength. People often confuse dBm with dB. A dBm is the signal power in relation to one milliwatt. A signal power of zero dBm is one milliwatt, a signal power of three dBm is two milliwatts, six dBm is four milliwatts, and so on. Do not be misled by minus signs. It has nothing to do with signal direction. The more negative the dBm goes, the closer the power level gets to zero.

A good link has a very small rate of frame loss. A re-transmission occurs when a frame is lost, directly impacting performance.

2.5.3 Hops

The hop count as such is not increased by the intersite connection architecture. For example, if we have our SAN extension based on DWDM, the DWDM components are transparent to the number of hops. The hop count limit within a fabric is set by the fabric devices (switch or director) operating system and it is used to derive a frame hold time value for each fabric device. This hold time value is the maximum amount of time that a frame can be held in a switch before it is dropped or *the fabric is busy* condition is returned. For example, a frame may be held if its destination port is not available. The hold time is derived from a formula using the error detect time-out value and the resource allocation time-out value.

The discussion on these fabric values is beyond the scope of this book. However further information can be found in *IBM TotalStorage: SAN Product, Design, and Optimization Guide*, SG24-6384.

If these times become excessive, the fabric experiences undesirable time outs. It is considered that every extra hop adds about 1.2 microseconds of latency to the transmission.

Currently, SVC Remote Copy Services supports three hops when protocol conversion exists. That means if you have DWDM extended between primary and secondary sites, three SAN directors or switches can exist between primary and secondary SVC.

2.5.4 Buffer credits

SAN device ports need memory to temporarily store frames as they arrive, assemble them in sequence, and deliver them to the upper layer protocol. The amount of frames that a port can hold is called its *Buffer Credit*.

Fibre Channel architecture is based on a flow control that ensures a constant stream of data to fill the available pipe.

When two FC ports begin a conversation they exchange information about their buffer capacities. An FC port sends only the number of buffer frames for which the receiving port has given credit. This not only avoids overruns, but also provides a way to maintain performance over distance by filling the pipe with in-flight frames or buffers.

Tip: A rule-of-thumb says that to maintain acceptable performance one buffer credit is required for every two km distance covered.

BB_Credit

During login, N_Ports and F_Ports at both ends of a link establish its Buffer to Buffer Credit (BB_Credit).

EE_Credit

In the same way during login all N_Ports establish End to End Credit (EE_Credit) with each other. During data transmission a port should not send more frames than the buffer of the receiving port can handle before getting an indication from the receiving port that it has processed a previously sent frame. Two counters are used for that: BB_Credit_CNT and EE_Credit_CNT. Both are initialized to zero during login.

Each time a port sends a frame it increments BB_Credit_CNT and EE_Credit_CNT by one. When it receives R_RDY from the adjacent port it decrements BB_Credit_CNT by one. When it receives ACK from the destination port it decrements EE_Credit_CNT by one. Should at any time BB_Credit_CNT become equal to the BB_Credit or EE_Credit_CNT equal to the EE_Credit of the receiving port, the transmitting port has to stop sending frames until the respective count is decremented.

The previous statements are true for Class 2 service. Class 1 is a dedicated connection, so it does not need to care about BB_Credit and only EE_Credit is used (EE Flow Control). Class 3 on the other hand is an unacknowledged service, so it only uses BB_Credit (BB Flow Control), but the mechanism is the same on all cases.

Here we can see the importance that the number of buffers has in overall performance. We need enough buffers to make sure the transmitting port can continue sending frames without stopping in order to use the full bandwidth.

This is particularly true with distance. At 1Gbps a frame occupies 4 km of fiber. In a 100 km link we can send 25 frames before the first one reaches its destination. We need an ACK (acknowledgment) back to the start to get our EE_Credit full again. We can send another 25 before we receive the first ACK. We need at least 50 buffers to allow for non stop transmission at 100 km distance.

The maximum distance that can be achieved at full performance depends on the capabilities of the FC node that are attached at either end of the link extenders. This is vendor specific. There should be a match between the buffer credit capability of the nodes at either end of the extenders. A *host bus adapter* (HBA) with a buffer credit of 64 communicating with a switch port with only eight buffer credits can read at full performance over a greater distance than it can write. This is because on the writes the HBA can send a maximum of only eight buffers to the switch port, while on the reads, the switch can send up to 64 buffers to the HBA.

2.5.5 TCP window size

TCP performance depends not upon the transfer rate itself, but rather upon the product of the transfer rate and the *round-trip time* (RTT). This bandwidth x RTT formula measures the amount of data that may fill the pipe. It is the buffer space required at the sender and receiver to obtain maximum throughput on the TCP connection over the path. This means the amount of unacknowledged data that TCP must handle in order to keep the pipeline full.

TCP implements reliable data delivery by retransmitting segments that are not acknowledged within some *retransmission timeout* (RTO) interval. Accurate

dynamic determination of an appropriate RTO is essential to TCP performance. RTO is determined by estimating the mean and variance of the measured RTT, which means the time interval between sending a segment and receiving an acknowledgment for it.

The easiest way to determine the round-trip time is to use the **ping** command from one host to another, and use the response times returned by the **ping**.

Tip: TCP Window Size = bandwidth x RTT.

If the value of the TCP Window Size is over estimated, link oscillations occur with drops. Retransmission starts from zero and ramps up until it hits the bandwidth size. You must be conservative with these values to avoid drops, or worse, entering in a repetitive situation as shown in Figure 2-3.

Today, implementation of FCIP and IFCP allow us to enter a value for the minimum bandwidth. This value determines the new start point from where TCP retransmission starts.

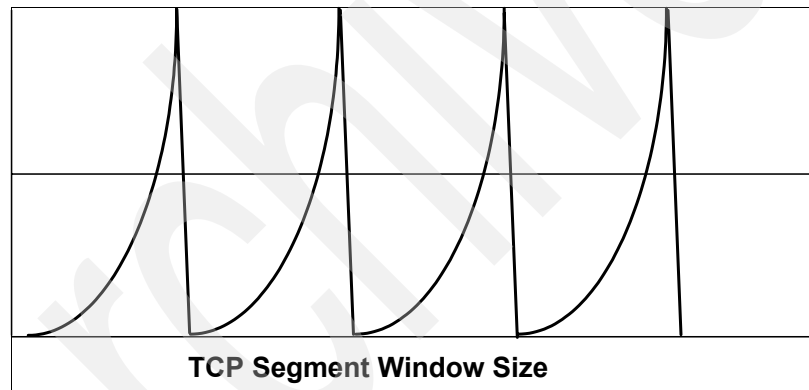


Figure 2-3 TCP/IP congestion mechanism

Refer to the following Web site for more information about TCP Window Size:

<http://www.ietf.org/rfc/rfc1323.txt>

2.5.6 TCP selective acknowledgment

Multiple packet losses from a window of data can have a catastrophic effect on TCP throughput. TCP uses a cumulative acknowledgment scheme in which received segments that are not at the left edge of the receive window are not acknowledged. This forces the sender to either wait for a round-trip time to find out about each lost packet or to unnecessarily retransmit segments that were

correctly received. With the cumulative acknowledgment scheme, multiple dropped segments generally cause TCP to lose its ACK-based clock, reducing overall throughput.

Selective Acknowledgment (SACK) is a strategy that corrects this behavior in the face of multiple dropped segments. With selective acknowledgment, the data receiver can inform the sender about all segments that arrived successfully. As a result, the sender needs to retransmit only the segments that were actually lost.

Refer to the following Web site for more information about TCP Selective Acknowledgments.

<http://www.ietf.org/rfc/rfc2018.txt>

Tip: We prefer equipment that supports SACK. If you already have a solution in place make sure it is enabled.

2.5.7 Path MTU and path MTU discovery

When one IP host has a large amount of data to send to another host, the data is transmitted as a series of IP datagrams. It is preferable that these datagrams be of the largest size that does not require fragmentation anywhere along the path from the source to the destination. This datagram size is referred to as the Path Maximum Transmission Unit (PMTU), and it is equal to the minimum of the MTUs of each hop in the path. Each media type has a default MTU that defines the maximum payload length for the media.

For example, the default MTU for Ethernet is 1500 bytes. You can override the default MTU for each interface when configuring the interface. Ethernet supports jumbo frames, which allow you to set the default frame size up to 9 KB. However, not all routers and switches support jumbo frames, and those that do may not have jumbo frame support configured on the appropriate interfaces.

A Fibre Channel frame is 2148 bytes, and for this reason, in a network with the MTU equal to 1500 bytes, it is necessary to break the frame into two segments. Segmenting a Fibre Channel frame and then reassembling it at the other end affects the maximum throughput performance of the Ethernet interface. An MTU of 2300 is preferred for transporting FCIP where sustained throughput is required. A smaller MTU, such as the default value of 1500, affects the throughput slightly as the device must segment and reassemble the Fibre Channel frames at each end of the tunnel.

PMTU Discovery tries to discover the largest MTU a path can support, by sending a sequence of frames with the *DF* (Don't Fragment) bit set in the IP header. If a frame is too big for an interface, an ICMP Can't Fragment message is

returned to the sender. This process is repeated until the largest MTU path supports without IP fragmentation.

Path MTU may not always work because it relies upon receiving ICMP Can't Fragment message. Layer 2 switches do not generate ICMP messages, and not all routers forward ICMP messages.

Refer to the following Web location for more information about Path MTU Discovery:

<http://www.ietf.org/rfc/rfc1191.txt>

Tip: Look for equipment that has PMTU discover. If you already have it make sure it is enabled.

Restriction: The MTU for the Gigabit Ethernet Interface supporting the FCIP tunnel must be configured to the lowest MTU of any interface in the tunnel path. Consult your network administration staff or your link provider.

2.5.8 Compression

Data compression reduces the size of data frames to be transmitted over a network link. Reducing the size of a frame reduces the time required to transmit the frame across the network. Data compression provides a coding scheme at each end of a transmission link that allows characters to be removed from the frames of data at the sending side of the link and replaced correctly at the receiving side. Because the condensed frames take up less bandwidth, we can transmit greater volumes at a time.

Compression is software or hardware implemented, although sometimes a device can have both. A software implementation of compression will use CPU cycles from the device responsible for transmitting data and also cycles from the device that is receiving the data. A hardware implementation saves extra cycles implementing compression in specialized pieces of hardware.

Data compression results depend on the kind of data being transmitted. If you are transmitting just 0s and 1s (zeroes and ones) you can get rates as good as 30:1. If you are transmitting encrypted data or random data, it is possible that you will not get any performance improvement. The typical compression data rate is 2:1.

As an example of compression, if you have a link with 100Mb/s and a compression rate of 2:1, then you can transmit 200Mb/s as opposed to the 100Mb/s that you can transmit without compression.

During implementation, test compression rates with representative data to get a realistic idea of the actual compression rate.

Tip: Compression generally doubles the performance.

2.5.9 Write Acceleration

Write Acceleration is a SCSI spoofing mechanism designed to improve application performance by reducing the overall service time for SCSI write input/output operations and replicated write I/Os over distance.

The performance improvement from Write Acceleration typically approaches 2:1 but depends upon the specific situation. In some situations with large write block sizes, the performance improvement can be more than 2:1 because multiple transfer readies per write are eliminated.

Write Acceleration reduces the number of FCIP WAN round trips per SCSI FCP Write I/O to one. Most SCSI FCP Write I/O exchanges consist of two or more round trips between the host initiator and the target SVC.

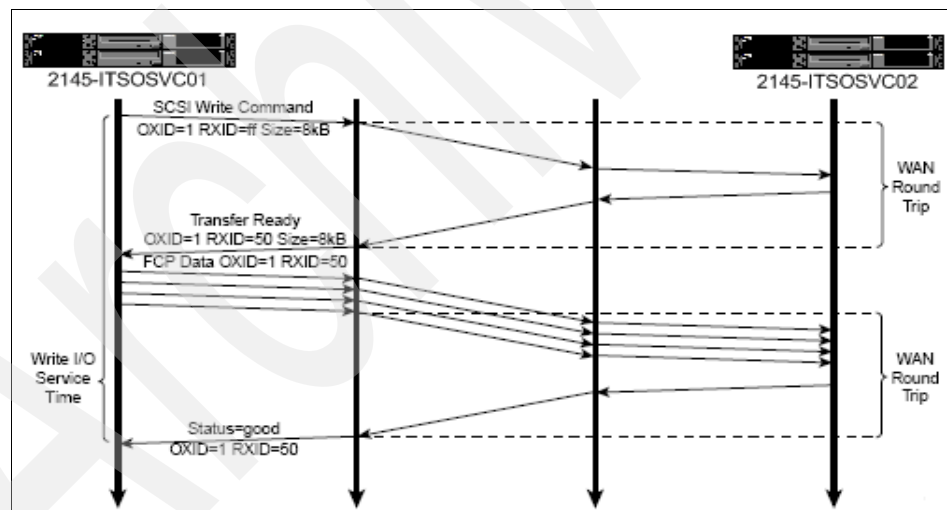


Figure 2-4 SCSI FCP Write without Write Acceleration

The protocol for a normal SCSI FCP Write without Write Acceleration (see Figure 2-4) is as follows:

1. Host initiator issues a SCSI Write command, which includes the total size of the Write (8 KB, in this example), an Origin Exchange Identifier (OXID), and a Receiver Exchange Identifier (RXID).

2. The target responds with an FCP Transfer Ready. This tells the initiator how much data the target is willing to receive in the next Write sequence and also tells the initiator the value the target assigned for the RXID (50, in this example).
3. The initiator sends FCP data frames up to the amount specified in the previous Transfer Ready.
4. The target responds with a SCSI Status=good frame if the I/O is completed successfully.

The protocol for Write Acceleration, shown in Figure 2-5, differs in the following manner:

1. After the initiator issues a SCSI FCP Write, a Transfer Ready is immediately returned to the initiator by the SAN switch. This Transfer Ready contains a locally allocated RXID.
2. At the remote end, the target, which has no knowledge of Write Acceleration, responds with a Transfer Ready. The RXID of this is retained in a local table.
3. When the FCP data frames arrive at the remote SAN switch from the initiator, the RXID values in each frame are replaced according to the local table.

The RXID for the SCSI Status=good frame is replaced at the local SAN switch with the value assigned in step 1.

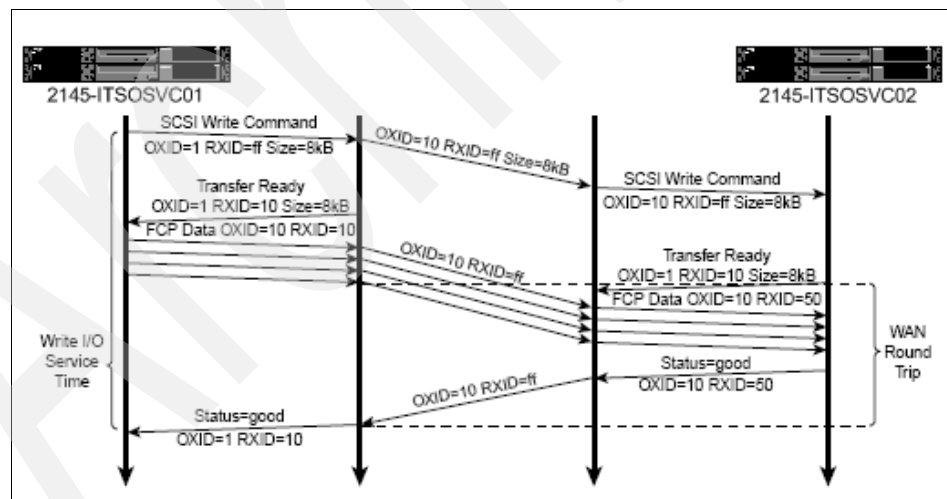


Figure 2-5 FCP SCSI operation with Write Acceleration

Write Acceleration can improve Metro Mirror performance because it requires two round trips to complete. Global Mirror code works differently, and it requires just one round trip to transfer data. Also write acceleration does not bring any

performance improvement. More details can be found in section 3.3.2, “Remote Copy layer” on page 70.

Tip: Write Acceleration can generally double the performance.

2.5.10 Encryption

When data is transmitted over distance on shared links we need to consider the data confidentiality and integrity. It is possible to encrypt data before transmission and decrypt it at the receiving side. Encryption can be performed in the SAN or on the network if you are using FCIP or IFCP. Encryption adds extra latency due to frame or packet manipulation, although this extra latency can be measured in microseconds.

As with compression, encryption can be implemented by software or by hardware. Software demands extra CPU cycles from the device processing Fibre Channel frames or IP packets. Hardware implementation relies on specialized hardware processing at wire speed.

2.5.11 Extended Fabrics and Routed Fabrics

Connecting existing SAN fabrics or extending a SAN fabric creates SAN design challenges. Basically, there are two ways to interconnect SVC clusters located at different sites:

- ▶ Extended Fabrics
- ▶ Routed Fabrics

Extended fabrics

Primary and secondary site fabrics can be merged in what we call an extended fabric. Fibre Channel control traffic flows between sites through interconnection technology, and any device can communicate with another one when adequately zoned. Fabrics can be extended up to 500 KM with long wave SFPs, DWDM, and dark fiber.

Extended fabrics have one main drawback, error propagation. After you link these SANs using an inter-switch link (ISL), you create essentially one large fabric. Any error that occurs on one side is transmitted to the other. This could be simple zoning changes, taking down one network for upgrade, and adding new equipment to a SAN. All these activities create state changes that are propagated across the IP link and introduces unwanted downtime and latency in applications.

Following are items that also need to be considered:

- ▶ Link oscillation that affects the entire network.
- ▶ Service disruptions from fabric reconfigurations and zone sets.
- ▶ Size of entire network is limited by FC fabric limits.
- ▶ Interoperability issues between multivendor fabrics.
- ▶ Reconfiguration may be required to connect previously separated fabrics.
- ▶ Principal switch selection over E_Port links.
- ▶ Larger SNS tables.
- ▶ Lengthy convergence time.
- ▶ Potential interoperability issues due to conflicting microcode versions.

Routed fabrics

In a SAN fabric that has implemented multiprotocol routing, data traffic is transported between specific initiators and targets on different fabrics without merging the fabrics into a single logical fabric. The traffic is checked to ensure that initiators do not access any resources across the fabrics other than the designated ones.

Routed fabrics minimize the impact of change in fabric services across geographically dispersed sites and limit fabric control traffic such as RSCNs and Build/Reconfigure Fabric.

Routed fabrics even allow the connection of fabrics that have the same domain ID.

Implementing a multiprotocol fabric requires multiprotocol routers, and also requires that the Inter VSAN Routing feature is using Cisco SAN switches.

Other benefits are interoperability, network integration, protocol conversion, isolation, network address translation, distance independence, and performance improvement.

Interoperability

Routed fabrics allow the connection of heterogeneous SAN islands by providing simultaneous support for multivendor implementations. They also provide investment protection for existing equipment along with the agility to adopt new technologies as they are required.

Network integration

Routed Fabrics unify the storage area network by joining SAN segments and consolidating access to resources, simplifying network design and reducing total cost of ownership.

Protocol conversion

SAN Routing liberates the SAN from the confines of a one-protocol solution, allowing storage networks to cost-effectively extend distances over the IP network using FCIP or iFCP. It may also support iSCSI integration.

Isolation

SAN routing enables storage area networks to scale beyond traditional limitations by providing essential border isolation for faults, broadcast storms, management and security, thereby increasing stability and network integrity.

Network address translation

An essential component of True SAN Routing, Network Address Translation, (NAT) maintains the autonomy of individual segments, shielding fabrics from addressing conflicts while enabling critical authorized information to pass through.

Distance independence

Unlike tunneling protocols that create one stretched fabric, SAN Routing preserves the autonomy of both local and remote SANs by providing zone and fault isolation over metro or wide area distances, enabling customers to deploy complex multi-site solutions while leveraging the cost benefits of available and affordable IP network services.

Performance improvement

Operating at wire speed, intelligent SAN routers offer performance enhancements over standard fabric implementations by providing extended buffers and efficiency algorithms such as compression, transmission aggregation, and bandwidth provisioning.

Note: We recommend a Routed fabric because of its stability.

2.6 Bandwidth sizing

When you design a storage replication solution across sites, sooner or later one fundamental question is asked: what bandwidth must be put in place to satisfy the remote write requests while maintaining adequate performance on the local production site? This section addresses that question.

2.6.1 Measuring workload I/O characteristics

To size the bandwidth required to replicate data between two different locations we need to know the amount of data that is written to disk: the write profile. We need to obtain this information for all the servers involved in replication and must add it together to get the I/O profile of multiple servers.

The write activity measurements must span a reasonable amount of time, days or weeks, and must be gathered in busy periods such as month or year end. A reasonable long sampling interval is five to 15 minutes, and a more detailed analysis can be performed for the peak hour.

There are many measures of I/O activity. For bandwidth sizing one measure is essential: the number of bytes per second that are written to disk. This can be calculated as the average I/O size multiplied by the number of I/Os.

2.6.2 Determining the bandwidth

In section 2.6.1, “Measuring workload I/O characteristics” on page 47, we discussed how to collect the performance data to determine write activity. These writes need to be replicated to the remote system.

Using I/O write profile to determine required bandwidth

If we have collected write activity from multiple systems and platforms we first need to summarize it to get a profile that aggregates write activity from the multiple systems. We need to sum the write performance data, interval by interval, for all the systems we want to replicate.

Collecting, consolidating, and generating reports from several servers brings challenges such as disparate platforms with different ranges of tools available, time synchronization between all servers (NTP can address this if in place), and how to consolidate all data from different tools and finally generate reports.

If you have storage from a single vendor, the task may be made easier by tools that are made available by the vendor. But there are other considerations that need to be taken into account. One of these is that data from the storage subsystem is replicated, and any performance information will not take the copies into account.

If you have more than one storage subsystem, you need to collect, consolidate, and generate reports from several subsystems, and this may involve different tools from different vendors.

It is easier if you have a tool in place is not platform specific and that collects, stores, and generates reports from hosts, switches, and even storage subsystems.

IBM Total Productivity Center (TPC) is a tool that can do this job. Data collection and performance analysis with TPC is covered in Chapter 4, “Performance considerations in SVC Business Continuity solutions” on page 119.

After we have the number of write operations and the average size of this operation, we can estimate the intersite link bandwidth. It is very important to have data collected from the busiest days from the business’ point-of-view. If we fail to estimate link bandwidth size correctly, application performance can be impacted.

Synchronous or asynchronous

For synchronous replication we need to determine the peak write rate and to choose a network bandwidth that is greater than this peak. The reason for this choice is not to impact the performance of the production environment's applications. If we choose anything less than the peak write rate, the primary application may incur delays due to writes at the remote site. A second factor to consider is additional latency for the I/O operations and this translates into higher response times.

Important: Synchronous replication requires link bandwidth greater than overall write peak IOs.

Asynchronous replication is more complex. In general it supports adequate operations with less bandwidth. The penalty is that RPO is greater than zero and depends on the bandwidth and the arrival of writes. Also, depending on the replication implementation, data writes may need to be stored in a temporary storage area or log and the amount of log space must be determined.

2.6.3 Rule of thumb

If no information is available we can give a first rough estimate of writes using rules of thumb based on industry averages. Note that your installation may be completely different.

Computer Measurement Group (<http://cmg.org>), is a worldwide association of IT system performance professionals. CMG user groups meet several times a year, worldwide, to exchange IT performance information. Over the past five years, public CMG papers document that a reasonable worldwide average disk performance measurement is every 1 GB of disk data that produces (on

average) a little less than 1 I/O per second per GB. This estimate is reasonable for both mainframe and open system environments. This metric is called *access density*.

We can use this access density metric to derive a straight forward rule of thumb (ROT) for estimating the amount of production I/Os that are generated by a given amount of disk data. Using the access density above we can estimate the amount of write data per second that a given amount of data produces.

There are two rule of thumb numbers based on using a conservative average access density of 1 I/O per second per GB:

- ▶ One TB of production OLTP disk storage generates about 1 MBps of write data.
- ▶ One TB of production BATCH disk storage generates about 6.75 MBps of write data.

The calculations involved in deriving these two estimates are outside the scope of this discussion.

Worldwide, CMG white papers have shown that these estimates are valid. Using this information, we can calculate and extrapolate the number of transactions or batch updates that might be represented by a given amount of production disk mirrored storage.

SAN Volume Controller Mirroring Solutions

SAN Volume Controller (SVC) offers the following three Copy Services:

- ▶ FlashCopy
FlashCopy creates virtual disk (VDisks) point-in-time copies and gives the application continuous operation during backups.
- ▶ Metro Mirror
Metro Mirror is a synchronous Remote Copy function of SVC and is suitable for disaster recovery solutions for limited distances. SVC Metro Mirror keeps consistent and current images of primary VDisks.
- ▶ Global Mirror
For long distances, SVC Global Mirror offers an asynchronous copy function. SVC Global Mirror keeps consistent images of primary VDisks at all times, even when SVC clusters are separated by long distances.

In this chapter we focus on SVC Remote Copy Services because data replication plays an integral part in any Business Continuity solution.

This chapter is logically divided into theoretical and practical sections.

In the first sections we review basic SVC concepts and introduce new concepts of Remote Copy Services. In later sections we discuss how Remote Copy Services work.

Finally, we will present a real-life implementation of SVC Remote Copy service.

Note: In this chapter we use the term Remote Copy when the subject matter applies to both Metro Mirror and Global Mirror. Where necessary, we make the distinction between Metro Mirror and Global Mirror.

3.1 Function

SVC Remote Copy Services mirror data from one SVC cluster to another SVC cluster.

Mirroring data is actually replicating data from one SVC cluster to another. The objective is to have a consistent image of the production systems through out.

Copy operations can be implemented in two different ways:

- ▶ Synchronous

Synchronous data mirroring requires that a strict write order sequence be followed. Each update to the primary subsystems must also be updated in the secondary subsystems before another transaction can process. This results in near perfect data currency but can result in some lag time or latency between transactions.

- ▶ Asynchronous

Asynchronous data mirroring allows additional transactions to process without ensuring receipt in the secondary subsystems. The lag in data currency could be seconds, minutes, or hours depending on the specific technology used, but there is usually low or no impact on transaction processing.

Synchronous data replication is implemented by Metro Mirror while asynchronous data replication is implemented by Global Mirror.

Besides guaranteeing a consistent copy of production data at all times, SVC Remote Copy Service allows SVC VDisk's role to change and for data to flow from secondary to primary after the flow switches.

The source of data copy is always called primary, and the target of the copy is called secondary. Role changes mean that with specific commands we can switch VDisk roles from primary to secondary, and then back again. An example is a switch from the primary site to the secondary site for a drill or an operational

procedure. In this situation, secondary VDisks enable write operation for the mapped hosts.

Remote Copy operations are manipulated in pairs of VDisks. Pairs of VDisks using mirrored data from one SVC cluster to another are called *relationships*.

Remote Copy Services are manipulated through relationships. We can create, delete, start, and stop a relationship. When a relationship is started, data is copied from primary VDisks to secondary VDisks.

3.2 Concepts of operation

In this section we introduce basic concepts and terms indispensable to work with SVC Remote Copy Services. Section 3.3, “Remote Copy internals” on page 62 discusses SVC Remote Copy Services in detail.

3.2.1 SVC objects

Remote Copy Services commands have an argument *relationship*. Relationships are composed by pairs of VDisks. VDisks are a set of extents from a specific MDisk group. Managed disk groups are formed by MDisk. Managed disks are LUNs from raid controllers presented to the SVC cluster.

This section discusses SVC physical and logical objects and how they are correlated.

LUN

A *LUN* is a unique identifier used on a SCSI bus that enables it to differentiate between other devices called LUNs. We refer to it as the unit of storage that is defined in a storage subsystem such as an IBM System Storage Enterprise Storage Server® (ESS), IBM System Storage DS4000, DS6000™, and DS8000™ series Storage Server, or storage servers from other vendors.

Managed disk

A *managed disk* (MDisk) is a LUN presented by a RAID controller and managed by the SAN Volume Controller. The MDisk must not be configured so that it is visible to host systems on the SAN.

Managed disk group

The *managed disk group* (MDG) is a collection of MDisk that jointly contain all the data for a specified set of VDisks.

Extents

MDisks are chopped in small units called *extents*. Extent size ranges from 16 MB to 512 MB. Each MDG has one unique extent size and cannot be changed after being created.

Virtual disk

A *virtual disk* (VDisk) is a SAN Volume Controller device that appears to host systems attached to the SAN as a SCSI disk.

Grains

When data is copied from the source VDisk to the target VDisk, it is copied in units of address space known as *grains*. In the SVC, the grain size is 256 KB.

Tracks

Remote Copy controls read and write access implementing lock/unlock in tracks of VDisks. The track size is 32 KB.

Relationship

When creating a Remote Copy relationship, initially the master VDisk is assigned as the primary and the auxiliary VDisk the secondary. This implies that the initial copy direction is mirroring the master VDisk to the auxiliary VDisk. After the initial synchronization is complete, the copy direction can be changed if appropriate.

A relationship must be either Metro Mirror or Global Mirror. Metro Mirror is the default type.

Consistency Groups

Relationships can be grouped in Consistency Groups. After it is grouped in a Consistency Group, relationships are manipulated uniformly using the Consistency Group.

All the relationships in a Consistency Group must be either Metro Mirror or Global Mirror Relationships. A Consistency Group cannot contain a mixture of Relationship types.

The definition of a Consistency Group is set when the first relationship is added to the group. That is, the group automatically assigns the 'type' of that relationship, where type is Metro or Global.

3.2.2 Summary of data consistency

Remote Copy Services must guarantee data consistency at all times. Data consistency necessity drives how read/write operations occur in different situations. The key factors of data consistency are read stability and write ordering.

Data consistency

Data consistency in a remote copy solution means that a secondary VDisk at the recovery point contains the same data as the primary VDisk.

Consistent data

The consistent or inconsistent property describes the relationship of the data on the secondary to that on the primary VDisk. It can be considered a property of the secondary VDisk itself.

For example, a secondary VDisk is described as consistent if it contains data that could have been read by a host system from the primary, if power 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*.

From the point-of-view of an application, consistency means that a secondary VDisk contains the same data as the primary VDisk at the recovery point (the time at which the imaginary power failure occurred).

Read stability (mirror write consistency)

Read stability is a key concept of data consistency. It means that reads from the same sector of primary and secondary VDIs at a particular time must bring the same information, and if not, writes occurs between the reads.

I/O ordering (write ordering) and dependent writes

Many applications that use block storage have a requirement to survive failures such as losses of power and software crashes, and not lose data that existed prior to the failure. Since many applications need to perform a large number of update operations in parallel to that of 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 is designed with the concept of dependent writes. These are writes where it is important to ensure that an earlier write completed before a later write is started. Reversing the order of dependent writes can undermine the application algorithms and lead to problems such as detected, or undetected, data corruption.

Sequence number

Write operations for VDisks that belong to Global Mirror relationships are ordered at the primary SVC cluster. Each parallel operation receives a number called *sequence*, and it can contain more than one write operation with the condition that the write is to a different grain.

Consistency Groups

Certain uses of Remote Copy require the manipulation of more than one relationship. Metro Mirror or Global Mirror Consistency Groups provide 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 copied VDisks because the applications have related data that 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.

Remote Copy commands can be issued to a Consistency Group that affects all relationships in the Consistency Group, or to a single relationship if not part of a Consistency Group.

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.

Although it is possible that Consistency Groups can be used to manipulate sets of relationships that do not need 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; however, this may not be the case if the relationship is 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, Remote Copy 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, Remote Copy still refuses to grant access to its secondary, even though this is safe in this case because the Remote Copy policy is to refuse access to the entire Consistency Group if any part of it is inconsistent.

3.2.3 Modalities of Remote Copy

SVC has two Remote Copy modalities:

- ▶ Synchronous
- ▶ Asynchronous

Synchronous Remote Copy

Synchronous Remote Copy ensures that updates are committed at both primary and secondary before the application gives a completion update. This ensures that the secondary is fully up-to-date.

However, this means that the application is fully exposed to the latency and bandwidth limitations of the communication link to the secondary. Where this is truly remote this extra latency can have significant adverse affects on application performance.

1. Write to source SVC cluster cache.
2. Write to target SVC cluster cache.
3. Signal write complete from the remote SVC cluster.
4. Post I/O complete to host server.

Metro Mirror implements synchronous data replication for SVC.

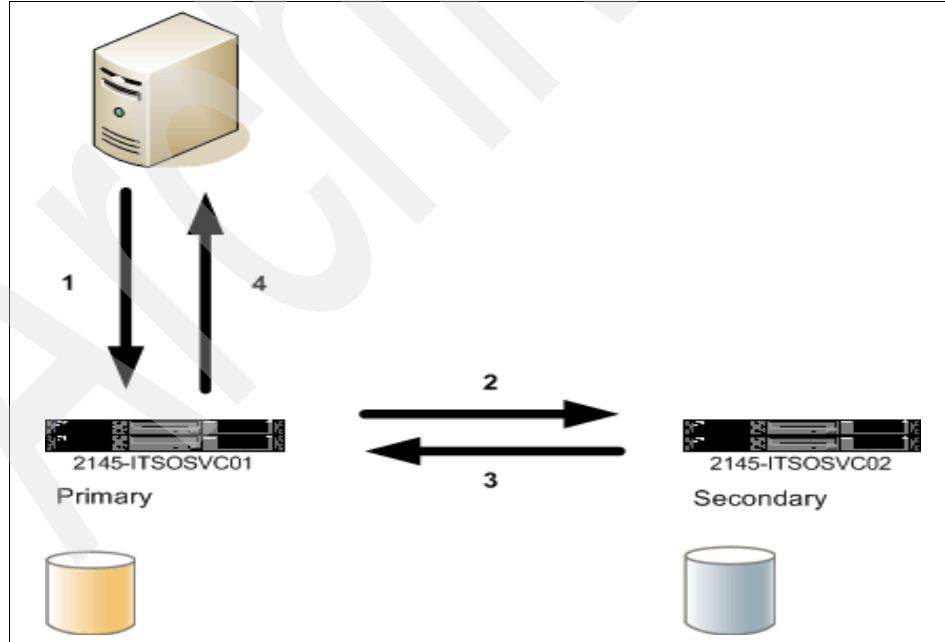


Figure 3-1 I/O write operation for Metro Mirror Remote Copy Service

Asynchronous Remote Copy

In Asynchronous Remote Copy the application is given completion to an update before that update is committed at the secondary. Hence, on a failover, some updates might be missing at the secondary. The application must have some external mechanism for recovering the missing updates and reapplying them. This mechanism may involve user intervention.

Asynchronous Remote Copy provides comparable functionality to a continuous backup process that is missing the last few updates. Recovery on the secondary site involves bringing up the application on this recent backup and then re-applying the most recent updates to bring the secondary up to date.

1. The host server makes a write I/O to the source (local) SVC.
2. The write returns as completed to the host server's application.
3. Later, in a non-synchronous manner, the source SVC sends the necessary data so that the updates are reflected on the target (remote) SVC.
4. Write operation completion is given to the source SVC.

SVC Remote Copy Service implements asynchronous data replication using the Global Mirror function.

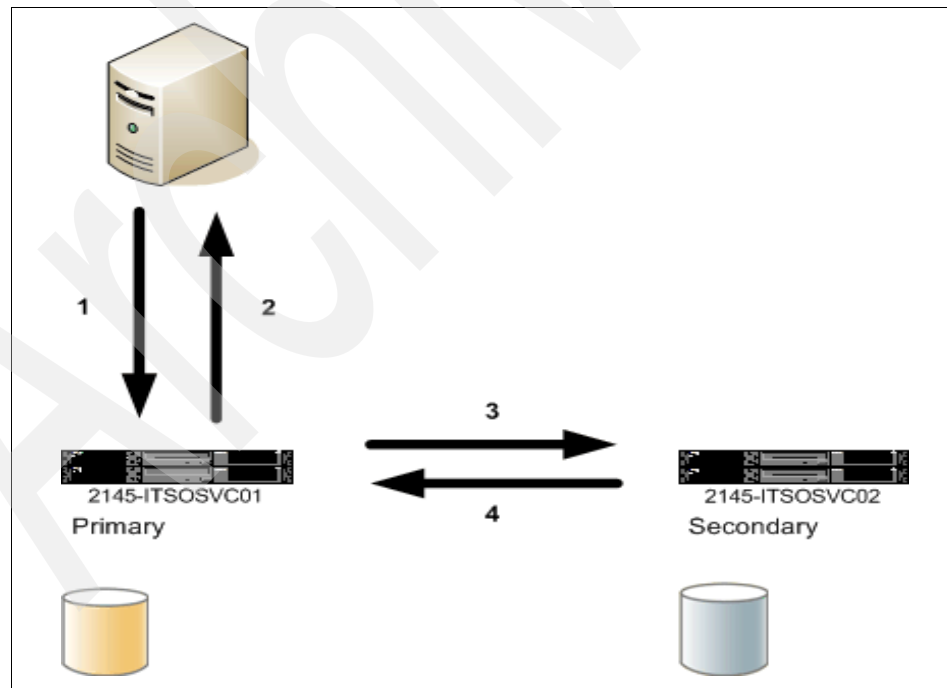


Figure 3-2 Data communication for Global Mirror Remote Copy Service

From the previous procedure, note how in an asynchronous type technique, the data transmission and the I/O completion acknowledge are independent processes. This results in virtually no application I/O impact, or at most to a minimal degree only. This is convenient when required to replicate over long distances.

3.2.4 Summary of relationships and Consistency Group states

In this section we explain the different states of a Remote Copy relationship. The meaning of each state is important to the Remote Copy process.

Relationships and Consistency Group states

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*.

Figure 3-3, shows an overview of the states that apply to a Remote Copy relationship in the connected state.

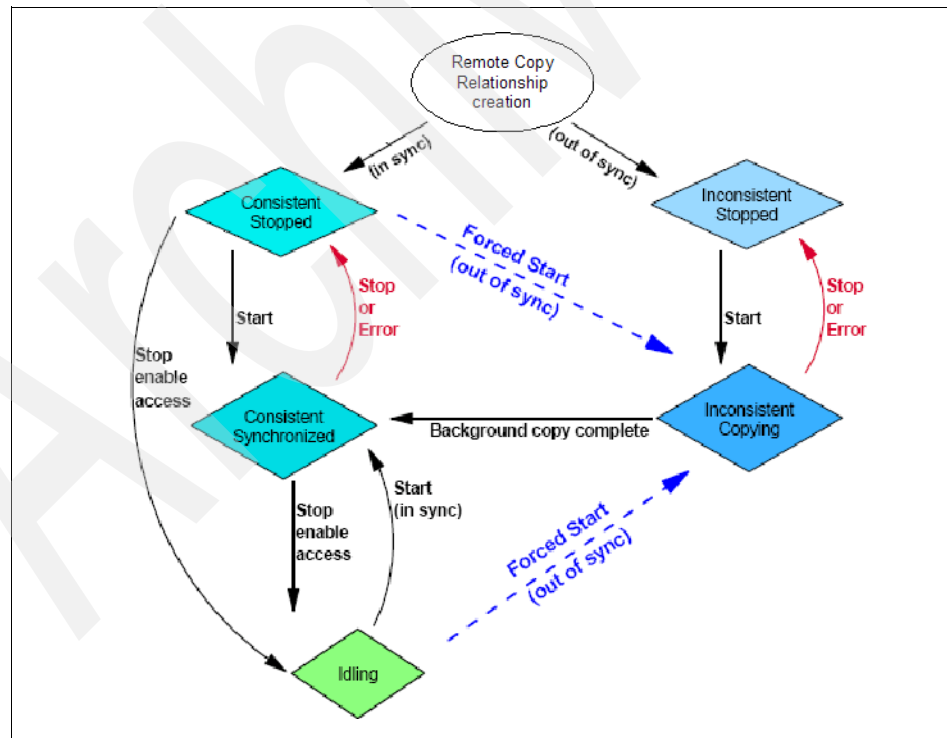


Figure 3-3 Relationship states

When creating the Remote Copy relationship, you can specify if the auxiliary VDisk is already in sync with the master VDisk and if the background copy process is skipped. This is especially useful when creating Remote Copy relationships for fresh new VDisks.

The states showed in Figure 3-3 on page 59 are described in detail in the following subsections.

InconsistentStopped

This is a connected state. In this state, the primary is accessible for read and write I/O but the secondary is not accessible for either. A copy process needs to be started to make the secondary consistent.

InconsistentCopying

This is a connected state. In this state, the primary is accessible for read and write I/O, but the secondary is not accessible for either read or write I/O.

A background copy process runs, which copies data from the primary to the secondary VDisk.

If the background copy process completes a stand-alone relationship, or on all relationships for a Consistency Group, the relationship or Consistency Group transitions to *ConsistentSynchronized*.

ConsistentStopped

This is a connected state. In this state, the secondary contains a consistent image, but it might be out-of-date with respect to the primary.

ConsistentSynchronized

This is a connected state. In this state, the primary VDisk is accessible for read and write I/O. The secondary VDisk is accessible for read-only I/O. Writes that are sent to the primary VDisk are sent to both primary and secondary VDisks.

Either good completion must be received for both writes, the write must be failed to the host, or a state transition out of ConsistentSynchronized must take place before a write is completed to the host.

Idling

This is a connected state. Both master and auxiliary disks are operating in the primary role. Consequently both are accessible for write I/O.

IdlingDisconnected

This is a disconnected state. The VDisk or disks in this half of the relationship or Consistency Group are all in the primary role and accept read or write I/O.

The main priority in this state is to recover the link and make the relationship or Consistency Group connected once more.

InconsistentDisconnected

This is a disconnected state. Secondary VDisks are disabled for read/write operations.

ConsistentDisconnected

This is a disconnected state. Secondary VDisks can accept read I/O but not write I/O.

Empty

This state only applies to Consistency Groups. It is the state of a Consistency Group that has no relationship and no other state information to show.

It is entered when a Consistency Group is first created. It exits when the first relationship is added to the Consistency Group at which point the state of the relationship becomes the state of the Consistency Group.

A deep discussion about relationship or Consistency Group states are out of the scope of this IBM Redbooks publication. It is well covered on *IBM System Storage SAN Volume Controller*, SG24-6423.

3.2.5 Remote Copy implementation differences

IBM System Storage storage array controllers such as DS4000, DS6000, and DS8000 have Copy Services similar to the SVC. Copy Services are named in the same way through System Storage products: FlashCopy for point-in-time copies, Metro Mirror for synchronous data replication, and Global Mirror for asynchronous data replication.

These services have the same purpose, but they have different code and are implemented differently. For example, Global Mirror for DS8000 implements asynchronous replication followed by a point-in-time copy where SVC Global Mirror implements just the asynchronous replication portion. There are other implementation details and you should consider it when comparing different products Copy Services.

You can find more details about SVC Remote Copy Services in 3.3, “Remote Copy internals” on page 62.

Refer to the following IBM Redbooks publications for more details about DS4000 Copy Services:

- ▶ *IBM System Storage DS4000 Series and Storage Manager*, SG24-7010
- ▶ *DS6000 Copy Services on IBM System Storage DS6000 Series: Copy Services in Open Environments*, SG24-6783
- ▶ *DS8000 Copy Services on IBM System Storage DS8000 Series: Copy Services in Open Environments*, SG24-6788

SVC supports intracluster, which allows the testing of Remote Copy Services inside a single SVC cluster. With intracluster we can assess application impact due to Metro Mirror or Global Mirror code, and even distance delay, to simulate that in the real implementation SVC clusters are located in different locales. We can simulate mirroring data from one VDisk composed by MDisks from a storage controller located at a primary site to a second VDisk with MDisks from a storage controller located at a secondary site, but connected to the same SVC cluster. SVC intracluster facilities are used to allow this scenario. This scenario works, but it is not recommended because if the primary site SVC cluster is unavailable, secondary VDIs are not accessible.

Note: Intracluster is not discussed in this book because it does not represent a recommended BC solution.

3.3 Remote Copy internals

This section discusses some aspects of the SVC and SVC Remote Copy Services operations, including the functionality of SVC Remote Copy.

3.3.1 Intercluster communication

Each SVC cluster can be configured to recognize a single other cluster as its partner. When both clusters are correctly configured and can communicate with each other over the fabric, they establish further communication facilities between the nodes on each of the clusters. This comprises of the following:

- ▶ A single control channel that exchanges and coordinates configuration information. This is called *intercluster link*.
- ▶ I/O channels between each of the nodes in the clusters.

These channels are maintained and updated as nodes appear and disappear, and as links fail and are repaired to maintain continuous operation where possible.

Remote Copy relies on a pre-configured remote cluster partnership to work and the whole communication occurs over Fibre Channel logins.

Fabric

All intercluster communication is performed through the SAN. Prior to creating intercluster Remote Copy relationships, you must create a partnership between the two clusters. All SVC node ports on each SVC cluster must be able to access each other to facilitate the partnership creation.

We use an example to show how SVC is supposed to be connected for high availability, performance, and Business Continuity. The same scenario is implemented in section 3.7, “Scenario implementation” on page 85.

We have two sites, and at each one we have two-node SVC clusters, two SAN switches, one storage array controller, and one “intersite communication device”.

Each SVC node has four ports named ports 1, 2, 3, and 4.

For example, taking just one node we connect both HBAs port 1 to switch SAN01 and both HBAs port 2 to switch SAN03.

Best practice says to use the same module/port in the switch side for both fabrics. In our example, HBA port 1 is connected on module 1 port 1 on SAN01 and HBA port 3 is connected on module 1 port 1 on SAN03.

Figure 3-4 on page 64 shows physical connections from SVC01 cluster to local fabrics:

1. SVC01 node 1 has two connections to the switch SAN01 and two connections to the switch SAN03.
2. SVC01 node 2 has two connections to the switch SAN01 and two connections to the switch SAN03.
3. Switch SAN01 has SVC01 node 1 HBA port 1 connected to module 1 port 1 and SVC01 node 1 HBA port 3 connected to module 2 port 1.
4. Switch SAN03 has SVC01 node 1 HBA port 2 connected to module 1 port 1 and SVC01 node 1 HBA port 4 connected to module 2 port 1.

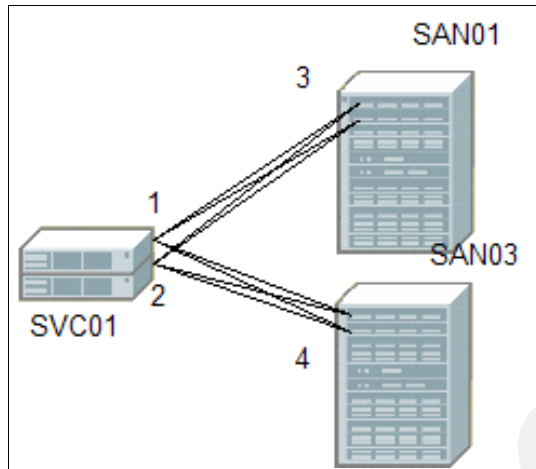


Figure 3-4 Node connections to SAN switches

Table 3-1 summarizes the physical connections from both clusters — the primary and the secondary. The primary nodes are node 1 and node 2, and the secondary nodes are node 3 and node 4.

Table 3-1 SAN connections summary (SAN03 and SAN04 are not shown)

	SVC01Node 1	SVC01Node 2	SVC02Node 1	SVC02Node 2
HBA port 1	SAN01 module 1 port 1	SAN01 module 1 port 2	SAN02 module 1 port 1	SAN02 module 1 port 2
HBA port 2	SAN03 module 1 port 1	SAN03 module 1 port 2	SAN04 module 1 port 1	SAN04 module 1 port 2
HBA port 3	SAN01 module 2 port 1	SAN01 module 2 port 2	SAN02 module 2 port 1	SAN02 module 2 port 2
HBA port 4	SAN03 module 2 port 1	SAN03 module 2 port 2	SAN04 module 2 port 1	SAN04 module 2 port 2

After the physical connections are defined, we need to create the logical connections.

Zones

There are different zones we need to put in place to allow the correct and supported operation of an SVC. Zones apply for local and for remote fabrics.

SVC Zones are zones that contain only SVC node ports. It is used for node communication. In our case we have Fabric01 (switch SAN01) and Fabric02 (switch SAN03).

Front End Zones (Host zones) are zones that contain SVC ports and host ports. It allows hosts to login into SVC and have VDisks mapped to it.

Back End Zones (Storage Zones) are zones that contain SVC ports and storage array controller ports.

Intercluster Zones are zones that contain ports from different SVC clusters. Intercluster zones are explained in “Intersite communication” on page 65.

Intersite communication

Intersite communication is discussed in great detail in section 2.5, “Distance factors for intersite communication” on page 34. Intersite communication devices allow local and remote fabric devices to communicate. Figure 3-5 on page 66 shows local and remote site connections through L1 and L2 links. SAN01 communicates with SAN02 using R01, L1 and R02. R01, L1, and R02 are our intersite communication devices. We have the same configuration for our counterpart fabric. SAN03 communicates with SAN04 through R03, L2, and R04.

It is possible to create zones in Fabric01 with devices logged into SAN02. These are intercluster zones. In our case, we have intercluster zones for SVC cluster nodes with ports located on SAN01 and SAN02 and for SVC cluster nodes with ports located on SAN03 and SAN04.

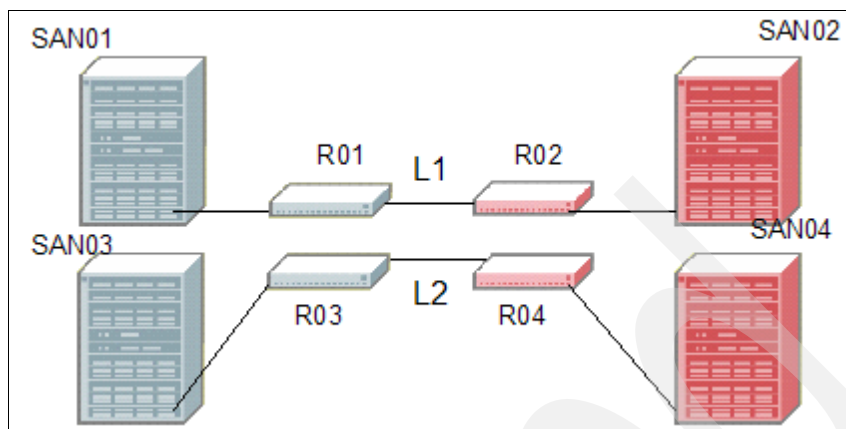


Figure 3-5 Local and remote site connections through L1 and L2 links

I/O channels

Intercluster zones define I/O channels. An I/O channel is the term used to describe the connections between nodes of SVC clusters that deliver mirror write I/Os between the clusters for intercluster Remote Copy. I/O channels are created directly by local and remote fabric arrangement.

Figure 3-9 on page 73 shows two SVC clusters connected through Fabric01 and Fabric02. SVC01 node 1 and SVC02 node 1 have the following four connections to each fabric:

- ▶ HBA port 1 connected to Fabric01
- ▶ HBA port 2 connected to Fabric02
- ▶ HBA port 3 connected to Fabric01
- ▶ HBA port 4 connected to Fabric02

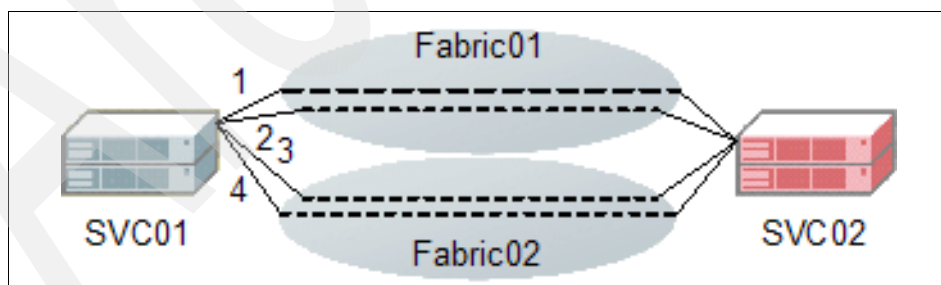


Figure 3-6 SVC01 Node 1 I/O channels

Intercluster link

On each cluster, one node is selected to control communications between the local and remote cluster. This node is called the *focal point*. Communication

between focal points flow between one of the I/O channels. This particular I/O channel receives the name of intercluster link.

The *intercluster link* carries the traffic between the focal points. Focal point traffic is distributed among the other nodes that belong to the cluster. Figure 3-7 on page 68 shows Intercluster link parameters.

The *intercluster link bandwidth* defines the allowed amount of link that can be used by the background copy processes. This governs the rate at which background copy is performed. The default value is 50 MBps, but it can be changed. Host write I/O is not subject to this same quota. Remember that at the time of this writing, we have SVC nodes equipped with 400 MBps HBAs.

If the background copy bandwidth is set too high for the link capacity, the background copy I/Os can back up on the link and delay synchronous secondary writes and lead to an increase in foreground I/O latency.

If the background copy bandwidth is set too high for the storage at the primary site, background copy read I/Os overload the primary storage and delay foreground I/Os.

If the background copy bandwidth is set too high for the storage at the secondary site, background copy writes at the secondary storage and again delays the synchronous secondary writes of foreground I/Os.

Important: To set the background copy bandwidth optimally, make sure that you consider all three resources (the primary storage, the intercluster link bandwidth, and the secondary storage).

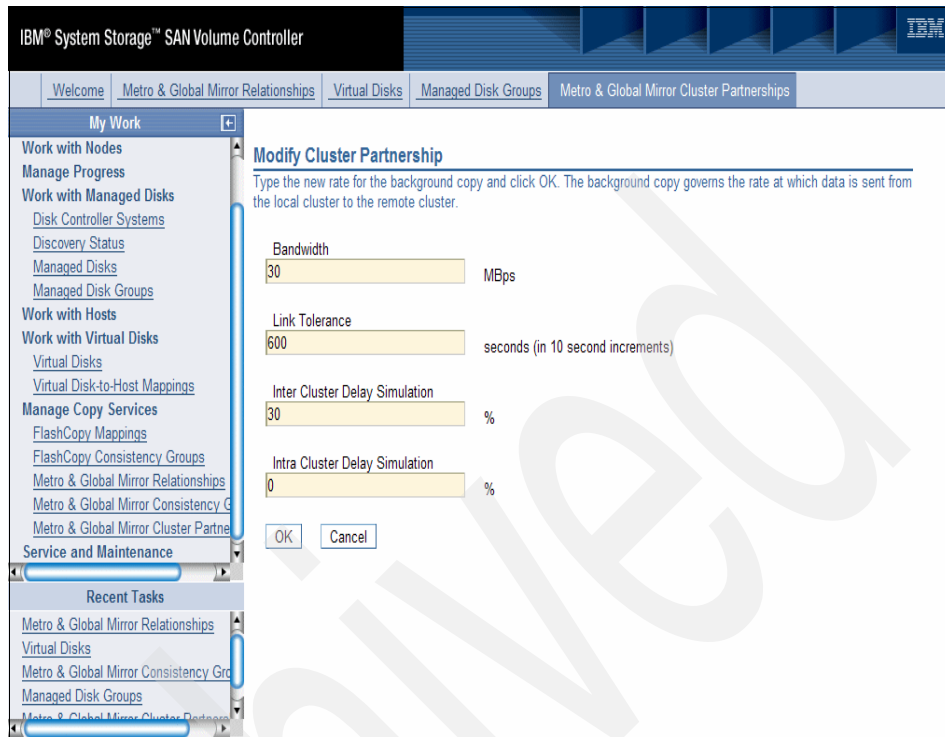


Figure 3-7 Intercluster link parameters

SVC Remote Copy Services implementation avoids an adverse effect on system behavior at the primary and to accomplish that it can postpone copy operations. A whole grain of tracks on one VDisk is processed at around the same time but not as a single I/O. Multiple grains are copied simultaneously, enough to satisfy the requested rate, unless the resources available cannot sustain the requested rate.

Background copy proceeds from low LBA to high LBA in sequence. This is to avoid conflicting with FlashCopy, which operates in the opposite direction. It is expected that it does not conflict with sequential applications since it tends to vary its disks more often.

No particular algorithm is used to attempt to load balance traffic between nodes. Targets for I/O are chosen based on factors such as order of appearance. It is anticipated that this is not an issue since the throughput of the intercluster connection is expected to be much less than that of any single node.

Delays imposed due to the intersite distance have a direct impact on application performance in a Metro Mirror implementation. This time is added to the disk

write response and it is limited to 20 ms round trips at the time of this writing. Check section 3.5, “Remote mirroring limitations” on page 80 for more details and information.

For Global Mirror relationships, the maximum supported round-trip time is 80 ms. Global Mirror relationships write response time delay does not directly impact the application performance, although it does impact it indirectly. Read section 3.5, “Remote mirroring limitations” on page 80 for more details and information.

Link tolerance specifies the maximum period of time that the system tolerates delay before stopping Global Mirror relationships. Specified values are between 60 and 86400 in increments of 10 seconds. The default value is 300.

The software that maintains the I/O channels monitors errors that are experienced on the links and deliberately excludes that link if the rate of error exceeds a certain threshold. Following are the principle classes of error that lead to an excluded link:

- ▶ An intermittent hardware problem on a non-redundant link that leads to a link reset for the SVC cluster.
- ▶ Lost frames or delay exceeding 1.5 seconds affecting the heartbeats between the two clusters.
- ▶ A problem leading to excessive data frames being dropped (even where non-data frames are passed successfully).
- ▶ A series of failures that mean that progress is not made, such that a message is not delivered for more than 15 seconds.

If the clusters are connected through a non-redundant link, the errors lead to a heartbeat timeout being detected by one cluster and that cluster then closes all process logins between the two focal point nodes. Such events are monitored versus the use of a threshold. Events that happen after 30 seconds of the first event are ignored, so that a large but short term disruption is tolerated. If within 10 minutes, errors occur more than every 30 seconds then the link is excluded.

When a link is excluded the focal point node deliberately prevents an intercluster link from being created between the local cluster and the remote cluster. The cluster logs an error. The other clusters report partially configured in its configuration state for the remote cluster. It is likely that the other cluster also detected the same set of errors and excluded the link on the same basis, in which case both clusters report as partially configured. If only one cluster detects an error and excludes the link, then that cluster continues to report as fully configured.

We can use the link tolerance during a link maintenance to avoid what is described above.

Inter Cluster Delay Simulation applies to Global Mirror and permits a delay simulation to be applied on writes sent to Secondary VDisks. This permits you to test performance implications before deploying Global Mirror and obtaining a long distance link. Specify a value from 0 to 100 in 10 second increments. The default value is 0 and disables this feature completely.

3.3.2 Remote Copy layer

The main I/O path inside SVC for Remote Copy purposes is contained in the following software modules:

- ▶ SCSI front end
- ▶ Remote Copy
- ▶ Cache
- ▶ FlashCopy
- ▶ Virtualization
- ▶ SCSI back end

A single host I/O operation received from the SCSI front end navigates the layers until it reaches the SCSI back end, if necessary. There are several questions in the path. Is it a read or a write? Is there a Remote Copy relationship for the target VDisks? Is the data already on cache? Is there a FlashCopy going on? Each layer is in charge of a specific function and after some processing is done, this specific layer transmits the necessary operations to the layer underneath it or to a remote SVC cluster.

Figure 3-9 on page 73 shows the major modules where the I/O from the hosts to the storage controllers flow.

We briefly describe each major component, but our focus is on the Remote Copy layer.

SCSI front end

I/O is received from hosts in the SCSI front end. LUN mapping is implemented in this layer. Host SCSI requests are translated into SVC format and transmitted to the next layer.

Cache

The cache may process the I/O as a cache hit in which case data is transferred to or from the host until completion.

The cache sends I/O to the FlashCopy to process Stage and Destage Requests. A *Stage Request* is simply a read and a *Destage Request* is simply a write. The operations are renamed at this layer to indicate that they are often performed asynchronously and quite separately from host I/O.

FlashCopy

The FlashCopy layer, when in use, is responsible for implementing point-in-time copies of production VDisks.

Storage Virtualization

Storage Virtualization is responsible for mapping VDisks to MDisks.

SCSI back end

The SCSI back end sits at the bottom of the Filter stack and processes requests to Managed Disks that are sent to it by the Virtualization Layer above. The SCSI back end is responsible for addressing SCSI commands to RAID controllers out on the SAN.

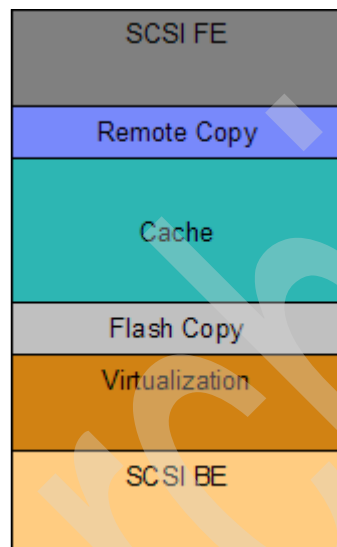


Figure 3-8 SVC I/O path

Remote Copy

When Remote Copy is active it communicates with another SVC node to implement Remote Copy functions. This communication happens through the intercluster link.

Remote Copy is an in-band copy solution, and this means it intercepts read/write operations before read and writes hit cache. That is why it is positioned right above the cache and below the SCSI Front End (Figure 3-9 on page 73). It also receives read I/Os before the cache and when necessary it delays them as part of lock/unlock mechanisms that guarantee data consistency.

VDisks that are part of a relationship are decomposed in two different ways for different purposes: tracks for data consistency control and grains for Remote Copy. The size of a track is 32 KB and a grain has a size of 256 KB.

When data is copied from the source VDisk to the target VDisk, it is copied in units of *grains*. Bitmaps contain one bit for each grain. The bit records whether the associated grain was split yet by copying the grain from the source to the target. The rate at which the grains are copied across the source VDisk to the target VDisk is called the *copy rate*. By default, the copy rate is 50 percent, though this can be altered.

Metro Mirror

Write operation I/Os are delivered from the primary SVC cluster back-end interface to the front-end interface of the secondary cluster, using the same I/O protocols as are used to deliver write I/O to back-end controllers.

The SCSI back-end round-robins I/O between the logins available to it in order to deliver I/O to the remote nodes in the remote cluster's I/O group.

Because writing at secondary nodes happens at almost the same time as the host write operation, two operations are necessary. First the data travels and is written on the secondary SVC cluster. The second operation is a write completion acknowledge to the primary SVC cluster.

Figure 3-10 on page 74 shows interacting layers between a primary SVC cluster and a secondary SVC cluster. SVC01 acts as a host writing to the SVC02 cluster.

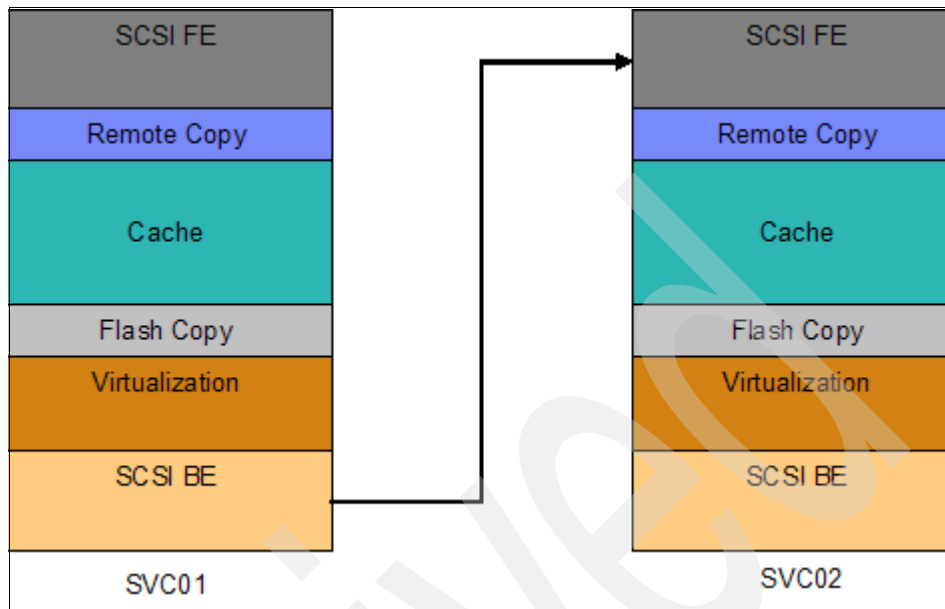


Figure 3-9 Remote Copying with Metro Mirror write operations

In a write operation, a request for exclusive access to the track is sent to the node in charge of a specific VDisk before a write takes place. This node is called the *owner*. The owner ensures that only one node is granted the track lock at any time. This means that Remote Copy might hold off I/Os that may otherwise be handled in parallel by cache. Obtaining the write lock is performed concurrently with fetching the data for the write I/O from the host.

Having obtained the write lock, a record of the write lock is made in non-volatile memory on both the owner node and the partner node.

Having secured the non-volatile record in two nodes, the write I/O is submitted, in parallel, to the cache and to the secondary copy of the Remote Copy operation.

When both the cache and remote write are completed, the non-volatile record is removed locally, and a message is sent to the owner node to remove it from there.

Special consideration is needed when a background copy is being performed. If a write I/O takes place to a track that has not yet been copied, then the track lock is still obtained, but the Remote Copy write is not performed. If a background copy operation is scheduled for a track that has a write I/O, then the background copy I/O is delayed until the track lock is released. If a track lock for a write I/O is

requested while a background copy operation is in progress, the track lock is delayed.

Global Mirror

Remote Copy asynchronous implementation relies on the same protocols used to implement SVC clustering to communicate with nodes on the remote SVC cluster.

Control operations are exchanged between pre-elected nodes and data flows through peer-to-peer communication. These require only one round-trip to deliver a write and receive its completion.

Data is moved and controlled by the Remote Copy layer. See Figure 3-10.

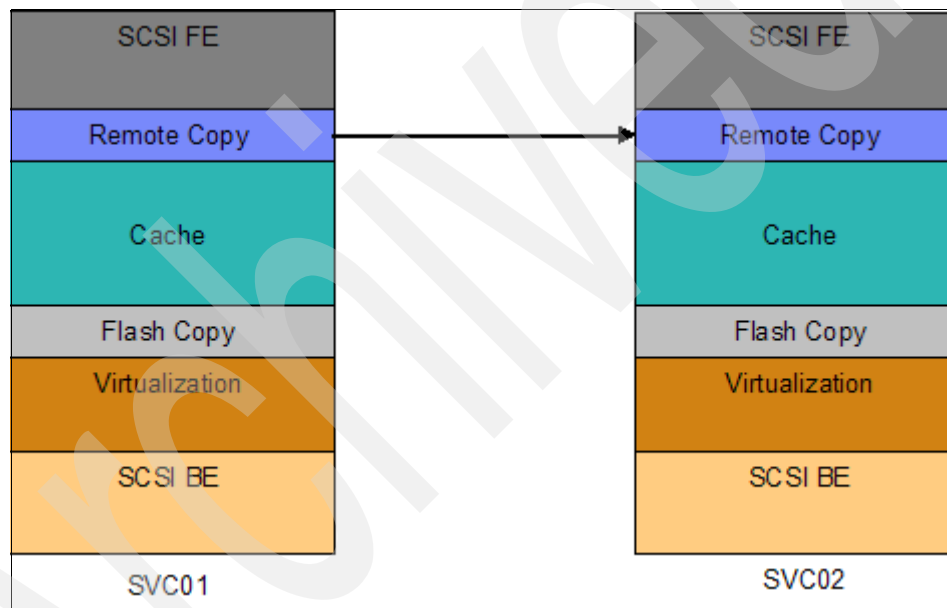


Figure 3-10 Remote Copying with Global Mirror write operations

Concurrent write operations receive a sequence number. Operations for different tracks can share the same sequence number.

Data is committed to primary storage using the normal cache algorithm. A log of I/O and its sequence number is committed on both nodes of the I/O group, while the write operation is transmitted to the remote SVC cluster.

At the secondary SVC cluster, write operations are applied following strictly the write ordering from the primary SVC cluster.

I/Os that overlap already active I/Os are serialized. This is called *colliding writes*, and in normal workloads they are rare for a wide range of applications. I/Os that are not overlapped are allowed to proceed in parallel. I/Os to different tracks run in parallel.

3.3.3 Read stability and write ordering

Read stability together with write ordering are the key concepts of data consistency.

Read stability means that two consecutive reads from the same sector, from primary and secondary, should bring the same information if no writes occur between the reads.

Write ordering simulates the application write ordering at the primary VDisk and at the secondary VDisk. It is a way to guarantee data consistency at the secondary by writing data in the same order it happened on primary.

Write ordering is implemented only in Global Mirror relationships and it is implemented by giving a sequence number to each operation on the primary SVC cluster, and applying writes on the secondary following this sequence number. A sequence can contain more than one write operation if the write operations have different tracks as target.

SVC utilizes an algorithm called *freeze/run* to guarantee read stability.

Freeze/run algorithm

The algorithm is called *freeze/run* because it requires ongoing I/O to be suspended (*freeze*) on a single relationship or across an entire Consistency Group before being allowed to run again. A freeze/run cycle is required for either of the following events:

- ▶ User intervention.
- ▶ An I/O command to either primary or secondary fails such that a host write is not known to be copied to both disks and the primary VDisk is online.

In such circumstances a freeze is performed on the primary of the affected relationship:

- ▶ New write operations from hosts are suspended.
- ▶ Host writes where the data was not written, with successful completion, to both primary and secondary are stalled. In other words, completion to the host is delayed.

- ▶ All ongoing write activity is allowed to complete until it is either stalled or completed to the host, but only if both primary and secondary writes are completed.
- ▶ New read I/O is suspended.

When the freeze is achieved across all relationships in a Consistency Group and any system transients have completed, then a decision is made as to how to proceed. An attempt might be made to perform a recovery copy (copy data from primary from sectors that had I/Os in flight during the error condition) which, if successful, allows I/O to resume without loss of synchronization. Otherwise, assuming the primary is still online, a run is performed with the relationship ConsistentStopped. As a result a run I/O is resumed, which means the following:

- ▶ Host I/Os are allowed to complete.
- ▶ New I/Os are allowed to proceed issuing writes only to the primary.
- ▶ Read I/Os are allowed to proceed.

Difference map (Global Mirror)

The *difference map* performs change recording. This records which areas of the disk were not copied from primary to secondary or were changed since they were copied, as well as recording which areas have writes in flight when *synchronized* or *copying*.

The difference map is implemented as a non-volatile bitmap. One bit corresponds to a grain of data and each grain corresponds to 256 KB of aligned disk space.

A difference map is maintained on all online nodes that are members of the I/O group of a primary VDisk. Space for a difference map is permanently reserved for the secondary VDisk for use when it needs to act as a primary.

The grain state at a specific moment can be clean, out of sync, or dirty.

Clean means that there is no I/O in flight and the primary and secondary are synchronized.

Out of sync means that the primary and secondary are synchronized, but I/O is in flight and must complete before being marked clean. It could also mean that after a disruption or failure, RecoveryCopy is required to complete successfully before declaring clean.

If a grain is *dirty*, the primary and secondary are not synchronized.

Non-volatile Journal

The *non-volatile journal* data structure maintains details of all writes that are active on consistent synchronized relationships. These details include the VDisk, LBA, count of sectors, and for Global Mirror the sequence number that is assigned.

The contents of the non-volatile journal are maintained on the nodes in the I/O group of the primary VDisk for the relationship and are updated at the same time as the bitmap in the difference map. The same messages that maintain the locks also drive the maintenance of the non-volatile journal.

The non-volatile journal has the same availability as any of the different maps associated with the VDIs in the same I/O group. In the event that the nodes of the I/O group are offline, then all relationships using the non-volatile journal change from consistent synchronized to consistent stopped. Then the non-volatile journal is de-allocated and is re-established when the nodes re-appear.

3.3.4 Global Mirror time restrictions

The Global Mirror algorithms operate so as to maintain a consistent image at the secondary. To accomplish this, Global Mirror keeps a record of all write I/Os sent to the secondary. One I/O operation at the primary VDisk is depicted in Figure 3-11 on page 78.

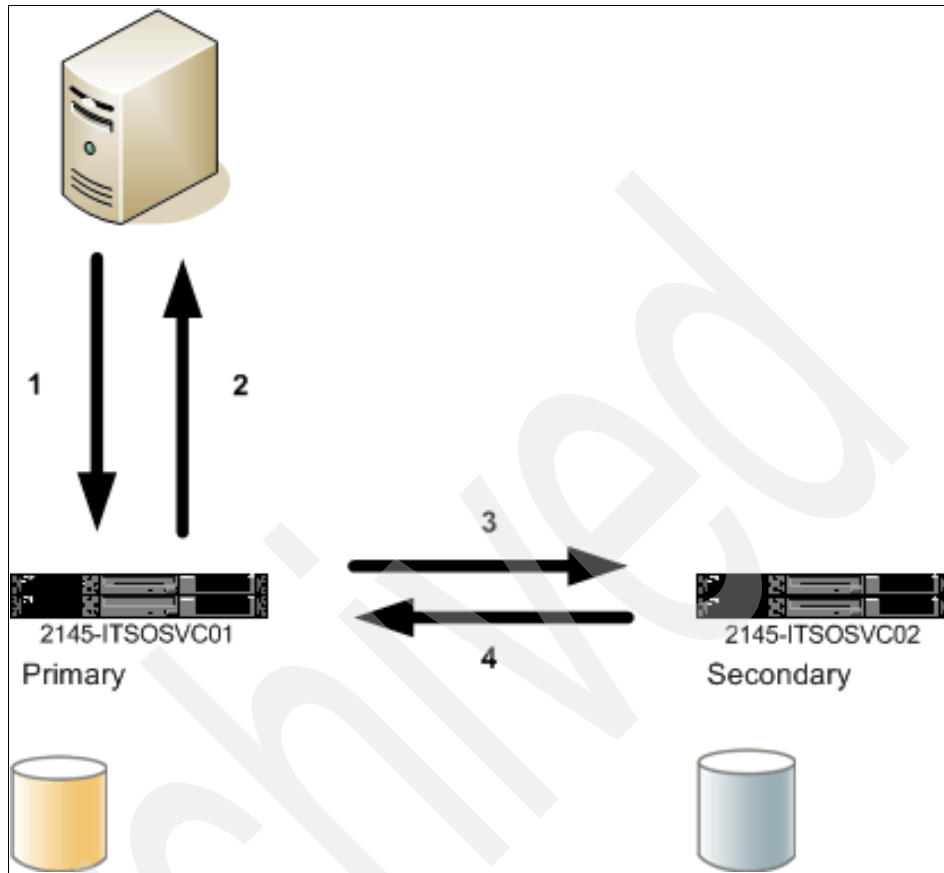


Figure 3-11 SVC Global Mirror write operation

Following is the sequence for the write operation:

1. The host server makes a write I/O to the primary SVC.
2. The write returns as completed to the host server's application.
3. At the same time, that is, in a non-synchronous manner, primary SVC sends the necessary data so that the updates are reflected on the secondary SVC VDisks.
4. The target SVC returns write complete to the source SVC when the updates are secured in the target SVC cache. The source SVC resets its Global Mirror change recording information.

The primary SVC cluster monitors the time it takes for the secondary SVC cluster I/O to complete. The maximum supported time for this operation is 80 ms. If the

operation does not complete in 80 ms, SVC Remote Copy suspends the relationship due to an I/O error.

Note: Global Mirror does not support Intersite latency greater than 80 ms.

3.4 Remote mirroring compatibility

Compatibility applies to SVC 4.1.1 release only.

3.4.1 Operating systems

Remote Copy Services are transparent for the host servers attached to the SVC. All major operating systems are supported. Operating systems supported by the SVC can be found at the following Web site:

<http://www-03.ibm.com/servers/storage/support/software/sanvc/installing.html>

3.4.2 HBAs

All major vendor HBAs are currently supported for SVC attachment. HBA models, driver version, and firmware level compatibility matrix can be found at the following Web site:

<http://www-03.ibm.com/servers/storage/support/software/sanvc/installing.html>

3.4.3 RAID controllers

All major vendor storage subsystems are currently supported for SVC attachment. These can be found at the following Web site:

<http://www-03.ibm.com/servers/storage/support/software/sanvc/installing.html>

3.4.4 SAN switches

Supported SAN switches and firmware versions can be found at the following Web site:

<http://www-03.ibm.com/servers/storage/support/software/sanvc/installing.html>

3.4.5 Intersite communication

Intersite communication is discussed in great detail in section 2.5, “Distance factors for intersite communication” on page 34.

Currently, SVC Remote Copy services support any technology for intersite communication that satisfies the latency requirement of Metro Mirror or Global Mirror and has bandwidth available to support normal operation and peak workload.

Supported technologies can be found at the following Web site:

<http://www-03.ibm.com/servers/storage/support/software/sanvc/installing.html>

3.5 Remote mirroring limitations

Limitations apply to SVC 4.1.1 release only.

In this section we present Remote Copy Service limitations. For a complete list of current SVC limitations, visit the following Web site:

<http://www-03.ibm.com/servers/storage/support/software/sanvc/installing.html>

3.5.1 Relationship per VDisk

One VDisk can participate in only one relationship, and it can be primary or secondary.

3.5.2 Number of relationships

SVC Remote Copy supports 1024 relationships. The relationship limit applies to the total number of relationships, where the total is the sum of Metro Mirror and Global Mirror relationships.

3.5.3 Number of Consistency Groups

SVC Remote Copy supports up to 256 Consistency Groups. A Consistency Group limit applies to the total number of Consistency Groups, where the total is given by the sum of Metro Mirror and Global Mirror Consistency Groups.

3.5.4 Number of relationships per Consistency Group

A Consistency Group can hold up to 512 relationships at this time.

3.5.5 Data mirroring capacity

The SVC cluster can mirror up to 16 TB of data. This limitation comes from the bitmap space, where for each relationship blocks of 8 GB of bitmap space are allocated. It does not matter if the VDisk size is less than 8 GB, it has at least 8 GB bitmap space allocated to control grain operations. For example, one relationship with VDIs of 5 GB, requires 8 GB of bitmap space.

3.5.6 Distance limitation

The SVC supports the use of distance extender technology to increase the overall distance between local and remote clusters. This includes DWDM and FCIP extenders. If this extender technology involves a protocol conversion, then the local and remote fabrics should be regarded as independent fabrics, limited to three hops each. The only restriction on the interconnection between the two fabrics is the maximum latency allowed in the distance extender technology.

Remote Copy Services are not distance limited but latency limited. It is the time for write completion that matters for the SVC code.

Metro Mirror supports latency up to 20 ms. Metro Mirror is a synchronous remote copy, and latency has direct impact on I/O operations.

Global Mirror supports link latency up to 80 ms. Latencies above 80 ms relationships are suspended due to I/O errors.

If we take 1ms of latency for each 100 Km to give an idea of distance, using dark fiber (as an example) for connection, this corresponds to approximately 8000 Km for Global Mirror.

3.5.7 Background copy

Background copy is delivered by the intercluster link, which is one single I/O channel between the clusters' focal point. Today, SVC nodes are equipped with 4 Gbps HBAs, which up limits the background copy processes to 4Gbps.

3.6 Remote copy requirements

Note: Updated requirements for SVC are located at the following Web location:

<http://www-1.ibm.com/support/docview.wss?rs=591&uid=s591S1002858>

3.6.1 Licensing

Remote Copy Services is a licensed feature. If you currently do not have the license, contact your IBM sales representative for more information or visit the following Web site:

http://www-306.ibm.com/common/ssi/OIX.wss?DocURL=http://d03xhttpc1001g.boulder.ibm.com/common/ssi/rep_ca/5/897/ENUS106-425/index.html&InfoType=AN&InfoSubType=CA&InfoDesc=Announcement+Letters&panelurl=&paneltext=

Figure 3-12 on page 83 shows Metro and Global Mirror features enabled.

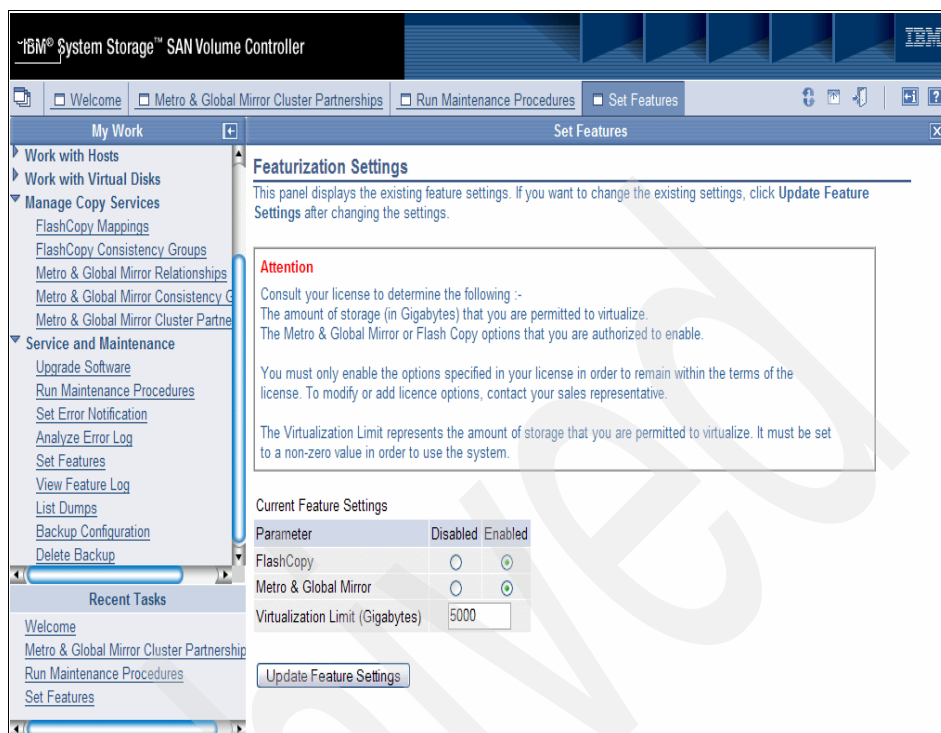


Figure 3-12 SVC current feature settings

3.6.2 Fabric requirements

Remote Copy Services allow for communication between distinct clusters. This communication is carried out using I/O channels composed of Fibre Channel paths between nodes of HBAs. That is only possible when SVC nodes from different clusters have their ports in the same zone. We name this kind of zone an intercluster zone. The requirement is that at least one I/O channel exists between distinct clusters. We recommend that each port of the primary SVC cluster talks with the equivalent port at the secondary, in a symmetrical configuration, using counterpart fabrics for high availability.

The local or remote fabric must not contain more than three hops in each fabric. Any configuration that causes this to be exceeded is unsupported. When a local fabric is connected to a remote fabric for Metro Mirror, the hop count between a local node and a remote node must not exceed seven.

For example, node A in fabric A wants to connect to node B in fabric B. Within fabric A, node A takes two hops to reach the ISL that connects fabric A to fabric B. The hop count is at two at this point in time. Traversing the ISL between fabric

A and fabric B takes the hop count to three. After it reaches fabric B it takes three hops to reach node B. The hop count total is six (2+1+3), which is within our limits.

Another example might be where three hops were used within fabric A, and three hops were used in fabric B. This means that there can only be one hop between fabric A and fabric B; otherwise, the supported hop count limit of seven is exceeded. Alternatively, for example, fabric A consists of one hop and fabric B also consists of one hop. This leaves up to five hops that can be used to interconnect switches between fabric A and fabric B.

If multiple ISLs are available between switches, it is recommended that these ISLs are trunked. The switch vendor's recommendation for trunking should be followed.

General fabric requirements for SVC can found at the following Web site:

<http://www-1.ibm.com/support/docview.wss?rs=591&uid=ssg1S1002858>

3.6.3 Intersite communication and storage controller requirements

To ensure that hosts do not perceive the latency of the long distance link, the bandwidth of the hardware maintaining the link and the storage at the secondary site must be provisioned. This supports the maximum throughput delivered by the applications at the primary that are using Remote Copy Services. If the capabilities of this hardware are exceeded, the system becomes backlogged and the hosts receive higher latencies on their write I/O.

Remote Copy implements a protection mechanism to detect this condition and halts the secondary traffic (Global Mirror specific) to try to ensure that such misconfiguration does not impact application availability.

Intersite communication is discussed in detail in sections 2.5, "Distance factors for intersite communication" on page 34 and 2.6, "Bandwidth sizing" on page 46.

Back End storage array controllers planning for SVC Remote Copy Services is covered in section 4.3, "Design considerations and planning" on page 130.

3.6.4 SVC configuration requirements

Because the SVC is used either in a new SAN or attached to an existing SAN, the number of different configurations in which it is used is large, and it is not practical or possible for us to describe every requirement for any particular

scenario. General requirements for the SVC can be found at the following Web site:

<http://www-1.ibm.com/support/docview.wss?rs=591&uid=ssg1S1002858>

Important: Besides Remote Copy requirements, each element's requirements must be satisfied.

3.7 Scenario implementation

Following are the steps to implement SVC Remote Copy Services. Our scenario is a new database installation and the steps are as follows:

1. Primary Cluster and Secondary Cluster zoning.
2. Primary Cluster and Secondary Cluster partnership.
3. Primary Storage Array LUN creation.
4. Primary Storage Array LUN mapping.
5. Secondary Storage Array LUN creation.
6. Secondary Storage Array mapping.
7. Primary SVC cluster MDisk recognition.
8. Primary SVC cluster MDisk group creation.
9. Primary SVC cluster VDisk creation.
10. Primary SVC cluster VDisks mapping.
11. Secondary SVC cluster VDisks mapping.
12. Remote Copy Consistency Group creation.
13. Start data copy (start relationships).
14. Check data copy status (relationships status).

For example, if we are implementing a disaster recovery solution for an existing database, in this case the steps to use are:

1. Primary Cluster and Secondary Cluster Partnership.
2. Secondary Storage Array LUN creation.
3. Secondary Storage Array mapping.
4. Primary SVC cluster MDisk group creation.
5. Primary SVC cluster VDisks mapping.
6. Secondary SVC cluster VDisks mapping.
7. Remote Copy Consistency Group creation.

In this section, we approach SVC Remote Copy Services in a practical way. No discussions or extensive explanations follow each step.

It is important, at this point, that the primary SVC cluster can reach the secondary SVC cluster.

3.7.1 Fabric configuration

Local and remote fabric implementation is described in section 3.3.1, “Intercluster communication” on page 62.

Local and remote SVC clusters connect to redundant fabrics formed by counterpart fabrics, and we take care to use the same switch ports to have a symmetric configuration.

All necessary zones are in place, including intercluster zones, and they strictly follow each equipment vendor’s best practices and SVC configuration rules (3.6.4, “SVC configuration requirements” on page 84).

3.7.2 Intersite Communication configuration

Intersite communication is discussed in detail in section 2.5, “Distance factors for intersite communication” on page 34.

In 3.7.1, “Fabric configuration” on page 86, we created an intercluster zone, allowing local SVC clusters to log into remote SVC clusters and vice-versa.

Before proceeding we must check the intercluster communication listing to find out if there are intercluster candidates.

To verify that both clusters can communicate with each other, we can use the **svcinfo lsclustercandidate** command. Example 3-1 confirms that our clusters are communicating, as ITSOSVC02 is an eligible SVC cluster candidate for the SVC cluster partnership.

Example 3-1 Listing available SVC cluster for partnership

```
IBM_2145:ITSOSVC01:admin>svcinfo lsclustercandidate
id            configured    cluster_name
000002006040469E no            ITSOSVC02
```

3.7.3 Cluster partnership

After we are sure both SVC clusters are communicating as shown in Example 3-1, we establish SVC cluster partnership as in Example 3-2.

Example 3-2 Creating an intercluster partnership at primary

```
IBM_2145:ITSOSVC01:admin>svctask mkpartnership -bandwidth 100 ITSOSVC02
```

We do the same for the remote SVC cluster: **svcin**fo **lsc**luster**c**andidate followed by an **svctask** **mkpartnership**.

We need to establish this relationship for both clusters. If not, we end up with a partially configured intercluster, where we cannot create Remote Copy relationships.

We see this particular situation after executing **mkpartnership** for one cluster and before executing the second cluster with **svcin**fo **lsc**luster.

Example 3-3 shows how to view intercluster partnership configuration and status (fully_configured).

Example 3-3 Verifying partnership

```
IBM_2145:ITS0SVC02:admin>svcin
```

id	name	location	partnership	bandwidth
cluster_IP_address	cluster_service	IP_address	id_alias	
000002006040469E	ITS0SVC02	local		
9.43.86.40				
9.43.86.41		000002006040469E		
000002006180311C	ITS0SVC01	remote	fully_configured	100
9.43.86.29	9.43.86.30		000002006180311C	

If anything other than **fully_configured** is displayed, check the fabric for zone correctness, and the SVC log messages for errors.

3.7.4 Storage controllers LUN creation

In our example the application needs 40 GB of space, which is our small database. Our primary storage array has eight arrays, and we create LUNs of 5 GB on each array to distribute the I/Os across the maximum amount of spindles possible.

Figure 3-13 on page 88 shows the creation of a 5 GB LUN with recommended settings. The storage subsystem is an IBM System Storage DS4500.

ITSODS4500_A - Specify Capacity/Name (Create Logical Drive)

On this screen, you specify the capacity and unique name for an individual logical drive. You must indicate exactly how much of the array's available capacity you want to allocate for an individual logical drive.

Array RAID level: 5
Array available capacity: 55,555 GB

Logical Drive parameters

New logical drive capacity: 5 Units: GB

Name (30 characters maximum): ITSOSVC01LUN7

Advanced logical drive parameters:

☒ Use recommended settings
☐ Customize settings (I/O characteristics and controller ownership)

< Back Next > Cancel Help

Figure 3-13 LUN creation window at the primary storage array

It is usual to have predefined volume (LUNs) sizes in a site that correlates with best practices for administration of specific equipment. For example, when using Shark storage arrays equipped with 70GB SCSI disks, it is best to use LUNs of 17.5 GB, 35 GB, 70 GB, and 140 GB disk size—we have considered the formatted size of the disk.

We change the names of LUNs to facilitate our administration from storage subsystem default to our convention. If your site does not have a standard name convention in place, you should consider it. The LUN name convention is *<Name of Host>LUN<lun unit number>*.

Even though it is not always possible, it helps to reduce the administration overhead if you can have the same configuration in both sites. In our implementation example it is possible to create the same LUNs number on the secondary array controller and have them mapped to the same LUNs.

LUNs are mapped to the four connections of each SVC cluster node. Two connections on each Fabric and each node. Figure 3-14 on page 89 shows a LUN mapping: Host Group ITSOSVC01 contains all connections from both

nodes. In Figure 3-14, we can see the name of the LUN, what it is mapped to, and where (LUN).

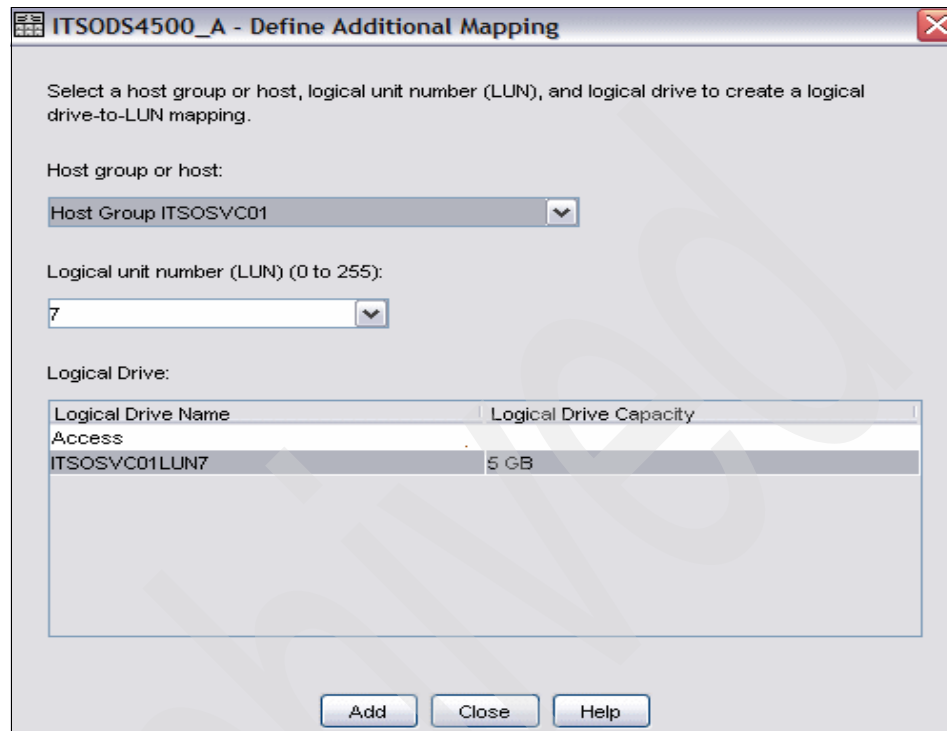


Figure 3-14 Primary Storage Array Controller LUN Mapping

After both controllers present eight LUNs each for your SVC clusters respectively, we can move to the SVC steps.

3.7.5 MDisk creation

The first step here is to check if all MDisk are available. We are looking for eight unmanaged MDisk. If not, we can discover the MDisk as shown in Example 3-4.

Example 3-4 Detecting new MDisk

```
IBM_2145:ITSOSVC01:admin>svctask detectmdisk
```

After we force the SVC to look for new MDisk, we can list them. We have the result sorted by mode (unmanaged). We can see the result in Figure 3-15 on page 90. Our new MDisk are *mdisk4* through *mdisk11*.

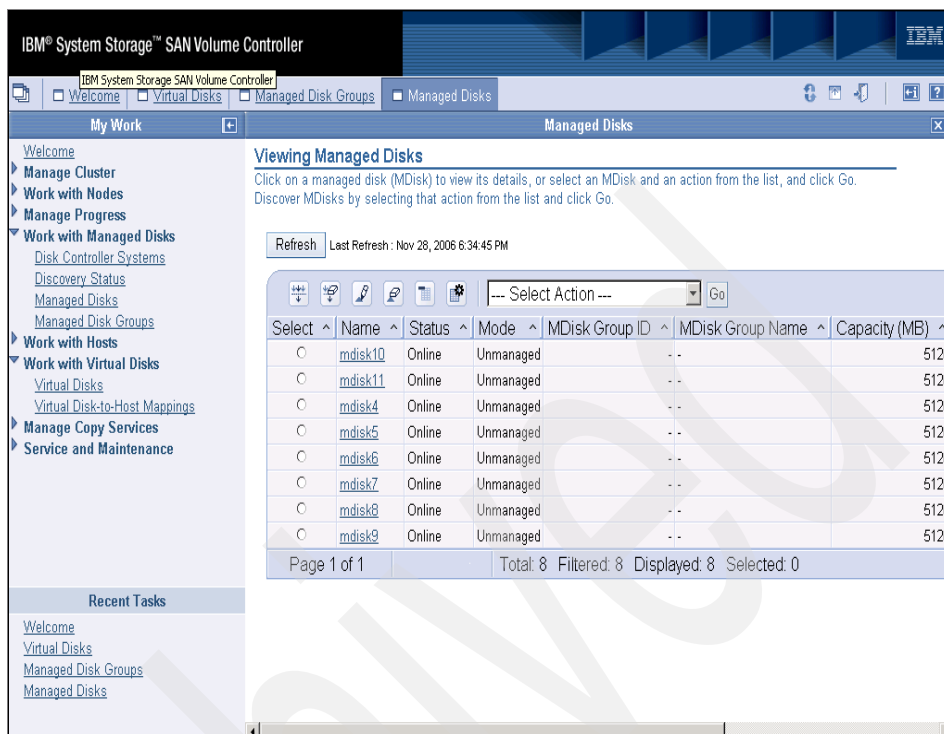


Figure 3-15 Primary SVC unmanaged MDisk list

We can change the name of MDisk to facilitate some sort of operation or simply to fit in a specific site name convention. Our convention is *<short name of application>mdisk<unit>*.

Figure 3-16 on page 91 shows the rename operation started from the MDisk window.

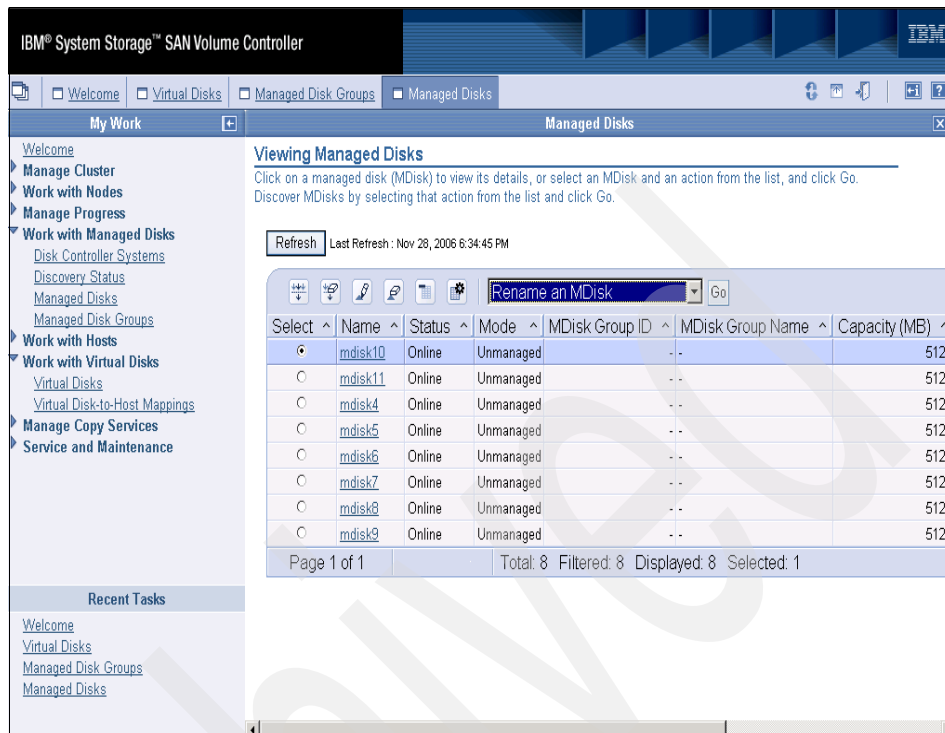


Figure 3-16 Managed Disks Rename an MDisk drop menu option

Renaming huge amounts of MDisks takes some time. We encourage you to use the CLI and maybe scripts to accomplish that. Figure 3-17 on page 92 shows the rename process from mdisk10 to app01mdks10. *App01* stands for a specific application. We can take advantage of that on sort operations.

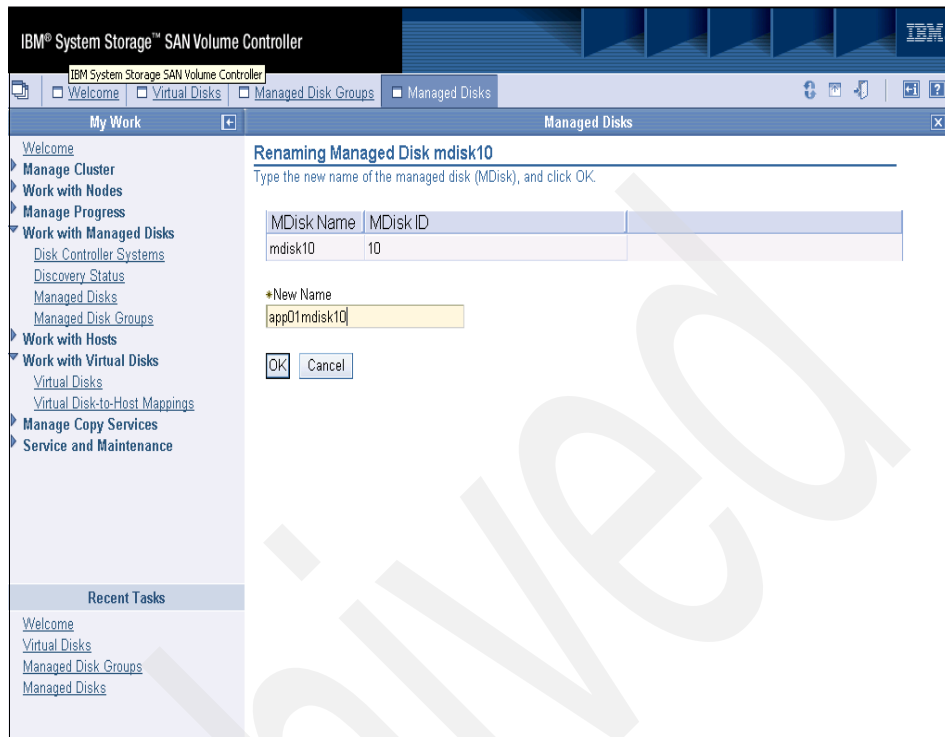


Figure 3-17 *mdisk10 is called app01mdisk10 to facilitate management*

It is a good practice to differentiate primary VDisks from secondary. It helps to manage the environment and helps to protect against mistakes. In our case, we use an *rc* prefix on the secondary to indicate that they are Remote Copy targets. It could be *s* (for secondary) and *p* (for primary) in front of the primary VDisks' names. Figure 3-19 on page 94 shows a window with the rename process using CLI as the secondary. The new name of MDisk starts with *rc*.

Example 3-5 Renaming a MDisk using CLI

```
ITS0SVC02:admin>svctask chmdisk -name rcapp01mdisk4 mdisk4
```

After we have all MDisks renamed on both clusters, we can move to the MDisk group creation, where we group MDisks together to create striped VDisks.

3.7.6 MDisks Group Creation

We create a new MDisk group from where we get the extents for the VDisks for our APP01 application.

Figure 3-18 shows the creation steps after selecting **MDisk Group Management** → **Create an MDisk Group**.

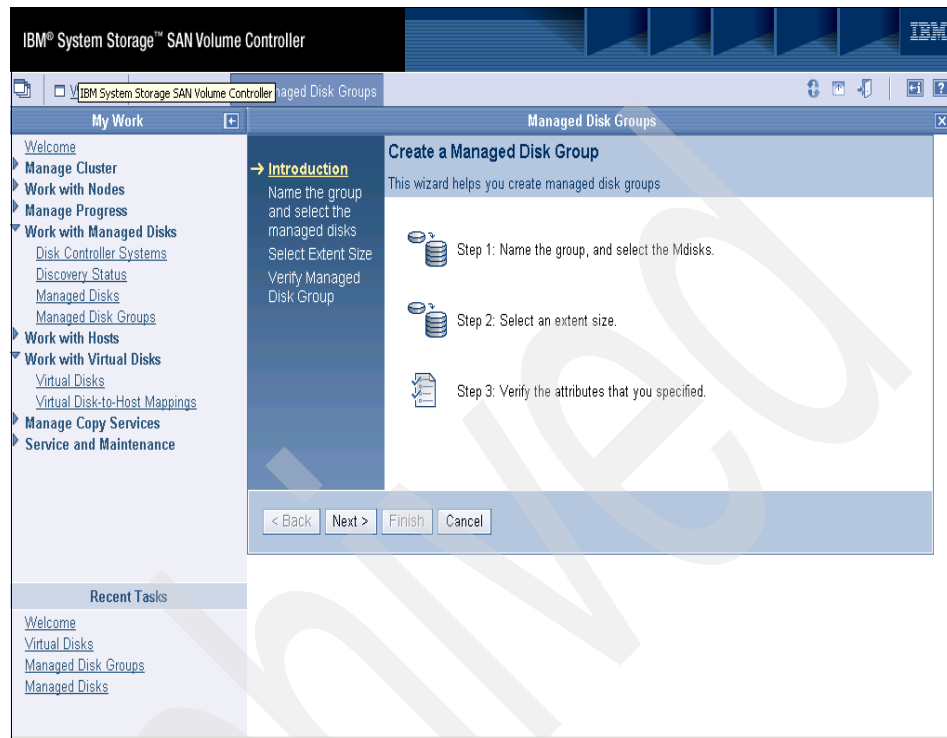


Figure 3-18 Two steps to create a managed disk group

The first step is to give the MDisk Group a name. In our case we name it MDG01APP01 and pick up the disks that belong to the new MDisk Group. Figure 3-19 on page 94 shows step 1.

MDG01APP01 is made up from MDG<sequence number for the sample application><short name of application>.

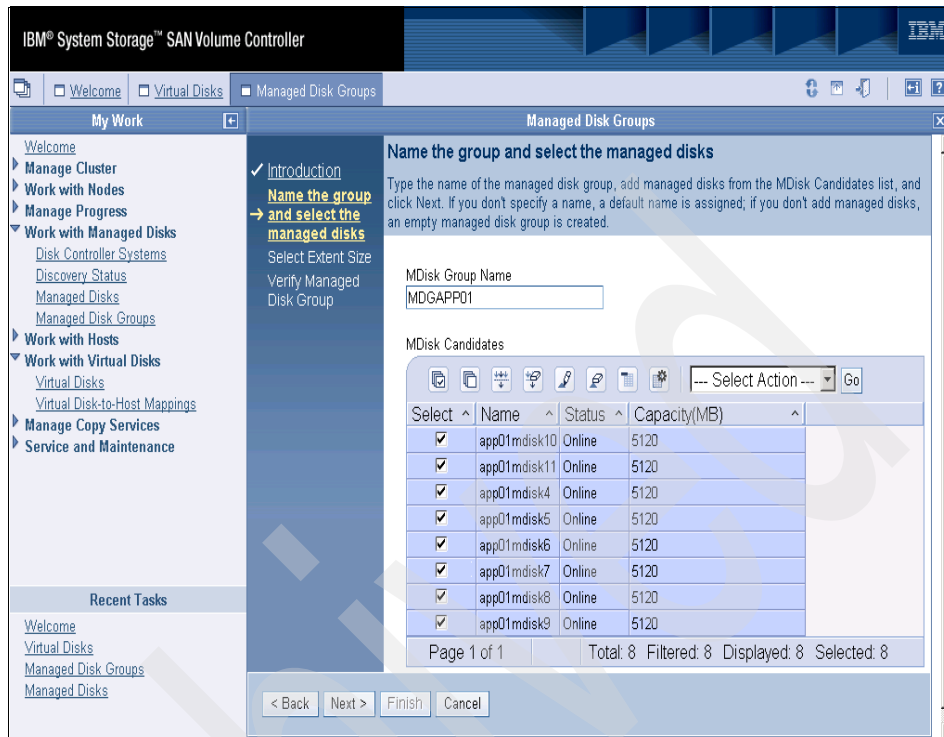


Figure 3-19 MDisk Group Name and MDisks Candidates pick up window

After naming the new MDisk Group and selecting all eight MDisks, select the extent size. VDisks are made up of extents. Possible values range from 16 MB to 512 MB. We choose 128 MB. Figure 3-20 on page 95 shows the extent size selection menu.

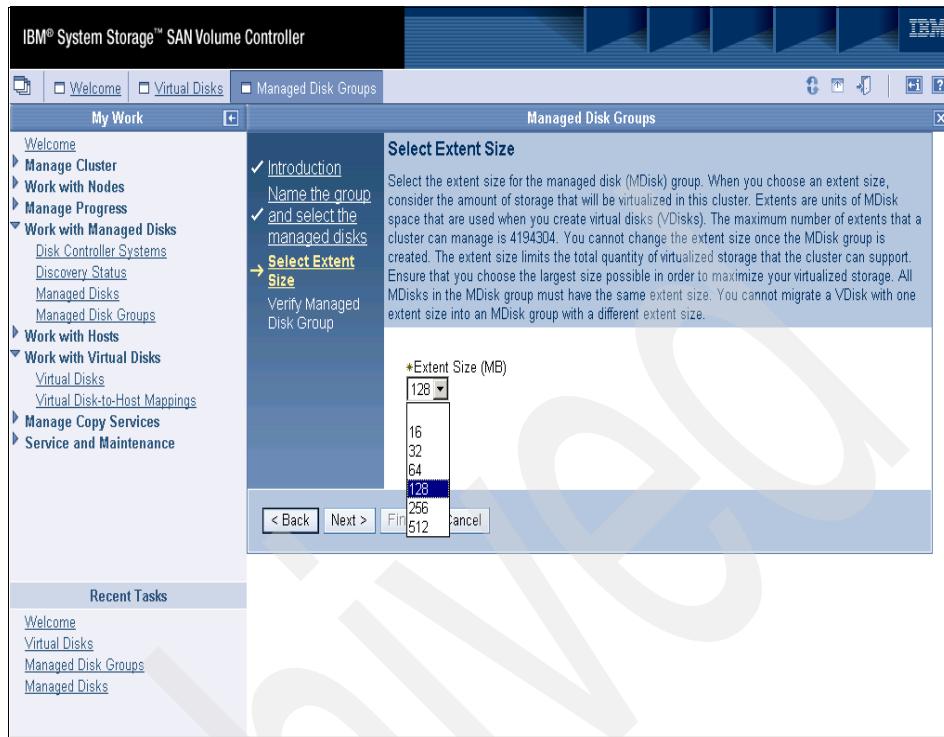


Figure 3-20 Extent size menu option. Extent size was choose to be 128MB for MDG01APP01

Extent size has a direct relationship with the maximum supported capacity of a SVC cluster. A unique SVC cluster can support up to 2^{22} extents at this writing time.

3.7.7 VDisks creation

There are six steps to create a VDisk when you choose to work with the GUI. These steps are shown on the VDisk Wizard window shown in Figure 3-21 on page 96.



Figure 3-21 GUI Wizard lists steps to create a VDisk

Steps 1 and 2 were done previously. Step 3 is shown in Figure 3-22 on page 97. The I/O Group, Preferred Node, and MDisk Group names are selected.

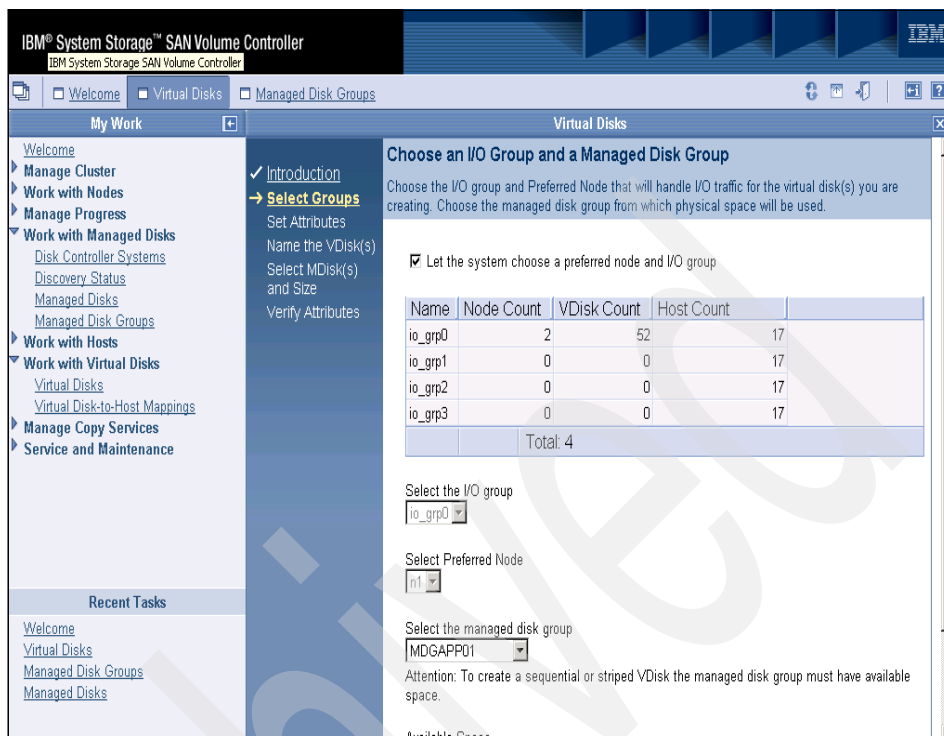


Figure 3-22 The wizard asks for I/O Group, Preferred Node, and MDisk Group name

Figure 3-23 on page 98 shows what is necessary to complete step 4—the type of VDisk, cache mode, unit device identifier, and how many VDIs we are creating.

The VDisk has three policies: Sequential, Striped, and Image. When a VDisk is created using a *sequential* policy, its extents are allocated from a single specified managed disk. When a VDisk is created using a *striped* policy, its extents are allocated from the specified ordered list of managed disks. *Image* mode provides a direct block-for-block translation from the Managed Disk to the VDisk with no virtualization. This mode is intended to allow virtualization of MDIs that already contain data written directly and not through a SVC.

There are two cache modes: readwrite and none. When VDIs are created using *readwrite* cache, all read and write I/O operations that are performed by the VDisk are stored in cache. When VDIs are created not using cache (none), all read and write I/O operations that are performed by the VDisk are not stored in cache.

Unit identifier allows us to attribute one integer number for each VDisk.

It is possible to create more than one VDisk when using the GUI. It is as simple as specifying the number of VDisks desired on the specified field.



Figure 3-23 The wizard asks for the type of vdisk, cache mode, unit identifier, and number of vdisks to create

Figure 3-24 on page 99 shows step 5. The prefix for the VDisks name is entered. Our VDisks are called app01vdisk1, app01vdisk2, and so on.



Figure 3-24 VDisk name prefix when creating more than one VDisk

Figure 3-25 on page 100 shows step 6. It shows which MDisk candidates we pick, the size, and if we asked to format new VDIs.



Figure 3-25 Managed Disk Candidates, size, and format check box

After clicking the **Finish** button, the VDisks are created. Figure 3-26 on page 101 shows the created VDisks.

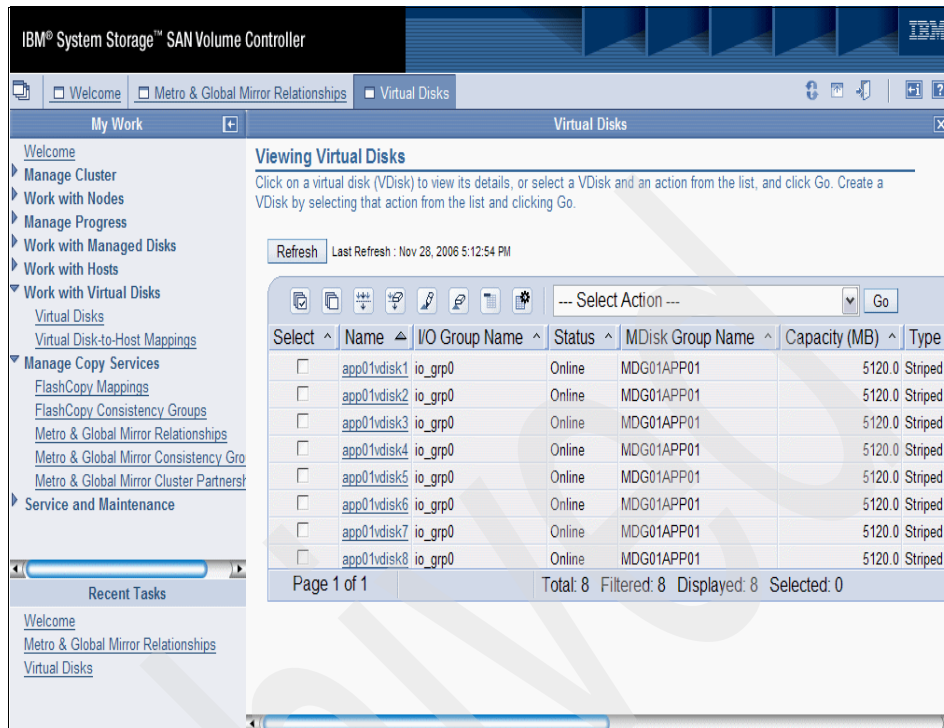


Figure 3-26 Primary SVC cluster VDisks

Figure 3-27 on page 102 lists the VDisks filtered by *rcapp01** from the secondary SVC cluster. All eight VDisks are online and come from the MDG01APP01 MDisk group.

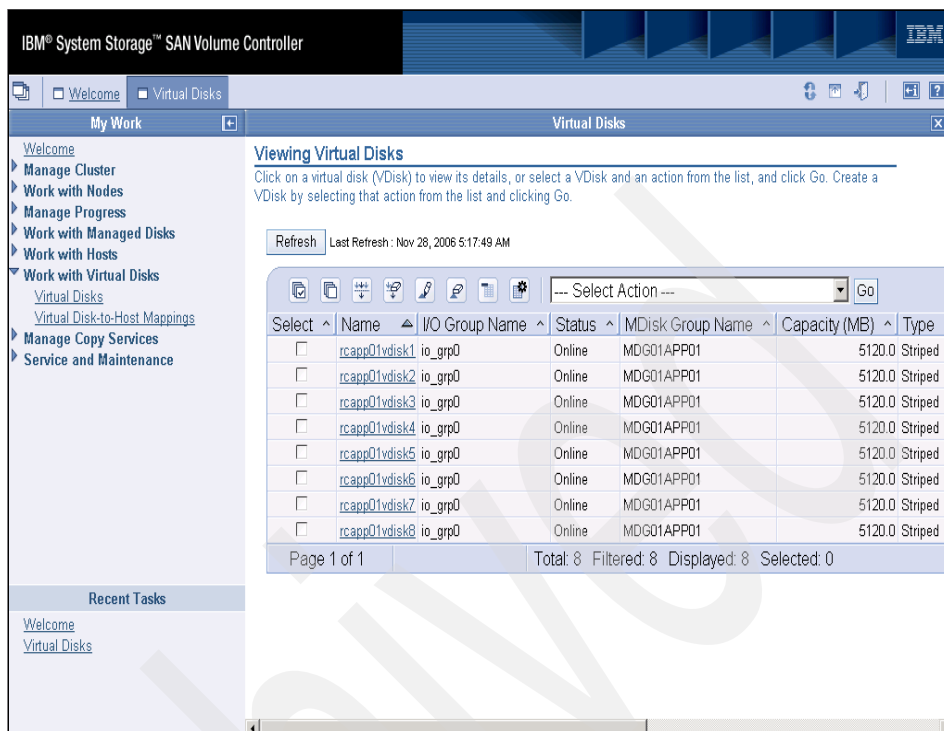


Figure 3-27 Secondary SVC cluster VDisks

Figure 3-26 on page 101 and Figure 3-27 show our application APP01 production VDisks (app01 VDisks) and disaster recovery VDisks (rcapp01 VDisks). Our naming convention makes it easy to identify who is master and who is auxiliary and mainly create necessary relationships because all VDisks are correctly ordered and numbered.

In our scenario, app01vdisk01 has auxiliary rcapp01vdisk01, and so on.

3.7.8 Creating Consistency Groups

In this section we create a Consistency Group to hold APP01 relationships. A Consistency Group is necessary to guarantee application consistency through all application VDisks. There are three steps:

1. Select the auxiliary cluster.
2. Add mirror relationships.
3. Verify settings before creation.

These steps are shown in Figure 3-28 on page 103.

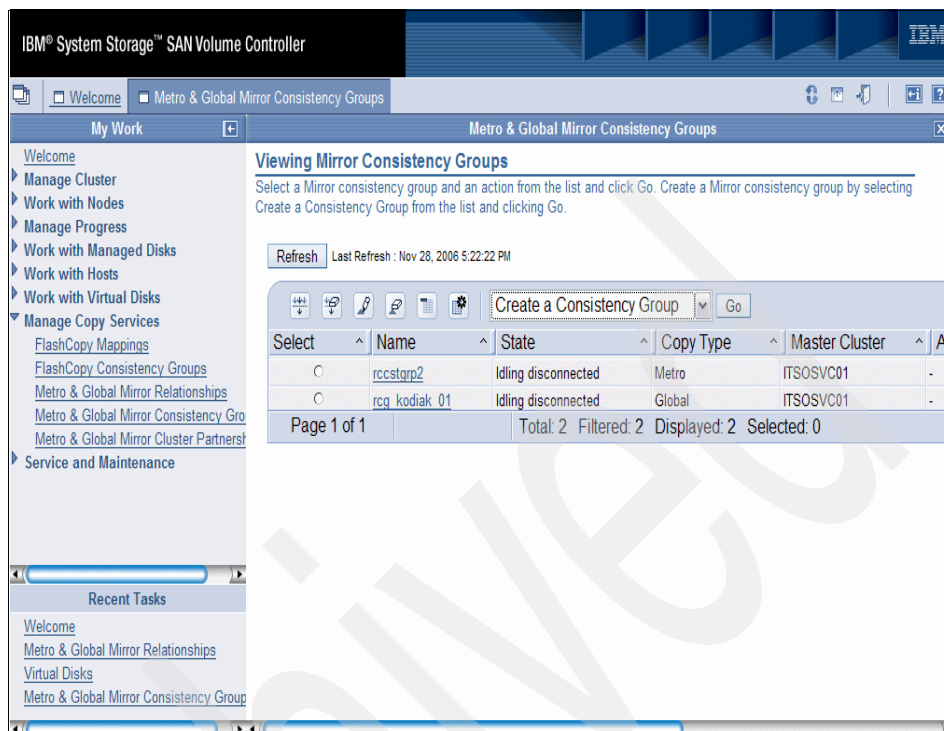


Figure 3-28 Consistency Group wizard

The Consistency Group name must reflect the application's name. In our case, the Consistency Group name is APP01CG. Figure 3-29 on page 104 shows this. Check the **Create an inter-cluster Mirror consistency group** option.

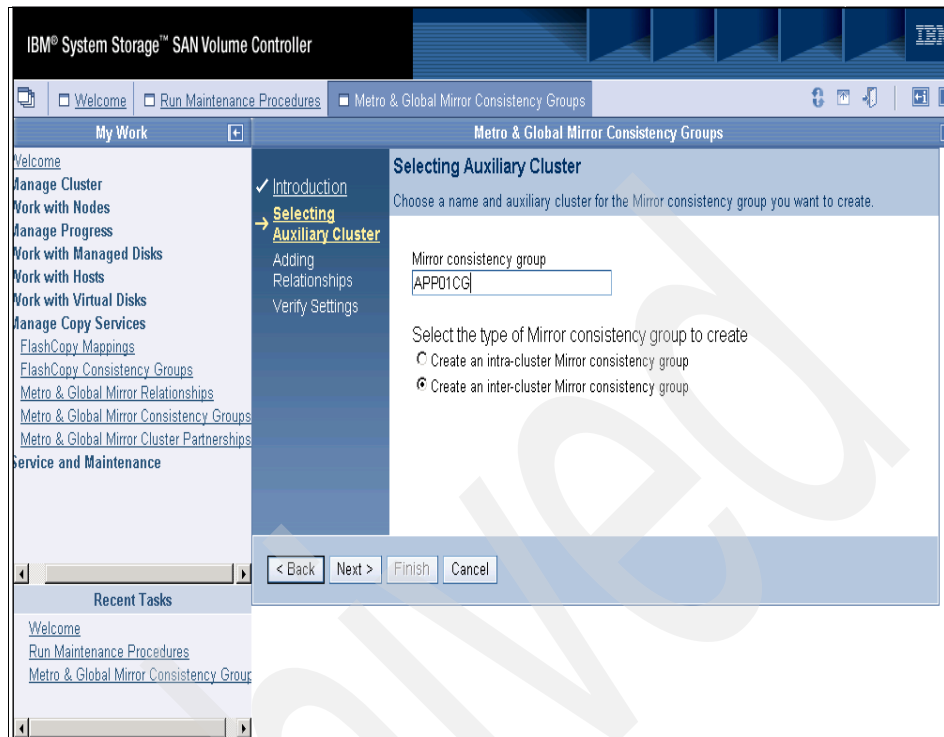


Figure 3-29 Mirror Consistency Group name and cluster police

The next window allows you to pick relationships to belong to the Consistency Group. We do not have any at this point (we do not show this window for that reason).

After the creation of the Consistency Group, we have a Consistency Group named *APP01CG*, as shown in Figure 3-30 on page 105.

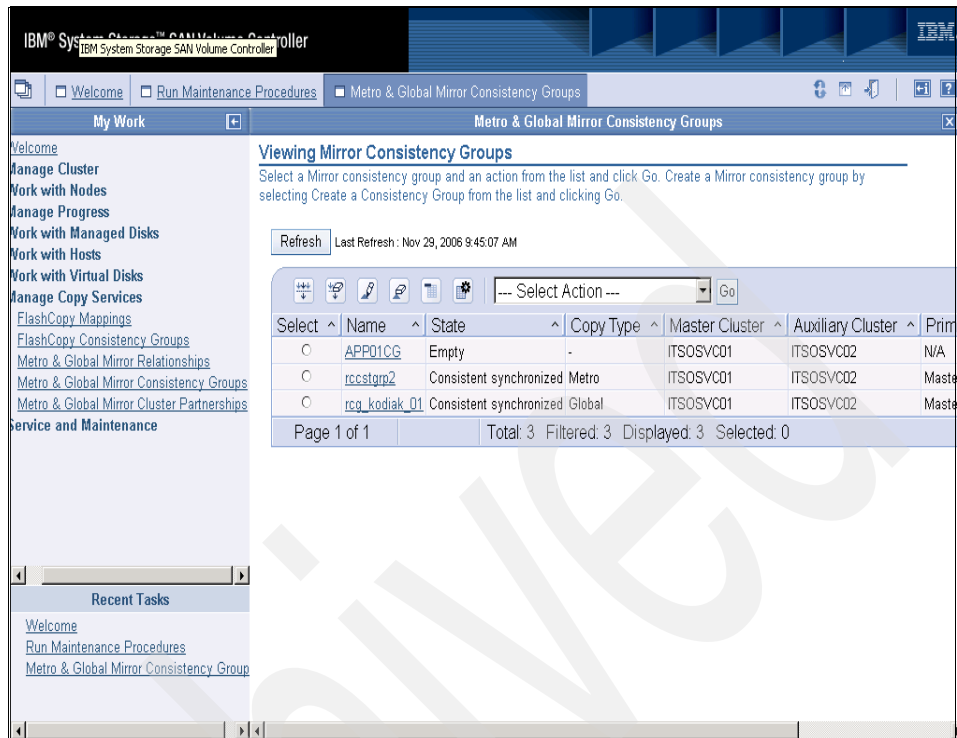


Figure 3-30 Consistency Group

APP01CG needs to be populated with relationships. The Copy type is defined when the first relationship is added.

3.7.9 Relationship creation

Figure 3-31 on page 106 shows the Manage Copy Services menu located on the left panel of the SVC window:

- ▶ FlashCopy Mappings
- ▶ FlashCopy Consistency Groups
- ▶ Metro & Global Mirror Relationships
- ▶ Metro & Global Mirror Consistency Groups
- ▶ Metro & Global Mirror Cluster Partnership



Figure 3-31 SVC GUI Manage Copy Services menu

From **Manage Copy Services** (Figure 3-31), select **Metro & Global Mirror Relationships**. Select **Create a Relationship** from the menu (Figure 3-32 on page 107).

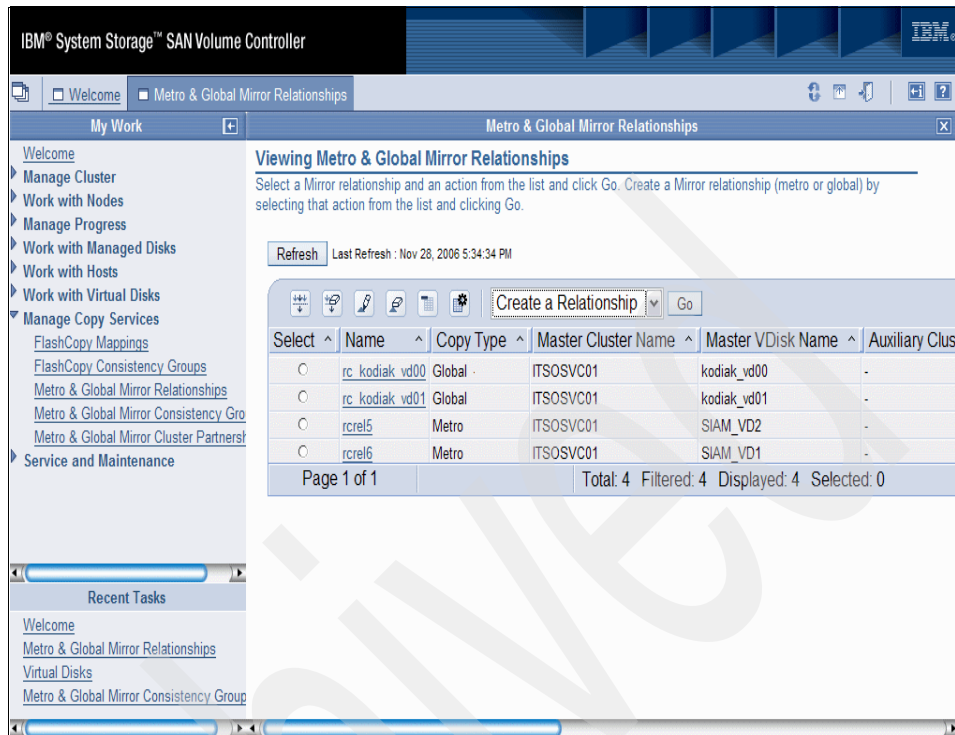


Figure 3-32 Creating a Relationship from the GUI

The first window from the GUI wizard is shown in Figure 3-33 on page 108. The Metro Mirror relationship or the Global Mirror relationships are created basically in the same way. There is just one window option that distinguishes one from the other. If you are using the CLI, you have to specify the option **-global** at the end of **mkrcrelationship**.



Figure 3-33 Creating a Relationship Wizard

There are practically five steps to create a relationship using. It is just one command from the CLI (**mkcrrelationship**).

In the first step we input the relationship name, copy type, and cluster relationship (inter or intracluster).

The relationship name should follow a well-defined and documented naming convention. If you do not have one, take a moment to consider it before proceeding. In our case, we name each relationship the following way: *<application short name>RL<sequence number>*, where sequence number matches the VDisks sequence number. The purpose of the naming convention is to facilitate relationship management. We quickly know that APP01RL1 is the relationship containing VDisks app01vdisk1 and rapp01vdisk1.

The Copy type is Metro Mirror or Global Mirror, Synchronous or Asynchronous. For long distance (latency), consider using Global Mirror.

Metro Mirror

The Metro Mirror relationship is the default type. Figure 3-34 shows a Metro Mirror window.



Figure 3-34 Creating a Metro Mirror relationship

Global Mirror

Select the **Global Mirror Relationship** option to create a Global Mirror relationship (Figure 3-35 on page 110). If you are creating relationships using the CLI, do not forget to put the **-global flag** at the end of the **mkrcrelationship** command. If not, a Metro Mirror relationship is created.

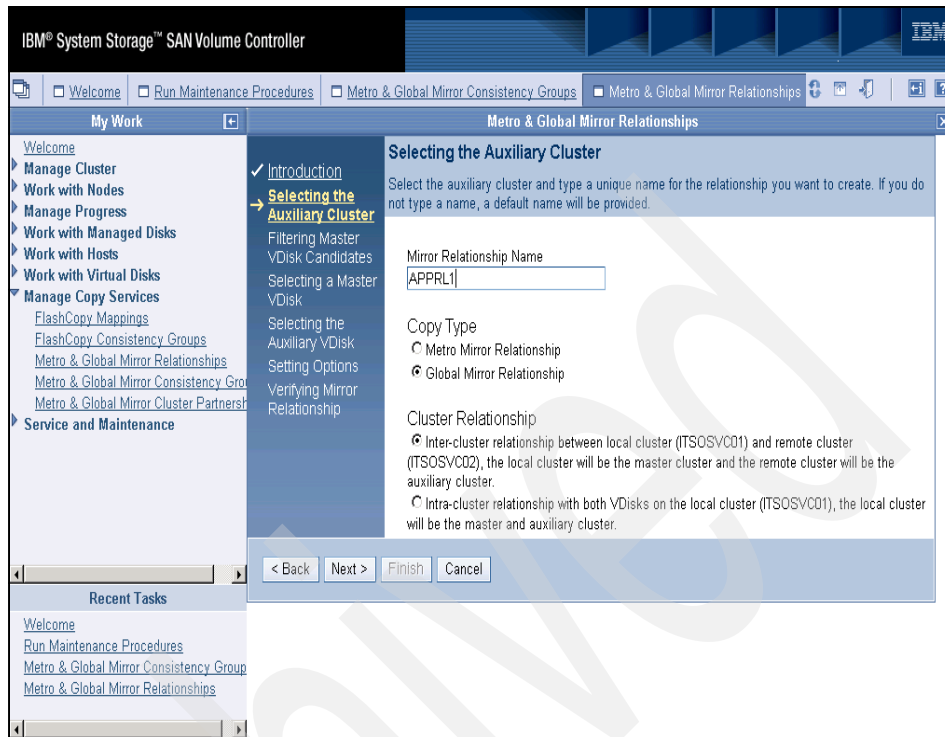


Figure 3-35 Creating a Global Mirror relationship

After we select the name and type of copy, we select master and auxiliary VDisks. In our case, we are looking for app01vdisk1 and rcapp01vdisk1. We can use the filter window to facilitate the search process as shown in Figure 3-36.

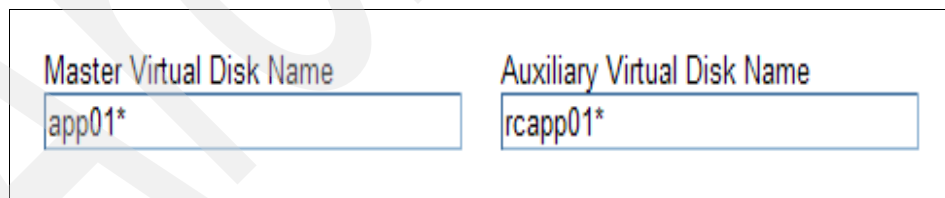


Figure 3-36 GUI VDisks candidates filter window.

After master and auxiliary VDisks are selected, (app01vdisk1 and rcapp01vdisk1) we have two options: *Synchronized* check box and *Consistency Group*. (See Figure 3-37 on page 111).



Figure 3-37 Synchronized check box and Consistency Group menu

In our case, both VDisks, master and auxiliary, do not contain any information, so select the Synchronized box. If you are creating a relationship where the master contains data, do not select the Synchronized option. The Synchronized option means that the relationship state is stopped consistently.

The Consistency Group menu allows us to pick the previously created Consistency Group APP01CG.

We created the first relationship of the eight necessary to support our application APP01. The other seven are created using the CLI. See Example 3-6.

Example 3-6 Creating a Global Mirror relationship

```
IBM_2145:ITS0SVC01:admin>svctask mkrcrelationship -master app01vdisk8
-aux rcapp01vdisk8 -name APP01RL8 -cluster 000002006040469E -sync
-global
```

3.7.10 Starting a relationship or Consistency Group

Starting a relationship or Consistency Group puts secondary VDisks offline. The remote host does not have write access to secondary volumes. Data flows from primary to secondary and overwrites all information that can be contained on secondary VDisks. This way we rename remote VDisks with a prefix, so that it can be easily identified as a target. After selecting a specific relationship or Consistency Group (as we do), select **Start Copy Process** → **Go** (Figure 3-38).

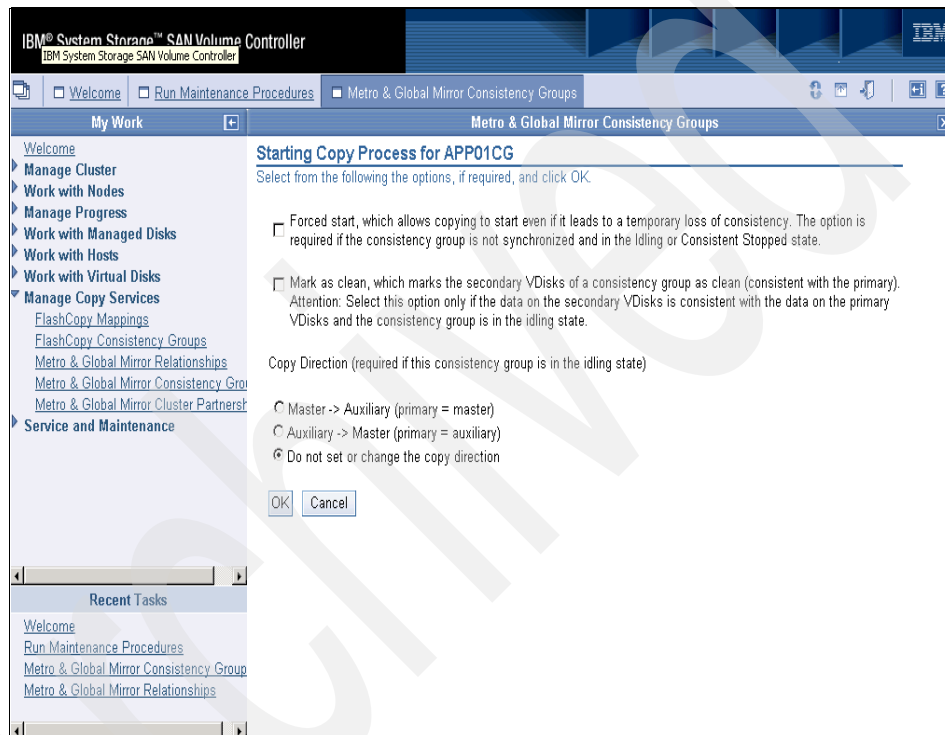


Figure 3-38 Start a relationship or Consistency group options

Forced start asks for user confirmation because for a certain period of time, consistency is lost. The period of time is the time required for the background process to copy the necessary grains from primary to secondary or from secondary to primary, depending on the copy direction chosen.

Mark as clean can be used if no writes occur on primary and secondary. Both VDisks contain the same information.

Copy direction allows us to switch copy direction if required. Switching copy direction is discussed later in this section.

3.7.11 Stopping a relationship or Consistency Group

If, for example, we want to enable remote VDisks for I/O, we need to stop a consistent synchronized relationship or Consistency Group. To stop a Consistency Group, select the Consistency Group, and then select **Stop Copy Process** (See Figure 3-39).

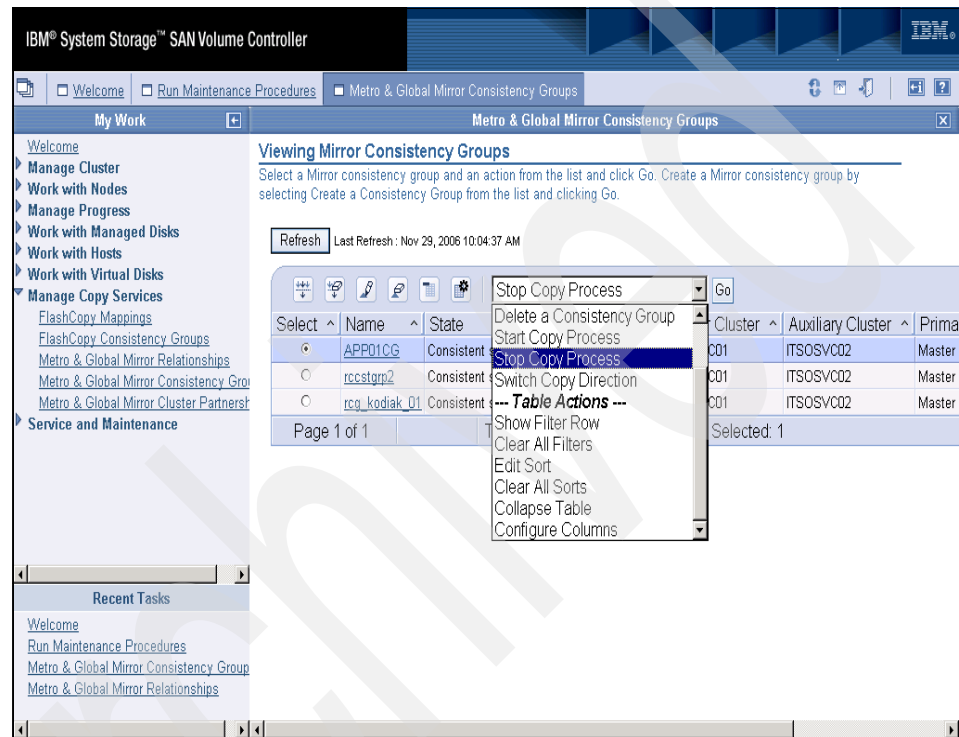


Figure 3-39 Stopping APP01CG consistent synchronized Consistency group

Select the option for write access (See Figure 3-39). If not, the secondary stays offline to the remote SVC cluster hosts.

Figure 3-40 on page 114 shows us stopping the copy process.

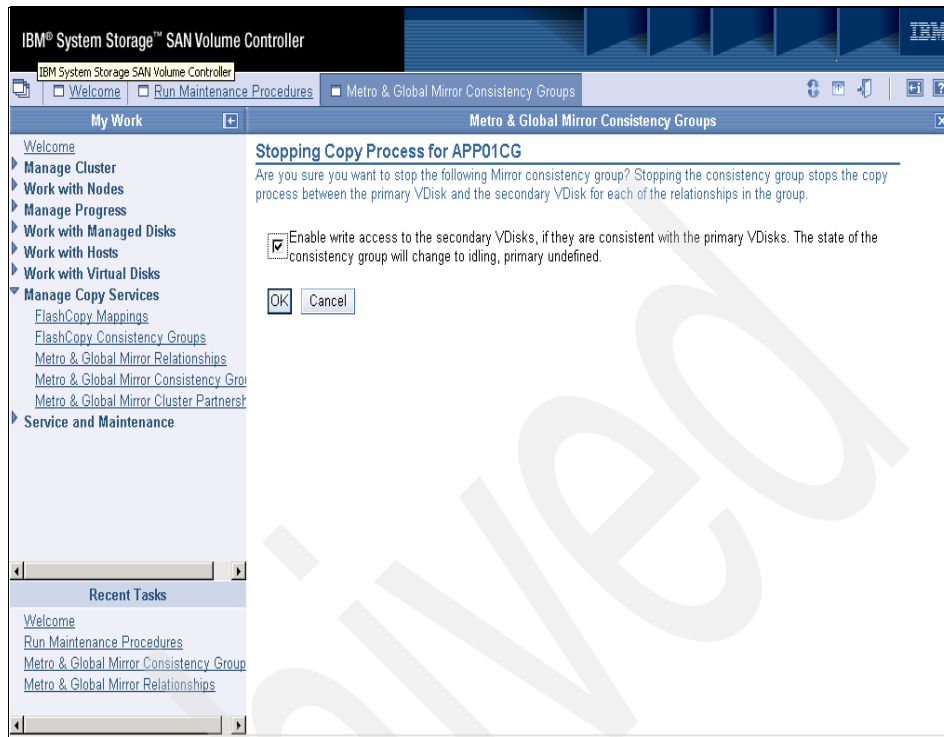


Figure 3-40 Stopping Copy Process for a Consistency group

The application host server can access its disks, and the application on the remote site server is started.

3.7.12 Switching copy direction

We will consider a situation or procedure where production switches to the secondary site for a certain period of time and then switches back to the primary site. In this scenario, data flow is from auxiliary to master VDisks for a certain period of time. This can be done by first stopping data replication from primary to secondary, enabling the secondary for host server write, and then (after application validation and other procedures), switching relationships or Consistency Groups.

Example 3-7 on page 115 shows how to check the APP01CG state and stop enabling host access.

Example 3-7 Shows how to stop a consistent synchronized Consistency Group

```
IBM_2145:ITS0SVC01:admin>svcinfolsrcconsistgrp -delim : -filtervalue
name=APP01CG
id:name:master_cluster_id:master_cluster_name:aux_cluster_id:aux_cluste
r_name:primary:state:relationship_count:copy_type
253:APP01CG:000002006180311C:ITS0SVC01:000002006040469E:ITS0SVC02:maste
r:consistent_synchronized:8:global
IBM_2145:ITS0SVC01:admin>svctask stoprconsistgrp -access APP01CG
IBM_2145:ITS0SVC01:admin>svcinfolsrcconsistgrp -delim : -filtervalue
name=APP01CG
id:name:master_cluster_id:master_cluster_name:aux_cluster_id:aux_cluste
r_name:primary:state:relationship_count:copy_type
253:APP01CG:000002006180311C:ITS0SVC01:000002006040469E:ITS0SVC02::idli
ng:8:global
```

The change is carried out using the **svctask switchrconsistgrp** command for a Consistency Group or **svctask switchrcrelationship** for a relationship, which is shown in Example 3-8.

Example 3-8 Changing relationship

```
IBM_2145:ITS0SVC01:admin>svctask switchrconsistgrp -primary aux
APP01CG
IBM_2145:ITS0SVC01:admin>svcinfolsrcconsistgrp -delim : -filtervalue
name=APP01CG
id:name:master_cluster_id:master_cluster_name:aux_cluster_id:aux_cluste
r_name:primary:state:relationship_count:copy_type
253:APP01CG:000002006180311C:ITS0SVC01:000002006040469E:ITS0SVC02:aux:c
onsistent_synchronized:8:global
```

When switching copy direction, we must remember that data from auxiliary VDisks flows and writes on master VDisks. For a certain amount of time, relationships or consistency groups are inconsistent. The other important thing is that master VDisks are offline.

3.7.13 Monitoring background copy progress and state

The **svcinfolsrcrelationship** command shows the relationship state and allows it to monitor background copy progress.

Example 3-9 on page 116 shows the output from one of our relationships, APP01RL1. State and progress are in bold. The relationship state is inconsistent, but Global Mirror is mirroring data from primary to secondary and it is eight percent of total.

Example 3-9 Viewing state and copy progress

```
IBM_2145:ITS0SVC01:admin>svcinfolsrcrelationship -delim : -filtervalue
name=APP01RL1
id:name:master_cluster_id:master_cluster_name:master_vdisk_id:master_vd
isk_name:aux_cluster_id:aux_cluster_name:aux_vdisk_id:aux_vdisk_name:pr
imary:consistency_group_id:consistency_group_name:state:bg_copy_priorit
y:progress:copy_type
43:APP01RL1:000002006180311C:ITS0SVC01:43:app01vdisk1:000002006040469E:
ITS0SVC02:4:rcapp01vdisk1:master::inconsistent_copying:50:8:global
```

You can obtain other information if you issue the **svcinfolsrcrelationship** command against a specific relationship, as shown in Example 3-10.

Example 3-10 Viewing relationship background copy progress and state

```
IBM_2145:ITS0SVC01:admin>svcinfolsrcrelationship APP01RL1
id 43
name APP01RL1
master_cluster_id 000002006180311C
master_cluster_name ITS0SVC01
master_vdisk_id 43
master_vdisk_name app01vdisk1
aux_cluster_id 000002006040469E
aux_cluster_name ITS0SVC02
aux_vdisk_id 4
aux_vdisk_name rcapp01vdisk1
primary master
consistency_group_id
consistency_group_name
state inconsistent_copying
bg_copy_priority 50
progress 36
freeze_time
status online
sync
copy_type global
```

The command **svcinfolsrcrelationshipprogress** shows how much of the data is already mirrored. Example 3-11 on page 117 shows output from the APP01RL1 relationship.

Example 3-11 The relationship background copy progress for APP01RL1

```
IBM_2145:ITS0SVC01:admin>svcinfo lsrelationshipprogress APP01RL1
id          progress
43          36
```

In our example, we have eight VDisks in a consistency group called APP01CG. As soon as a relationship belongs to a consistency group, it cannot be manipulated as a standalone relationship. Viewing copy state and progress is possible using **svcinfo lsrrconsistgrp** (Example 3-12).

Example 3-12 Viewing state and copy progress for APP01CG consistency group.

```
IBM_2145:ITS0SVC01:admin>svcinfo lsrrconsistgrp -delim : -filtervalue
name=APP01CG
id:name:master_cluster_id:master_cluster_name:aux_cluster_id:aux_cluste
r_name:primary:state:relationship_count:copy_type
253:APP01CG:000002006180311C:ITS0SVC01:000002006040469E:ITS0SVC02:maste
r:inconsistent_copying:8:global
```

You can run **svcinfo lsrrconsistgrp** directly to the consistency group, as in Example 3-13. In this case, you can see a verbose output showing relationships belonging to the consistency group.

Example 3-13 Viewing detailed listing of APP01CG consistency group

```
IBM_2145:ITS0SVC01:admin>svcinfo lsrrconsistgrp APP01CG
id 253
name APP01CG
master_cluster_id 000002006180311C
master_cluster_name ITS0SVC01
aux_cluster_id 000002006040469E
aux_cluster_name ITS0SVC02
primary master
state inconsistent_copying
relationship_count 8
freeze_time
status online
sync
copy_type global
RC_rel_id 43
RC_rel_name APP01RL1
RC_rel_id 44
RC_rel_name APP01RL2
RC_rel_id 45
RC_rel_name APP01RL3
```

```
RC_rel_id 46
RC_rel_name APP01RL4
RC_rel_id 47
RC_rel_name APP01RL5
RC_rel_id 48
RC_rel_name APP01RL6
RC_rel_id 49
RC_rel_name APP01RL7
RC_rel_id 50
RC_rel_name APP01RL8
```

If you have a considerable amount of standalone relationships and consistency groups, consider automating the whole process. There are a couple of ways to do that. You can consider scripting, which can be very time consuming, or use tools that are ready for you to use as in TPC for Replication. Automation is covered in Chapter 5, “Automation in a Business Continuity solution with SVC” on page 155, where TPC for Replication is discussed in detail.

3.8 Performance guidelines

Several factors are important and should be carefully planned in a new implementation or monitored in a solution already in production:

- ▶ Primary SVC cluster
- ▶ Primary storage array controllers
- ▶ Link bandwidth and latency
- ▶ Secondary SVC cluster
- ▶ Secondary storage array controllers.

Besides subsystems that compose the entire solution, there are best practices that must be considered when creating SVC objects, for example, dedicated MDisk groups. Section 4.3, “Design considerations and planning” on page 130 discusses what to consider when planning for SVC Remote Copy Services. Section 4.4, “Monitoring SVC copy service performance and identifying issues” on page 136 deals with SVC Remote Copy Services already implemented.

Performance considerations in SVC Business Continuity solutions

This chapter discusses the methods and guidelines for monitoring and evaluating SVC performance metrics. It provides an introduction to the data collection and reporting capabilities of *IBM Total Storage Productivity Center (TPC)* and their use in discovering aspects of SVC performance for the purpose of designing, operating, and examining SVC environments.

4.1 Considerations when using SVC in a Business Continuity solution

The SVC can form the main storage component for your Business Continuity solution. In addition to its high availability design, it offers Metro Mirror, Global Mirror (that we collectively call Remote Copy), and FlashCopy services to support the comprehensive Business Continuity efforts incorporating other elements of the IT infrastructure. This chapter focuses on the three copy services features and their interaction with performance aspects.

SVC is designed for versatile use of its copy service functionality in Business Continuity solutions. When you design a specific solution around SVC copy services, consider the following performance related items as part of your planning process:

1. Identify a set of volumes to mirror or copy.
2. Determine relevant I/O loads on identified volumes.
3. Select storage subsystems and interconnect methods.
4. Establish performance monitoring.

The first step is to identify the volumes that participate in Mirror relationships or FlashCopy mappings. Finding the set of volumes and how they are structured into Consistency Groups originating from the application level is not discussed here. It is mentioned only to set the scope for the other steps.

The next aspect involves the analysis of the current and expected I/O workload on these volumes. In section 4.3, “Design considerations and planning” on page 130, we discuss the relevant metrics, monitor, and design options to consider. The results of performance monitoring helps you in determining the I/O requirements for the target volumes and selecting appropriate target disk subsystems and long distance interconnects.

After you have implemented a Business Continuity solution with SVC Remote Copy or FlashCopy services, the operation and maintenance of the environment can benefit from establishing performance monitoring for SVC. Collecting performance metrics such as throughput, I/O rate, and response time on a VDisk, MDisk, and I/OGroup level can help you to detect trends of growth in demand and to adapt to these changes in your environment before they cause impacts on your applications. These and further metrics of performance collected over a period of time under normal operating conditions give you a baseline for comparison in situations with sudden change in I/O load, component failure, or unexpected performance issues.

With the additional enhancements in SVC V4.1, SVC provides a detailed set of performance counters collected on each SVC node. Data is collected at the SVC

node, node port, MDisk, I/O-Group, and VDisk level and covers data rates, I/O rates, latencies, queuing, cache, and inter/intra cluster communication aspects. TPC offers functionality to collect, monitor, and graph those SVC counters.

Section 4.2, “Introduction to TPC performance reporting for SVC” on page 121 presents an overview of the relevant TPC function needed to help you in planning, sizing, and operating an SVC Business Continuity solution. We recommend the use of TPC for collecting SVC performance data in a Business Continuity environment. The SVC provides performance statistics through a standardized SMI-S interface for use by any compliant management software.

4.2 Introduction to TPC performance reporting for SVC

TPC is an open storage infrastructure management solution designed to help reduce the effort of managing complex, heterogeneous storage infrastructures, improve storage capacity utilization, and administrative efficiency. It provides a central point of reporting and control for management of storage infrastructure capacity, performance, and availability.

Note: Enhancements in SVC V4.1 offer additional performance metrics for greater reporting detail that are useful for environments using Metro Mirror, Global Mirror, or FlashCopy. TPC V3.1.3 is required to collect the performance data of SVC V4.1 and was used in preparing this publication.

This chapter focuses on the use of TPC as a tool for SVC performance reporting. It is envisioned to be used in the early phases of establishing a Business Continuity solution, namely the determination of existing I/O workload conditions, as well as in the operational phase for performance reporting.

Reporting versus monitoring

The TPC features related to performance data collection, graphing, and analysis are often referred to as TPC performance monitoring. However, TPC is not an online performance monitoring tool; instead, it is a performance reporting tool. Nevertheless, TPC uses the term “performance monitor” for the configuration of the function that gathers performance data from a storage subsystem.

The difference between the term monitoring and reporting lies with the fact that TPC collects information at certain intervals, stores the data in its database, and allows you to initiate the generation of reports based on collected data. After the data is inserted and processed, you can view the information using default or user defined reports. You are not monitoring or looking at a live online stream of performance data being presented by TPC. TPC is also not an event driven

real-time notification system. However, it allows you to define performance related threshold alerts that can trigger further external event processing. Even though it all works like a monitor without user intervention, this is still performed at the intervals that you specified during the definition of the performance monitor job.

The following sections list the basic steps to integrate SVC into TPC and examine how to use TPC to view the SVC components and SVC performance data.

For more on TPC design, installation, operation, and other applications and components of TPC see the following IBM Redbooks publications:

- ▶ *IBM TotalStorage Productivity Center V3.1: The Next Generation*, SG24-7194
- ▶ *TotalStorage Productivity Center Advanced Topics*, SG24-7348
- ▶ *Monitoring Your Storage Subsystems with TotalStorage Productivity Center*, SG24-7364.

4.2.1 Adding SVC to TPC

Before you can use TPC to monitor SVC performance statistics, TPC needs to be configured to discover the SVC. The process of adding storage devices to a TPC server is described in several chapters of *IBM TotalStorage Productivity Center V3.1: The Next Generation*, SG24-7194. Following are the main steps and the chapter heading numbers:

1. Create an SVC CIMOM user with administrator authority for TPC (chapter 7.7: Configuring CIMOM for SAN Volume Controller).
2. Introduce SVC CIMOM into TPC (chapter 8.4.1:Configuring CIMOMs).
3. Configure data collection for SVC.
 - a. Collect asset information (chapter 8.5.1: Creating probes).
 - b. Collect performance data (chapter 8.5.4: Creating performance monitors).

You define the interval at which performance data is collected in Step b. You need to balance storage space need for collected performance data with your need for detailed and fine grained historic performance data. For a demanding Global Mirror environment, we recommend considering a short interval (five minutes) for better analysis opportunities in case of performance related issues.

Usually the interval at which TPC collects performance data to evaluate the thresholds for alert generation and the interval at which data is stored in the TPC database are the same. You can change this so that TPC collects data and checks for constraint violations at a shorter interval (for example, five minutes), but stores the data only at a longer interval (for example, 15 minutes). This

allows you to balance database storage space demand with your need for finer grained performance constraint violation reports.

Figure 4-1 shows the dialog to define this advanced data collection interval configuration found in the Navigation Tree. Go to **Disk Manager** → **Monitoring** → **Subsystem Performance Monitors** → **YourMonitor'sName**. Now select **Advanced...** in **Sampling and Scheduling**.

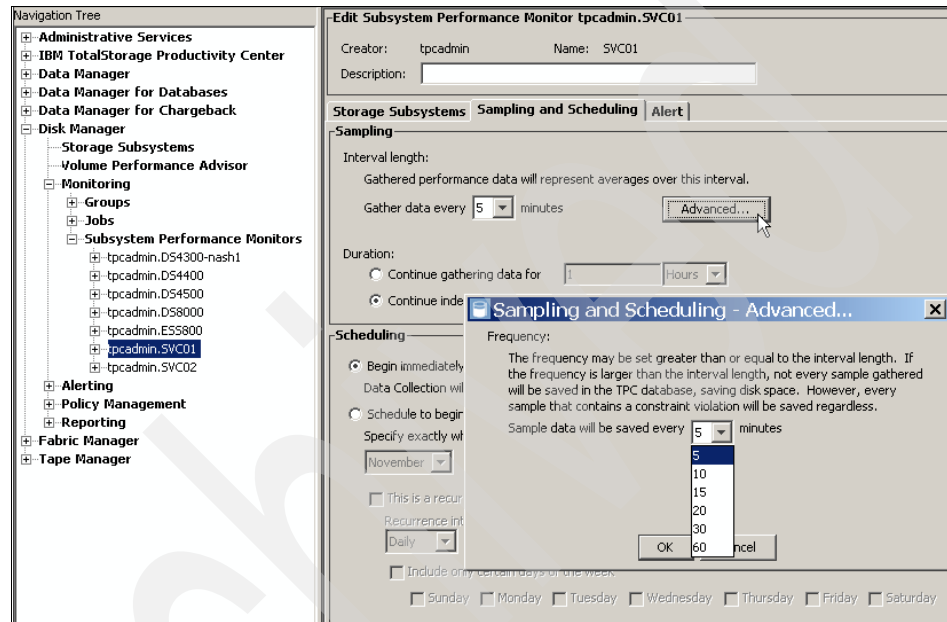


Figure 4-1 Advanced subsystem performance monitor sampling interval

4.2.2 Displaying SVC asset information

TPC presents the asset information in the **topology viewer** as customizable reports, tables, and charts. This can, for example, be used in finding the relationship between a particular VDisk associated with a certain host and the VDisk's underlying MDisk and back end storage subsystem to investigate a report of a user that experiences performance degradation for that VDisk.

Figure 4-2 on page 124 is the topology viewer showing the MDisk and MDisk Group aspect of the SVC.

1. Open the topology viewer for a storage system and go to **IBM TotalStorage® Productivity Center** → **Topology** → **Storage** in the Navigation Tree panel on the left of the TPC application window. The figure shows the back end aspect of the end-to-end relationship between VDisk and the back end subsystem volume. It visualizes the relationship between an MDisk Group

(MDG_DS45R5_8P), an identified MDisk (mdisk1), and its underlying DS4500 Backend Volume (ITSOSVC01_LUN01).

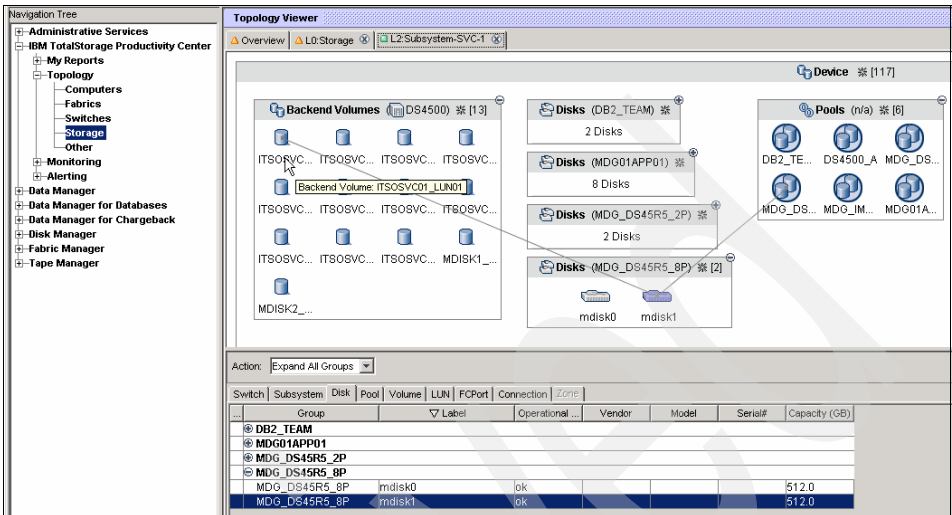


Figure 4-2 SVC in the TPC topology viewer

TPC can also present the relationship between the VDisks of *kodiak* and their underlying MDisks in a tabular form.

- Open the table view by clicking **Disk Manager** → **Reporting** → **Storage Subsystems** → **Volume to Backend Volume Assignment** → **By Volume**, in the Navigation Tree panel on the left of the TPC application window. Figure 4-3 on page 125 shows the filtering function of the TPC report that is opened by clicking **Filter...**, which allows you to narrow the selection of the volume to report on.

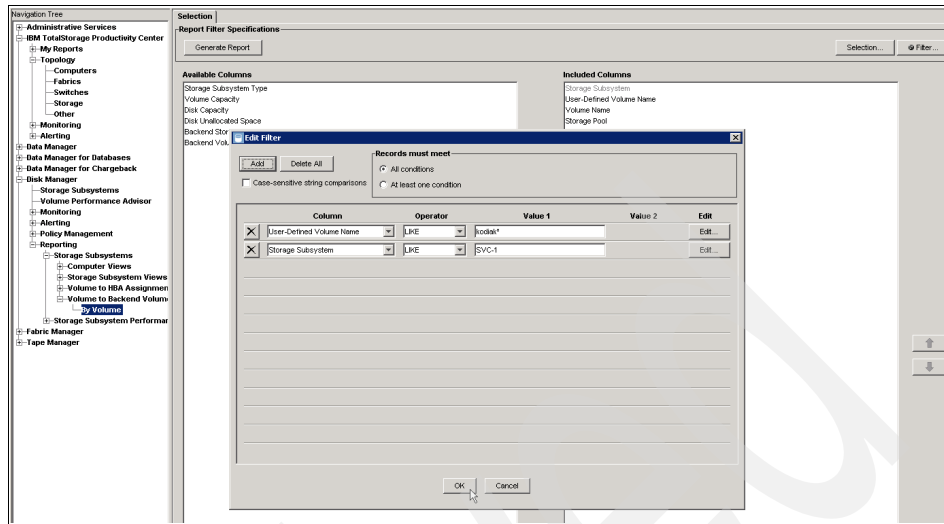


Figure 4-3 SVC back end volume report: VDisk filtering

3. Click the << and >> buttons shown in Figure 4-4 to select the desired column for the report. The reduced selection shown here is tailored to the purpose of our example.
4. Click **Generate Report** to produce the tabular report shown in Figure 4-5 on page 126.

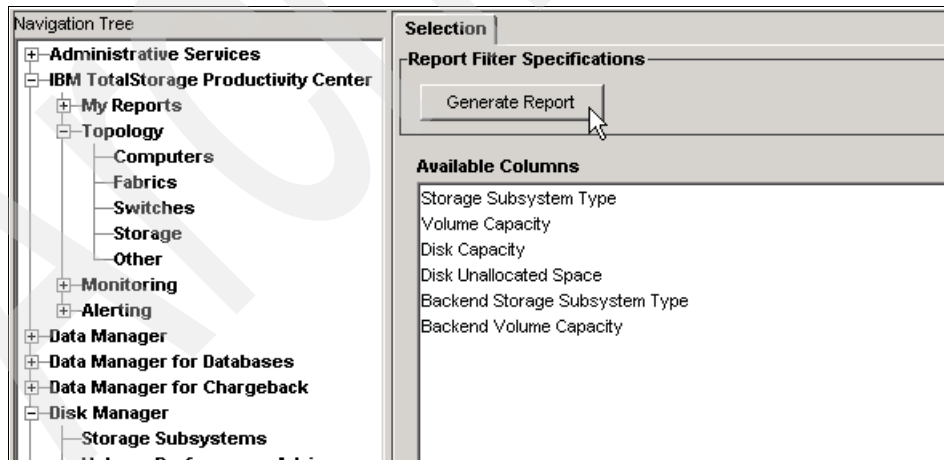


Figure 4-4 SVC backend volume report: Selecting the columns

The report shows all VDisks with a row for each MDisk that contribute extents to that VDisk.

Navigation Tree

Administrative Services

IBM TotalStorage Productivity Center

My Reports

Topology

Computers

Fabrics

Switches

Storage

Other

Monitoring

Alerting

Data Manager

Data Manager for Databases

Data Manager for Chargeback

Disk Manager

Storage Subsystems

Volume Performance Advisor

Monitoring

Alerting

Policy Management

Reporting

Storage Subsystems

Computer Views

Storage Subsystem Views

Volume to HBA Assignment

Volume to Backend Volume

Volumes

Storage Subsystem Performer

Fabric Manager

Tape Manager

Selection Volumes

Volumes: By Volume

Number of Rows: 4

Storage Subsystem	User-Defined Volume Name	Volume Name	Storage Pool	Disk	Backend Storage Subsystem	Backend Volume Name
SVC-1	kodak_vxd0	38	MDG_DS45R5_BP	mdisk0	DS4500	ITSOSVC01_LUN00
SVC-1	kodak_vxd0	38	MDG_DS45R5_BP	mdisk1	DS4500	ITSOSVC01_LUN01
SVC-1	kodak_vxd01	39	MDG_DS45R5_BP	mdisk0	DS4500	ITSOSVC01_LUN00
SVC-1	kodak_vxd01	39	MDG_DS45R5_BP	mdisk1	DS4500	ITSOSVC01_LUN01

Figure 4-5 SVC backend volume report: Generating the report

4.2.3 SVC performance statistics and presentation

This section introduces the TPC performance metrics for SVC and how to view them.

TPC performance metrics

TPC collects performance statistics data through a component called *Storage Subsystem Monitors*. You can find a brief reference for setting up SVC performance data collection in TPC and considerations for interval selection in section 4.2.1, “Adding SVC to TPC” on page 122.

TPC performance reporting is structured around a generalized model of a storage system. The following aspects and terms of the model used in TPC are relevant to the SVC:

- ▶ Storage Subsystem (SVC cluster)
- ▶ I/O Group
- ▶ Node
- ▶ Managed Disk Group
- ▶ Volume (VDisks)
- ▶ Managed Disk
- ▶ Ports

The SVC internally collects performance counters on a per SVC node basis at the level of VDisks, MDisks, ports, and the node itself. TPC aggregates and processes these raw counters into performance metrics at the VDisk, MDisk, port, and node level and further aggregates them to the MDisk group, I/O group, and SVC cluster level metrics. You can find a full list of all TPC performance metrics relevant to and supported for SVC in Appendix A, “TPC performance

metrics for SVC” on page 225. The remaining chapter discusses SVC performance in terms of these metrics.

TPC performance reports

You can view the collected performance data in several ways. We only introduce you to the basic use of reports and graphical presentation of performance data. Further useful topics for gathering workload data or investigating performance issues can be found in *Monitoring Your Storage Subsystems with TotalStorage Productivity Center*, SG24-7364.

The following series of window captures shows how to generate an example performance report on the VDisk level and how to draw a historic performance diagram. This can be useful in both operating and planning an SVC Business Continuity solution by helping you, for example, in finding peak I/O and data rate levels.

1. You begin the steps to generate a TPC performance report for SVC VDisks by selecting the **Disk Manager** → **Reporting** → **Storage Subsystem Performance** → **By Volume** option from the Navigation Tree (see Figure 4-6).
2. Select the time period you want to cover in the report (or leave the default of only reporting on the most current performance data sample), and select from the available SVC metrics.

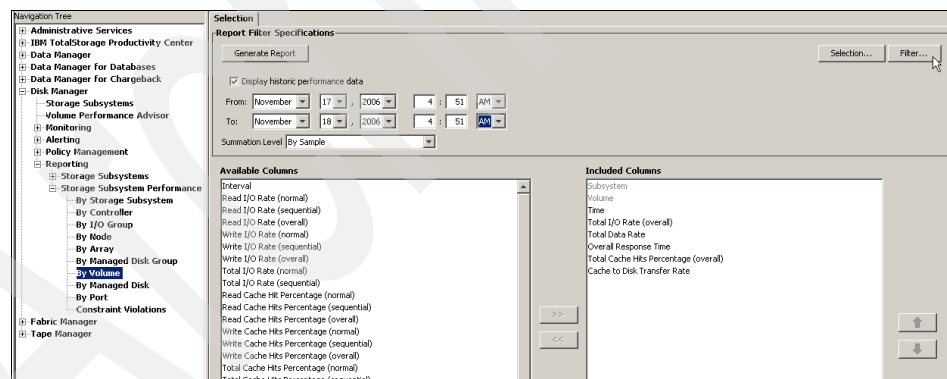


Figure 4-6 Generating a SVC performance report by VDisk

3. Click **Filter...** to restrict your report scope to a subset of the available performance data samples. Figure 4-7 on page 128 shows an example of how you can focus on VDisks with a specific name pattern. Note that you cannot filter the static properties of VDisks, such as their name or the subsystem they are defined on or on the actual performance data values themselves.

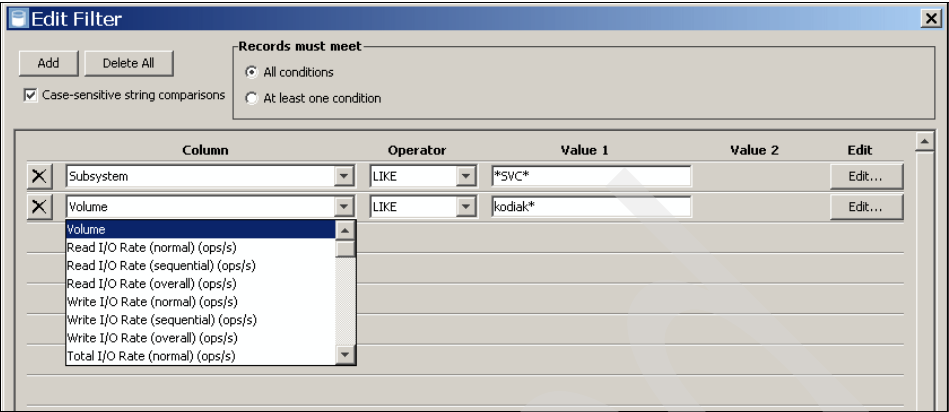


Figure 4-7 Performance report filter conditions

Figure 4-8 shows the resulting example performance report of an artificial workload generated by a test server, reverse sorted by the Total Data Rate metric to allow for easy identification of peak data rates of the server.

Selection: Volumes									
Storage Subsystem Performance: By Volume									
Number of Rows: 1386									
Subsystem	Volume	Time	Total I/O Rate (overall)	Total Data Rate	Overall Response Time	Total Cache Hits Percentage (overall)	Cache to Disk Tran		
SVC01	kodiak_vd00	Nov 17, 2006 5:55:33 AM	270.97 ops/s	1.06 MB/s	8.6 ms/op	80.33 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:55:33 AM	269.41 ops/s	1.05 MB/s	7.4 ms/op	80.06 %			
SVC01	kodiak_vd00	Nov 17, 2006 5:30:28 AM	262.07 ops/s	1.02 MB/s	10.9 ms/op	77.69 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:30:28 AM	261.98 ops/s	1.02 MB/s	10.6 ms/op	77.08 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:20:26 AM	250.85 ops/s	0.98 MB/s	6 ms/op	76.91 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:05:23 AM	250.02 ops/s	0.98 MB/s	5.9 ms/op	76.95 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:45:31 AM	249.98 ops/s	0.98 MB/s	5.9 ms/op	77.1 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:00:22 AM	249.92 ops/s	0.98 MB/s	12.2 ms/op	76.49 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:10:24 AM	249.86 ops/s	0.98 MB/s	6.1 ms/op	77.57 %			
SVC01	kodiak_vd00	Nov 17, 2006 6:00:34 AM	249.84 ops/s	0.98 MB/s	5.5 ms/op	80.07 %			
SVC01	kodiak_vd00	Nov 17, 2006 5:45:31 AM	249.82 ops/s	0.98 MB/s	5.8 ms/op	77.55 %			
SVC01	kodiak_vd00	Nov 17, 2006 5:35:29 AM	249.75 ops/s	0.98 MB/s	5.7 ms/op	77.04 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:40:30 AM	249.66 ops/s	0.98 MB/s	10.3 ms/op	77.65 %			
SVC01	kodiak_vd00	Nov 17, 2006 5:05:23 AM	249.63 ops/s	0.98 MB/s	5.1 ms/op	83.85 %			
SVC01	kodiak_vd00	Nov 17, 2006 5:00:22 AM	249.57 ops/s	0.97 MB/s	9.9 ms/op	82.26 %			
SVC01	kodiak_vd00	Nov 17, 2006 4:55:21 AM	249.54 ops/s	0.97 MB/s	7.4 ms/op	77.54 %			
SVC01	kodiak_vd00	Nov 17, 2006 5:40:30 AM	249.49 ops/s	0.97 MB/s	12.8 ms/op	77.04 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:35:29 AM	249.47 ops/s	0.97 MB/s	77.48 %				
SVC01	kodiak_vd00	Nov 17, 2006 5:20:26 AM	249.34 ops/s	0.97 MB/s	6.7 ms/op	77.24 %			
SVC01	kodiak_vd01	Nov 17, 2006 6:00:34 AM	249.34 ops/s	0.97 MB/s	5.3 ms/op	80.07 %			
SVC01	kodiak_vd00	Nov 17, 2006 5:10:24 AM	249.26 ops/s	0.97 MB/s	6.1 ms/op	79.59 %			
SVC01	kodiak_vd01	Nov 17, 2006 6:05:35 AM	249.05 ops/s	0.97 MB/s	11.2 ms/op	79.17 %			
SVC01	kodiak_vd01	Nov 17, 2006 4:55:21 AM	249.01 ops/s	0.97 MB/s	6.7 ms/op	76.77 %			
SVC01	kodiak_vd00	Nov 17, 2006 6:05:35 AM	248.8 ops/s	0.97 MB/s	10.4 ms/op	78.71 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:25:27 AM	238.43 ops/s	0.93 MB/s	6.2 ms/op	76.5 %			
SVC01	kodiak_vd00	Nov 17, 2006 5:25:27 AM	237.43 ops/s	0.93 MB/s	7.2 ms/op	77.74 %			
SVC01	kodiak_vd00	Nov 17, 2006 5:15:25 AM	231.4 ops/s	0.9 MB/s	12.5 ms/op	76.6 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:50:32 AM	231.37 ops/s	0.9 MB/s	6.4 ms/op	81.76 %			
SVC01	kodiak_vd01	Nov 17, 2006 5:15:25 AM	230.27 ops/s	0.9 MB/s	11.5 ms/op	80.24 %			
SVC01	kodiak_vd00	Nov 17, 2006 5:50:32 AM	229.03 ops/s	0.89 MB/s	7.7 ms/op	81.14 %			
SVC01	kodiak_vd01	Nov 17, 2006 6:10:36 AM	178.33 ops/s	0.7 MB/s	5.4 ms/op	77.26 %			
SVC01	kodiak_vd00	Nov 17, 2006 6:10:36 AM	178.07 ops/s	0.7 MB/s	7 ms/op	77.18 %			
SVC01	kodiak_vd00	Nov 17, 2006 7:40:54 AM	135.97 ops/s	0.53 MB/s	9.5 ms/op	76.73 %			
SVC01	kodiak_vd01	Nov 17, 2006 7:40:54 AM	135.44 ops/s	0.53 MB/s	8.3 ms/op	77.53 %			
SVC01	kodiak_vd01	Nov 17, 2006 9:16:12 AM	135.28 ops/s	0.53 MB/s	7.1 ms/op	76.75 %			
SVC01	kodiak_vd00	Nov 17, 2006 12:01:45 PM	134.74 ops/s	0.53 MB/s	10.3 ms/op	5.9 %			
SVC01	kodiak_vd01	Nov 17, 2006 11:36:40 AM	134.03 ops/s	0.52 MB/s	8.9 ms/op	76.61 %			
SVC01	kodiak_vd01	Nov 17, 2006 12:01:45 PM	133.72 ops/s	0.52 MB/s	8.8 ms/op	8.08 %			
SVC01	kodiak_vd00	Nov 17, 2006 11:36:40 AM	133.44 ops/s	0.52 MB/s	11.2 ms/op	77.87 %			
SVC01	kodiak_vd00	Nov 17, 2006 9:16:12 AM	132.82 ops/s	0.52 MB/s	8.1 ms/op	77.59 %			
SVC01	kodiak_vd01	Nov 17, 2006 8:11:00 AM	131.57 ops/s	0.51 MB/s	4.2 ms/op	82.01 %			
SVC01	kodiak_vd00	Nov 17, 2006 8:11:00 AM	129.66 ops/s	0.51 MB/s	5.1 ms/op	82.86 %			
SVC01	kodiak_vd00	Nov 17, 2006 8:36:05 AM	126.05 ops/s	0.49 MB/s	4.7 ms/op	83.19 %			
SVC01	kodiak_vd01	Nov 17, 2006 10:21:25 AM	125.91 ops/s	0.49 MB/s	5.6 ms/op	77.53 %			

Figure 4-8 Example performance report

4. Click the chart icon in the top left section of the report panel (highlighted in Figure 4-8 on page 128) to view a performance chart based on the performance report data and on the selected VDisks. In this example, the VDisk *kodiak_vd00* is selected.
5. Use Figure 4-9 to select your chart options.

Note: In the TPC 3.1.3 release used to prepare this book, all performance metric selected for charting must use the same unit type. Unit type examples are *percent*, *I/Os per second*, and *milliseconds*.

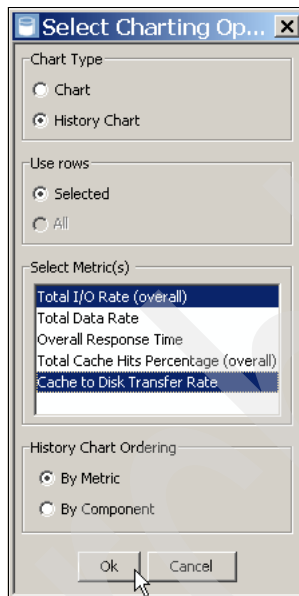


Figure 4-9 Performance chart options

Figure 4-10 on page 130 shows the example performance chart for the VDisk *kodiak_vd00* on Total I/O Rate (overall) and Cache to Disk Transfer Rate metrics.



Figure 4-10 Example performance chart

The following sections discuss SVC Business Continuity solution sizing approaches, as well as SVC and subsystem performance behavior situations during normal and overload conditions that make use of performance reports such as the one in Figure 4-10.

4.3 Design considerations and planning

This section highlights the performance and load metrics that are of foremost interest when planning for the use of SVC Metro Mirror, Global Mirror, or FlashCopy. We provide design guidelines and consideration points considering

the discussed performance metrics. We present methods for obtaining the presented performance statistics in an SVC environment using TPC and an example environment lacking SVC and TPC.

4.3.1 Key metrics

SVC needs to do additional processing for read as well as write operations on VDisks that are part of a Remote Copy relationship or a FlashCopy mapping when they are active. Depending on the current copy state, read operations might need to be redirected or write operations might cause a copy of a portion of the VDisk. The significant consideration factor is the write operation load on VDisks coming from servers. These write operations are the primary direct cause for the load on target VDisks and their underlying infrastructure consisting of MDisks, back end controllers and volumes, the storage network, and the long distance interconnect.

Consider the following generic metrics for all volumes that need to be sources in a Remote Copy relationship or a FlashCopy mapping.

- ▶ Write I/O rate on application volumes.
- ▶ Write data rate on application volumes.
- ▶ Total I/O rate on existing backend volumes.
- ▶ Total data rate on existing backend volumes.

The goal of observing these metrics is to determine the aggregated application write load and the remaining I/O load capacity of the existing storage subsystems.

For Remote Copy environments, you need to take the aggregated peak workloads of your application volumes over the full production cycle into consideration for sizing. This includes, for example, monthly batch processing, backup workload, high volume business intelligence tasks, and so forth.

For FlashCopy environments, consider the workload level around the times at which you intend to take the FlashCopy and the additional time needed for background copy when applicable.

The use of Remote Copy and FlashCopy does not only generate write load on target volumes, it can also generate additional read load on the source volumes. Gathering the total I/O rate and data rate on the existing backend volumes can help you in assessing whether they can handle additional load.

- ▶ Small write on source causes un-copied grain to copy (256k read from source).
- ▶ Background copy causes 256k reads (grain) (RC, FlashCopy).

- Read on FlashCopy target causes read on source when not copied.

4.3.2 Design guidelines

The application write loads, determined from the performance data observations, make up the base load component that the target storage subsystem and the SVC cluster interconnect network need to handle. Furthermore, consider the load components when sizing the bandwidth of the long distance interconnect and the I/O and data rate capabilities of the storage subsystems.

The background copy process adds the bandwidth and I/O load that you need to consider. This load appears both during initial synchronization and re-synchronization events of Remote Copy relationships and after the start of a FlashCopy mapping. The background copy rates are configurable, and you need to balance your objective for time to complete the background copy with the available bandwidth and remaining storage system I/O capacity.

You also need to consider the impact of the loss of interconnect or subsystem component redundancy. Your designed solution needs to provide the required bandwidth and I/O capacity, even under these conditions if you aim for a highly reliable and continuously available solution.

The performance balance between the primary and secondary subsystems of a Remote Copy solution needs to be considered. The secondary subsystem needs to satisfy the expected total I/O and data rate of Remote Copy writes with low response times. Section 4.4.1, “The secondary subsystem of a Remote Copy” on page 136 discusses a scenario concerning secondary subsystem performance.

In designing the solution, keep the growth of I/O and bandwidth demand in mind. Plan for growth when sizing the SVC cluster interconnect bandwidth. The balance consideration from primary to secondary storage subsystem must include a growth strategy that allows the possibly smaller secondary storage subsystems to keep up with the growth capabilities of a primary storage subsystem.

The design of the secondary SVC and storage subsystem configuration for Global Mirror needs to be particularly diligent. In a Metro Mirror configuration, the I/O workload put on the primary SVC is synchronously coupled to the secondary's performance and thus in a sense self regulating. However, in a Global Mirror configuration, a long response time from the secondary system cannot be tolerated for an extended period of time without causing the relationship to suspend (link tolerance threshold) and thus should be avoided by design.

Consider the following measures to facilitate reliable performance at the secondary system particularly in Global Mirror environments:

- ▶ Dedicated MDisk group

Use a dedicated MDisk Group for Global Mirror secondary VDisks. This helps to ensure that the I/O on other non-Global Mirror VDisks cannot overload MDisk and impact Global Mirror operation.

- ▶ Dedicated controller for Global Mirror MDisk

Use a dedicated storage system for Global Mirror MDisk. This helps to ensure that overloading other MDisk on that controller does not impact Global Mirror MDisk and VDisks. Examples for possible causes are cache monopolization, controller back-end saturation, or saturation of a shared RAID array.

Note: High-end controllers must provide good enough Quality of Service characteristics. The SVC also has specific algorithms in its I/O handling designed to reduce the impact of other MDisk's overload condition on Remote Copy operations.

- ▶ Measure WAN link latency under load

Determine latency of long-distance IP connection when operating at maximum required bandwidth. For example, the latency for an FCIP link measured under load might be higher than that of a **ping** test on an otherwise unused IP connection. You might, for example, use FTP file transfers to generate a sizeable load on the IP WAN link, and measure the response time by using a concurrent **ping**¹ or **hping2** test.

- ▶ Monitor performance

Use TPC or another adequate system to collect performance metrics and monitor those relevant to Global Mirror. The detailed collected data in TPC can also be a valuable source of information to the IBM support representatives helping you in a support situation. See section 4.4.5, "Identifying Global Mirror overload" on page 144.

- ▶ Use "real world" load to test.

Use real application workloads for testing Remote Copy and FlashCopy scenarios. Artificial workloads might miss or over emphasize specific workload aspects and cause over or under exercise of performance relevant aspects of your test-case configuration. For example, running load-generators such as IOMeter, instead of real-world applications such as

¹ Keep in mind that certain packet prioritization or other optimizations in your network might skew your results. Your IBM representative can assist you in WAN link measurements.

databases, mail servers, and so on, might result in markedly different I/O patterns, higher I/O concurrency, and latency sensitivity.

- ▶ Increased load on SVC nodes—in other words, you might need to add an I/O group. (Up to double observed under artificial case w/ 100 percent write and > 500 DDM on single I/O group. 2 I/O groups with Global Mirror then outperformed single I/O group without.)

4.3.3 TPC-based SVC performance reporting for Remote Copy sizing

We discuss relevant performance reporting data for Remote Copy sizing in the sections that follow.

Server load

SVC performance statistics present a good base for gathering the data needed on server write load characteristics as discussed in section 4.3.1, “Key metrics” on page 131. The use of TPC and SVC performance statistics offers a centralized approach to obtaining the server write performance metrics at a VDisk level. We recommend that you focus your attention to the following TPC SVC Volume metrics for all relevant VDisks:

- ▶ Write I/O Rate (overall)
- ▶ Write Data Rate

The write I/O rate and write data rate measure all write operation loads for VDisks presented by servers to the SVC SCSI Front End layer and Remote Copy layer². It represents the load that needs to be processed by Remote Copy operations and needs to be absorbed at the secondary SVC. It therefore represents the server load component for sizing the bandwidth of the SVC interconnect and I/O capacity for the secondary back end subsystem.

Note: We believe the use of these metrics is adequate to help you understand the server write load relevant for Remote Copy sizing purposes for most of the general workloads. However, for workloads with unusual cache interactions, additional metrics such as *Cache to Disk Transfer Rate* and *Dirty Write Percentage of Cache Hits* can give you additional insight into the required secondary subsystem I/O capacity. Your IBM representative can assist you in assessing those special cases.

Subsystem I/O capacity

For the purpose of assessing the utilization and remaining I/O capacity of the subsystem underlying the relevant VDisks, you can use the MDisk and MDisk Group metrics reported for SVC. Filter the output to the MDisk or MDisk Groups

² See 3.3.2, “Remote Copy layer” on page 70, for an overview of the SVC I/O filter stack.

that are relevant to the VDisks you want to use with Remote Copy or FlashCopy. The following metrics are most useful to compare to the designed I/O and data rate capacity of your subsystems:

- ▶ Back end Read I/O Rate
- ▶ Back end Write I/O Rate
- ▶ Total Back end I/O Rate
- ▶ Back end Read Data Rate
- ▶ Back end Write Data Rate
- ▶ Total Back end Data Rate

In addition to I/O and data rate metrics, the observation of average Back end Read Response Time and average Back end Write Response Time together with comparisons of these values to previous measurements or measurements of other subsystems can help you in assessing the loading factor of the subsystem. If the storage subsystem receives load from sources other than SVC, you need to take this additional load into account when assessing the utilization and remaining I/O capacity of the subsystem. You can use the TPC metrics for the respective subsystem to directly observe the overall load and to help you in the assessment process.

In addition to interactive reporting and analysis of performance data using the GUI, TPC offers a command line interface (TPCTOOL) that allows you to automate the process of exporting multiple performance metrics data series into a delimiter separated file. This file can be further processed, aggregated, and graphed using a spreadsheet application. The *Reporting with TPCTOOL*, REDP-4230 describes the process of obtaining performance data in this fashion and is a recommended read.

4.3.4 Subsystem-based performance reporting before introducing copy services

In cases where the SVC is not yet implemented, use the TPC for Disk to collect the performance metrics for your disk subsystem. Depending on your disk subsystem those metrics can be quite basic.

If you do not have TPC in your environment we recommend you use whatever disk performance tool is available with your disk subsystem, for example DS4000 Performance Monitor. For guidance consult the *DS4000 Best Practices and Performance Tuning Guide*, SG24-63632 at the following Web address:

<http://www.redbooks.ibm.com/abstracts/sg246363.html?open>

4.3.5 Host-based performance reporting before introducing copy services

You can monitor performance from the host side using monitoring tools such as IOMeter, iostat, and sar to collect performance stats. However, manual parsing and aggregation is necessary to find overall peaks.

4.4 Monitoring SVC copy service performance and identifying issues

After SVC copy services are established in a Business Continuity solution environment, the load metrics used in the design phase continue to be relevant as the key indicator for copy services load. However, additional SVC performance metrics become active and can be used for reporting. These metrics can help as an indicator for balanced operation of the SVC copy service operations. The following sections focus on specific scenarios in which the copy services infrastructure consisting of SVC, back end controllers and volumes, and possibly the storage network interconnection are driven out of balance. We present reporting metrics that can help you investigate the areas of interest in these situations.

A prerequisite for the investigation of developing performance issues is the continuous collection of performance data. This establishes a baseline of all performance metrics under normal operating conditions and allows for comparison and analysis of deviations should a performance issue arise. You need to balance the TPC performance monitoring interval for data collection between finer grained statistics available when using a short interval and the larger amount of storage space needed for the TPC database to hold the performance data. For a demanding Global Mirror environment, we recommend considering a short interval (five minutes) for better analysis opportunities in cases of performance related issues.

In addition to storing performance data as a baseline, you can also use each collected sample for checking certain metrics against user-defined constraints and report on constraint violations. For more on TPC features see section 4.4.6, “Using TPC alerts” on page 147.

4.4.1 The secondary subsystem of a Remote Copy

This is a case of a server using multiple VDisks in a Metro Mirror Consistency Group and copying data to a remote SVC cluster. The server experiences unsatisfactory I/O performance on the VDisks prompting an investigation of the issue at the storage level.

A review of IBM support requests revealed that the secondary subsystem is often initially overlooked as a cause for a performance degradation on the primary VDisk. The target subsystem must be capable of absorbing the full write rate at a low response time, because the synchronous nature of Metro Mirror exposes the server writing to the primary VDisk to the remote response times. The issue becomes most pronounced after the secondary SVC's cache for the secondary VDisks overflows due to the inability of the secondary storage subsystem to accept the sustained high write rate driven by the server. This causes the secondary subsystem response time to become visible at the remote SVC's secondary VDisk level and consequently at the primary VDisk level.

The following TPC SVC metrics can help you identify this situation. A marked increase in the following response time metrics can point to a problem at the remote site:

- ▶ Write Response Time at local SVC VDisk
- ▶ Port to Remote Node Response Time at local SVC
- ▶ Back end Write Response Time at remote SVC MDisk or MDisk Group

The increase of MDisk's response time by more than 50ms or to a level in excess of 100ms indicates a problem. Depending on the type of the secondary subsystem, TPC might offer additional performance metrics allowing you to further investigate the subsystem controller or back end level for overload indications.

The root cause for this situation can have a variety of origins. The following examples highlight frequently found reasons:

- ▶ High server I/O load
Growing I/O demand of the server using Metro Mirror exceeds the designed capacity of the secondary subsystem.

- ▶ Background copy
A Metro Mirror relationship is in the process of re-synchronizing the secondary to the primary data at a background copy rate overloading the secondary subsystem. Reduce the background copy rate or perform the operation during a period of lower operational I/O demand.

- ▶ Maintenance or defect
The secondary subsystem or SVC are subject to a maintenance operation or a defect that reduces their I/O capacity. Examples are events such as a microcode update, a disabled cache due to loss of redundancy, or a RAID rebuild.

- Concurrent workload

Additional I/O load is put on the secondary subsystem environment, for example, by servers at the remote site or by the remote SVC performing a VDisk migration.

The interaction between an overloaded secondary system and Global Mirror exhibits different behavior and is discussed in section 4.4.5, “Identifying Global Mirror overload” on page 144.

4.4.2 FlashCopy

The FlashCopy feature of SVC creates a point-in-time copy from a source VDisk to a target VDisk within a SVC cluster. The first write operation on a block of the source VDisk causes the grain (256KB segment of VDisk space) containing it to be copied to the target VDisk (grain split). SVC also uses a background process to split grains at a user configurable rate if desired. Read operations addressing unsplit grains on the target VDisk are redirected to the source VDisk.

The copy-on-write process can introduce significant latency into write operations. To isolate the host application that issued the write I/O from this latency, the FlashCopy indirection layer is placed logically below the SVC cache layer. This means that the copy latency is seen typically only on a destage from the cache rather than for original write operations from the host application that otherwise might be blocked waiting for the copy operation to complete. However, the latency becomes visible if the underlying storage subsystem of the target VDisk is unable to absorb the copy rate driven by host write operations and background copy.

Consider the following example containing an imbalance in the I/O capacity between the source and target VDisks that can cause a performance impact to the application.

The source VDisks are defined in an MDisk group backed by a high-end storage subsystem with an adequate I/O capacity for the I/O load driven by the server. The source VDisk is in a FlashCopy mapping with a target VDisk backed by a midrange storage controller with storage capacity optimized disk drives and limited I/O capacity. A FlashCopy is started and held active while the application workload continues to run. The server performs write operations at a significant rate into the SVC cache that gets destaged independently based on cache resource pressure. The source VDisk's back end subsystem can process the additional read load caused by the read operations that occur in the grain split copy process during destage. However, the target subsystem cannot absorb the grain writes at the rate that SVC intends to destage data from its cache. This can cause the write cache for the source VDisk to fill up over time if the write load continues at its high level. The time needed of grain splitting, which is at that

point dominated by the write response time of the target VDisk, becomes visible at the source VDisk level.

This overload situation for the target VDisk's subsystem is the cause for the performance impact seen at the source VDisk as well as the high cache resource demand for the source VDisk. Due to the constrained destage process, it puts pressure on other VDIs in the I/O Group that contend for cache memory. Examining the following SVC performance metrics can help you to identify and investigate this issue:

For source VDisk:

- ▶ Read Response Time
- ▶ Write Response Time
- ▶ Write I/O Rate

The performance issue is most likely discovered by a marked increase in the Write Response Time of the source VDisk. The response time is otherwise expected to be small when writes operations can be satisfied by SVC cache. The issue is likely to have little noticeable effect on the Read Response Time, because the source back end subsystem fulfilling the read requests is not considered to be overloaded in this scenario.

For the source VDisk's underlying MDIs (MDI Group):

- ▶ Back end Read I/O Rate
- ▶ Back end Read Response Time
- ▶ Back end Read Queue Time
- ▶ Back end Write I/O Rate
- ▶ Back end Write Response Time
- ▶ Back end Write Queue Time

You can use these and further metrics to validate the assertion that the source VDisk's back end storage subsystem is not the origin of the observed performance issue and that it is operating within its designed I/O capacity.

For the target VDisk's underlying MDIs (MDI Group):

- ▶ Back end Write Response Time
- ▶ Back end Write Queue Time
- ▶ Back end Total I/O Rate
- ▶ Back end Total Data Rate

You can use these and possibly further metrics to assess the overload condition on the targets back end storage subsystem. Look for High Back end Write Response Times or Write Queue Times and a Total I/O or Data Rate that is beyond the I/O capacity of the subsystem.

In a situation as described in this section, we recommend that you carefully inspect the source VDisk's Write I/O Rate at the time of the planned FlashCopy. The source and target VDisk's back end subsystems must be able to handle the read and write load caused by the copy activity of grain splits. Every source VDisk write operation that addresses an unsplit grain causes that grain to be copied upon cache destage. Also keep in mind that the background copy rate you defined for a FlashCopy mapping adds grain copy I/O load that you need to balance with the server workload and I/O capacity of the underlying storage subsystems. Where a FlashCopy mapping is experiencing a significant amount of grain splitting due to host write activity, it reduces the background copy rate appropriately to try to maintain the same overall copy rate for that mapping.

4.4.3 FlashCopy of Remote Copy secondary VDisks

Remote Copy master and auxiliary VDisks can be configured as the source in a Flash Copy mapping. This allows you to FlashCopy VDisks that are part of a Remote Copy relationship.

The additional load to the back-end storage subsystems during FlashCopy operation is dominated by the write operation in the I/O workload of the participating VDisks. Remote Copy target VDisks are a special case for FlashCopy because their I/O workload consists only of write operations. This section considers FlashCopy performance aspects that are more likely to be seen when the FlashCopy source VDisk is an active Remote Copy secondary VDisk. There are two areas of interest in this context:

- ▶ FlashCopy preparation phase (prestart)
- ▶ FlashCopy start and copying phase

FlashCopy preparation phase

FlashCopy uses a prepare phase (prestart) to establish a synchronization point to take the point-in-time copy (start). During the preparation phase, SVC sets the cache for the FlashCopy source VDisk in write-through mode and flushes all its modified data held in cache onto the underlying storage subsystem.

To reach the synchronization point quickly, the cache flush operations compete with other cache destage operations on equal terms, and neither takes precedence over the other. The storage subsystem must be able to absorb this increase in the write load in the FlashCopy preparation phase, because the additional write response time of the storage subsystem is visible at the VDisk level and consequently to a MetroMirror primary VDisk during that time. The impact on the storage subsystem depends on its remaining I/O capacity and the amount of modified data held in the cache at the time you initiate the preparation

phase. The following areas of interest can help you to estimate the relative amount of modified data to be flushed and assess potential impact:

- ▶ Recent I/O history

Consider the recent history of I/O operations from the server to the Remote Copy primary VDisk, and thus the write operations to the secondary VDisk. If the write rate is low prior to FlashCopy preparation, it is likely that fewer modified data is in cache.

- ▶ Concurrent I/O workload on secondary SVC

Consider the I/O workload of other VDisks on the secondary SVC competing for cache resources. If other locally used VDisks or other Remote Copy target VDisks that are not prepared at the same time have I/O activity, it is likely that a smaller amount of cache memory is allocated to the FlashCopy source VDisk in question.

- ▶ FlashCopy Consistency Group size

Consider all VDisks in a Consistency Group that are being prepared, because all FlashCopy mappings in that Consistency Group are being prepared at the same time.

- ▶ Back-end storage subsystem cache size

Consider the size of the backend storage subsystem cache in relation to the additional write load. A larger write cache might help the storage subsystem to more easily cope with the short increase in write operations.

Note: We do not expect the effects of FlashCopy related cache flush to appear to a significant extent under mixed workloads and typical SVC usage scenarios. The effect does not apply to Remote Copy relationships that are stopped prior to preparing the FlashCopy because the cache is likely to hold little modified data after a short period of time, and the primary VDisk is decoupled from the effects on the secondary VDisk.

A common application is creating a FlashCopy of a stopped Remote Copy secondary VDisk that is in a consistent state. This allows you to retain a consistent copy of your data during the phase of re-synchronization and inconsistency after you restart the Remote Copy relationship. A loss of the primary VDisk during that phase can leave you without a consistent copy of your data on the secondary SVC. The FlashCopy performance aspect discussed in this section is not likely to be seen in the common context just described, because the FlashCopy source VDisk is not actively being written.

FlashCopy copying phase

Write operations on a FlashCopy source VDisk cause a grain split whenever they address a grain that is not yet copied. The RemoteCopy target VDisk, which is the FlashCopy source in this scenario, receives only write operations and thus causes a disproportionate amount of grain splits when compared to a VDisk with a general server I/O workload.

The performance considerations for using FlashCopy that we discussed in section 4.4.2, “FlashCopy” on page 138 apply also to a Remote Copy secondary VDisk actively being written to. The back end storage subsystems of the FlashCopy source and target VDisks need to be able to accept the load caused by high grain split activity and potential FlashCopy background copy. Under subsystem overload conditions, the FlashCopy source VDisk might experience long write response times that can then become visible at the Remote Copy primary VDisk as described in 4.4.1, “The secondary subsystem of a Remote Copy” on page 136.

4.4.4 Identifying Global Mirror colliding write operations

Global Mirror is designed to preserve a consistent copy at the secondary VDisks at all times. The Global Mirror algorithm therefore needs to consider the observable order of write operations. The implementation used in the SVC as described in Chapter 3, “SAN Volume Controller Mirroring Solutions” on page 51 Global Mirror needs to synchronize certain concurrent or near concurrent writes to the same block (512 bytes) of a VDisk (colliding writes). This synchronization with the secondary SVC introduces latency to the affected write operations in the order of one round-trip delay and causes a response time impact to the primary VDisk.

Note: Current observations with a variety of workloads, including database, show little or no adverse effect on critical path application activity. However, this is based only on a limited number of development and early shipment installations. We therefore recommend collecting statistics data on the occurrence frequency of colliding writes.

Colliding write operations are represented in TPC in two *Global Mirror Overlapping Write* metrics. Global Mirror Overlapping Write I/O Rate measures the number of colliding write operations over time in a VDisk level. Global Mirror Overlapping Write Percentage measures, for each primary VDisk, the colliding write operations as a fraction of all Global Mirror write operations to a secondary VDisk.

Figure 4-11 on page 143 shows an example TPC graph of the Global Mirror Overlapping Write Percentage for the I/O Group that is performing Global Mirror on two VDisks loaded with an artificial write heavy random access pattern.

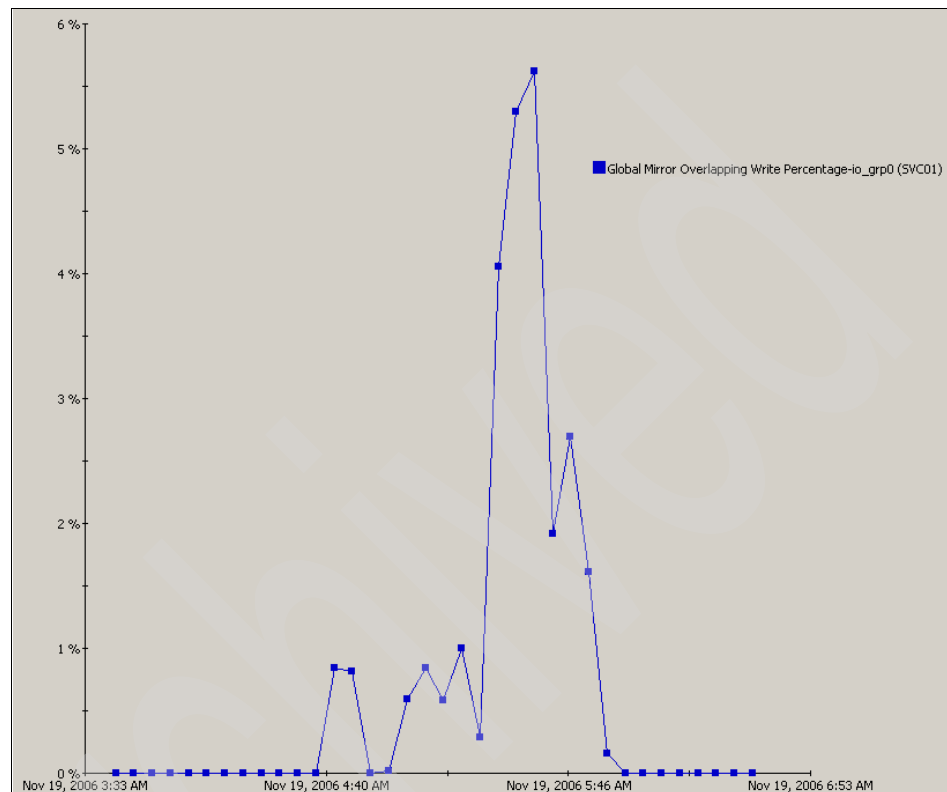


Figure 4-11 Global Mirror Overlapping Write Percentage example graph

We see the colliding write operation percentage stays very low during the initial phase of the increasing workload. During the high load phase, the graph shows about one percent of writes colliding, while this number increases to about five percent in the overload phase. This increases the latency of host write operations by an amount in the order of the inter cluster round-trip time. However, we cannot distinguish that component of the write latency from the apparently dominating factor of the back end response time in the order of hundreds of milliseconds seen at saturated storage subsystem.

The graph is intended to show that colliding writes can occur in a workload. However, the extent of colliding writes is greatly dependent on the write access pattern of the application. Furthermore, the potential impact of colliding write operations on application performance is dependent on the application's sensitivity to occasional increased write response times when re-writing the

same block. The recommendation to collect colliding writes statistics together with performance metrics is meant to help you in establishing baseline data for VDisks with different kinds of workload patterns. This might help to establish or rule out correlations between performance impacts and increased colliding write for particular workloads when investigating performance issues. It also emphasizes again the need to use realistic application workloads when testing Global Mirror performance. Artificial workloads might over or under estimate specific performance aspects present in the perspective real workload.

We suggest that you investigate situations of more than three percent colliding writes when the server perceives a performance problem. For VDisks with more than 10 percent colliding writes, we recommend you investigate potential performance impacts and consider possible further investigation and avoidance at the application level.

4.4.5 Identifying Global Mirror overload

This section focuses on possible problem areas and overload situations that can cause an impact on operating a Global Mirror environment. These overload situations in the SVC environment can cause the most heavily loaded Global Mirror relationships to stop and the SVC cluster to present a cluster error 1920.

Among the possible reasons to stop a Global Mirror relationship, the following performance related areas of interest must be investigated. The list is ordered by probability of the cause:

- ▶ SVC cluster interconnect bandwidth overload
- ▶ SVC cluster interconnect latency issue
- ▶ Secondary storage subsystem overload
- ▶ SVC node overload

The listed performance issues causing Global Mirror to stop are related to the inability of the primary cluster to perform Global Mirror write operations to the secondary cluster in a timely fashion. To protect the primary VDisks from exposing the deteriorating I/O performance to the application level, SVC stops the relationship when the user's configuration time of Global Mirror link tolerance time-out (gmlinktolerance).

Secondary storage subsystem overload

One of the probable overload causes of stopping Global Mirror relationships is the inability of the secondary storage subsystem to absorb the Global Mirror write rate. This situation is similar to the Metro Mirror situation described in section 4.4.1, "The secondary subsystem of a Remote Copy" on page 136. However, while it causes a performance impact for the time of the overload issue of the secondary system in the Metro Mirror case, it also causes Global Mirror

relationships to stop after the Global Mirror link tolerance time-out. The underlying possible reasons and areas of investigation for the overload causes are the same as in the Metro Mirror case and can be approached the same way.

SVC cluster interconnect bandwidth overload

This situation is characterized by a primary VDisk write data rate and consequently Global Mirror write data rate that exceeds the available bandwidth of the SVC cluster long distance interconnect. However, the lack of interconnect bandwidth is limiting the primary SVC's ability to transport Global Mirror data to the secondary SVC cluster. This causes a performance impact on the primary server and eventually stops the Global Mirror relationship.

The case of a saturated link can be identified by analyzing the performance statistics recordings of the time period around the stop time of the Global Mirror relationship for the following three indicated patterns:

Note: The following identification patterns assume that the available interconnect bandwidth at the time of the Global Mirror stop is known and is equal to the maximum interconnect bandwidth. This might not be the case in environments with a shared bandwidth interconnect or reduced bandwidth due to a loss of redundant links. Establishing the facts about the bandwidth available to SVC Global Mirror at the time of the stop event is outside the scope of this book.

1. You can observe that the write data rate of the *secondary* VDisks is close to the maximum interconnect bandwidth for a time period *before* the Global Mirror stop. This duration of time corresponds to the Global Mirror link tolerance. You can for example use a TPC performance report on the secondary SVC's MDisk Group that contains all secondary VDisks presenting the Write Data Rate metric. Including the further metrics Write I/O rate (overall), Backend Write Data Rate, and Backend Write Response Time to the report allows you to assess the back end controllers I/O behavior at the same time.
2. If you can observe that the write data rate of the *primary* VDisks is *not close* to the maximum interconnect bandwidth for a time period *before* the Global Mirror stopped, then your server write workload is not likely to be the saturating factor. The higher Global Mirror write data rate is most probably caused by background copy activities. Reduce the background copy bandwidth parameter for SVC cluster mirror partnership configuration so that the total Remote Copy data rate is within the interconnect's bandwidth capabilities. Alternatively, schedule background copy activity during times of lower Global Mirror write bandwidth demand.

3. If you can observe that the write data rate of the *primary* VDisks is *close* to the maximum interconnect bandwidth for a time period *before* the Global Mirror stopped, and marked *increases* after that, then your server write workload is likely to be the saturating factor. The higher write data rate shows the real I/O workload the servers attempted to perform, but the interconnect is unable to sustain. You can for example use a TPC performance report on the primary SVC's VDisks presenting the Write Data Rate metric to learn more about the actual bandwidth requirements of your servers at that time. Including further metrics like Read Response Time, Write Response Time, Write I/O rate (overall), Backend Write Data Rate, and Backend Write Response Time to the report can help you to see what response time impact your servers experienced during the overload period and how your primary back end storage operated at that time.

SVC cluster interconnect latency issue

This situation is concerned with the interconnect between the SVC clusters exhibiting an unacceptable amount of latency. This can, for example, be caused by a latency increase of an IP WAN connection due to high load or failure of the primary, short path. High interconnect latency is also seen when the interconnect bandwidth has exceeded, as described in the previous section. The Send Queue Time in particular might increase in this circumstance. This can be used as a symptom to diagnose a bandwidth issue where the interconnect bandwidth capability is not known or not certain. A bandwidth issue can manifest itself as high latency and an associated bandwidth drop. The reduced bandwidth is the result of what the interconnect can achieve. Where the bandwidth achieved is lower than expected as predicted from historical data, there is a high latency. The issue of degraded bandwidth must be taken up with the interconnect provider.

We recommend collecting performance statistics on the following metrics on a SVC node level:

- ▶ Port to Remote Node Send Data Rate
- ▶ Port to Remote Node Send Response Time
- ▶ Port to Remote Node Send Queue Time

The sum of Port to Remote Node Send Response Time and Port to Remote Node Send Queue Time are an indicator for the ports effective perceived round-trip latency to the secondary SVC. We recommend investigating the cause if the total time is observed having an average of over 80ms per operation, because this can cause errors in the Remote Copy communication.

We observed that the Port to Remote Node Send Response Time metric presents an elevated average response time at the order of 10ms per operation under very low load or idle SVC cluster relationship conditions due to one-time

effects. We suggest to qualify the relevance of the Port to Remote Node Send Response Time metric for the SVC cluster interconnect latency assessment by checking for a Port to Remote Node Send Data Rate of more than 200KB/s.

SVC node overload

The SVC node allocates internal processing time to perform I/O operations, Remote Copy, Flash Copy, and other services. Newer SVC engine models generally provide more processing capacity to support higher I/O and Copy Service performance potential. Global Mirror processing is one of the rather CPU time intensive services for an SVC node. Under very high Global Mirror workloads, SVC nodes serving primary or secondary VDisks can become possible contributors to long Global Mirror response times and stop Global Mirror relationships.

SVC records CPU utilization on a per node level as a direct measurement of the amount of processing performed by a node. Further performance metrics can serve as indirect indicators for high internal processing load on an SVC node. For the purpose of investigating Global Mirror overload situations, we recommend collecting statistics data on the following metrics on a node level:

- ▶ CPU Utilization
- ▶ Port to Local Node Message Send Time
- ▶ Port to Local Node Message Queue Time

High CPU load in an SVC node causes additional small latencies in a variety of processing steps. This becomes more visible for indirect measurement in timing metrics that have very low values under normal operating conditions and have little dependency on external latencies outside SVC. The sum of Port to Local Node Message Send Time and Port to Local Node Message Queue Time can be used to point at high internal processing load suggesting high Global Mirror I/O load. Average values of more than 1ms per operation on a SVC node level for the sum of the two metrics or a CPU utilization in excess of 50 percent can indicate a SVC node load condition contributing to a Global Mirror stop event; otherwise, contact your IBM support representative for further assistance.

4.4.6 Using TPC alerts

When operating a demanding Remote Copy or FlashCopy environment, you need to be aware of the overall condition of your copy services infrastructure. The previous sections described scenarios and situations that went out of balance and might have been detected earlier with the help of indicators for changing conditions. In addition to the alert and status notification provided by SVC through SNMP, TPC can help to detect changes in load conditions before they cause performance issues or frequent Remote Copy relationship suspension.

TPC offers an alerting system to help you take notice of conditions or events in your environment. The system is based on threshold definitions and constraint violation reports. The defined constraint conditions check against every collected sample of performance data, and violations are recorded for later reporting. Note that you can configure TPC so that it does not store all collected performance data into its database. This allows you to collect performance data and receive constraint violation alerts in five minutes, while storing the collected data in 15 minutes to reduce long term database storage space requirements.

We only show a small aspect of TPC alerting in this IBM Redbooks publication, namely the performance related storage subsystem alerts with triggering conditions related to SVC copy services applications. You can find a discussion considering further aspects of the TPC alerting system in *IBM TotalStorage Productivity Center V3.1: The Next Generation*, SG24-7194.

Creating a Storage Subsystem Alert

1. Use the context menu function **Storage Subsystem Alert** from the Navigation Tree option **Disk Manager** → **Alerting** → **Storage Subsystem Alerts** to create a new alert as shown in figure Figure 4-12.

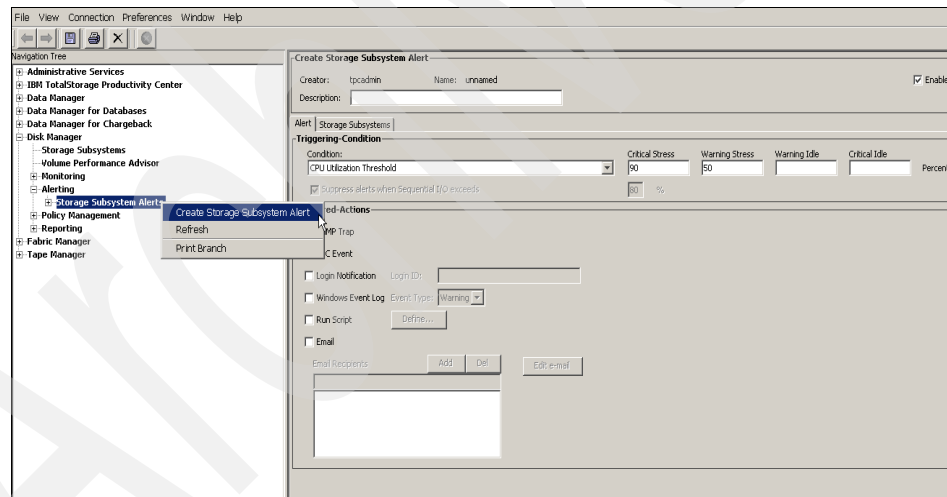


Figure 4-12 Creating a Storage Subsystem Alert

The thresholds are grouped into categories: stress, idle, levels warning, and critical as shown in the figure. Stress thresholds trigger when exceeded and are used here to indicate potential overload conditions. The idle level thresholds trigger when the condition value is lower then the defined values. We leave the idle threshold fields empty indicating no checking of this condition.

2. After you select an alert condition, enter the threshold values and select the triggered actions.
3. Continue on the Storage Subsystems tab to select the storage subsystems that you want to be subject to the threshold alerts. Save the alert using the **File** → **Save** option.

Table 4-1 shows the available threshold conditions for SVC in TPC.

Note: The presented threshold conditions apply to TPC 3.1.3 as used in preparing this book. See Appendix F of the *IBM TotalStorage Productivity Center User's Guide*, GC32-1775, for the current list applicable to your TPC release.

Table 4-1 TPC alert thresholds for SVC

Threshold (Metric)	Component type	Description
Total Backend I/O rate	SVC MDisk group	Sets thresholds on the average number of I/O operations per second for MDisk read and write operations for the MDisk groups. The Total I/O Rate metric for each MDisk group is checked against the threshold boundaries for each collection interval.
Total Backend Data Rate	SVC MDisk group	Sets thresholds on the average number of MB per second that are transferred for MDisk read and write operations for the MDisk groups. The Total Data Rate metric for each MDisk group is checked against the threshold boundaries for each collection interval.
Total I/O Rate (overall)	SVC I/O group	Sets threshold on the average number of I/O operations per second for read and write operations for the subsystem controllers (SVC clusters) or I/O groups. The Total I/O Rate metric for each controller or I/O group is checked against the threshold boundaries for each collection interval.
Overall Backend Response Time	SVC MDisk	Sets thresholds on the average number of milliseconds that it takes to service each MDisk I/O operation, measured at the MDisk level. The Total Response Time (external) metric for each MDisk is checked against the threshold boundaries for each collection interval.

Threshold (Metric)	Component type	Description
Total Data Rate	SVC I/O group	Sets threshold on the average number of MB per second for read and write operations for the subsystem controllers (SVC clusters) or I/O groups. The Total Data Rate metric for each controller or I/O group is checked against the threshold boundaries for each collection interval.
Total Port I/O Rate	SVC port	Sets thresholds on the average number of I/O operations or packets per second for send and receive operations for the ports. The Total I/O Rate metric for each port is checked against the threshold boundaries for each collection interval.
Total Port Data Rate	SVC port	Sets thresholds on the average number of MB per second for send and receive operations, for the ports. The Total Data Rate metric for each port is checked against the threshold boundaries for each collection interval.
CPU utilization	SVC node	Sets threshold on the average percentage of CPU utilization for the storage subsystem controllers (SVC clusters). The CPU Utilization metric for each SVC node is checked against the threshold boundaries for each collection interval.
Write-cache Delay Percentage	SVC VDisk	Sets threshold on the average percentage of cache write operations that are not processed in fast-write mode, measured at a VDisk level. The Write-cache Delay Percentage metric for each VDisk is checked against the threshold boundaries for each collection interval.

Suggested threshold metrics

The appropriate threshold metrics and threshold values are dependent on your environment and your monitoring requirements. However, based on the scenarios and areas of interest discussed above, we suggest you consider the thresholds in Table 4-2 on page 151 as a starting point to find the desired alerting levels for your specific configuration.

Table 4-2 Suggested threshold metrics

Threshold (metric)	Critical stress	Warning stress
Overall Backend Response Time	100	50
SVC CPU utilization	90	50
Write-cache Delay Percentage	10	3

The threshold on back end response time can help you to identify highly loaded MDisk and storage subsystems. Elevated SVC CPU utilization levels are of particular interest in Global Mirror environments because of their potential impact on Global Mirror operation latency and the link tolerance threshold functions algorithm that attempt to identify such impacts and prevent them from escalating into application impact.

Elevated SVC CPU utilization levels are of general interest in non-Global Mirror environments because high levels indicate there may be an opportunity to improve overall system performance by adding nodes to the cluster or by re-distributing the workload among the existing nodes.

The Write-cache Delay Percentage metric is primarily targeted at the cache effects during FlashCopy activities. The alerts can also be an indication of a storage subsystem unable to absorb the required write load causing the SVC cache to fill up.

Constraint violation report

The violation of defined threshold is recorded in the TPC database with each collected performance data sample.

Use the following steps to generate a report that summarizes the constraint violation frequency for enabled threshold metrics.

1. You can get to the dialog to generate the report by selecting **Disk Manager** → **Reporting** → **Storage Subsystem Performance** → **Constraint Violations** (see Figure 4-13 on page 152).

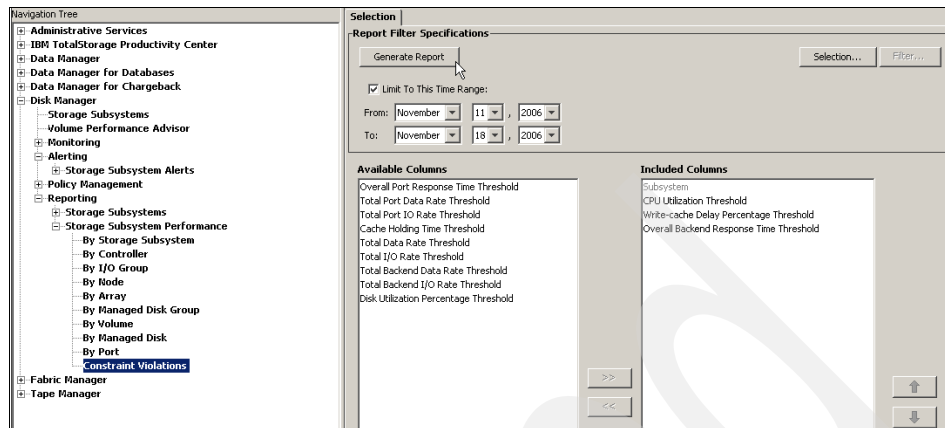


Figure 4-13 Generating a constraint violation report

Figure 4-14 shows an example constraint violation report for two SVC clusters with a subset of threshold metrics.

Selection Constraint Violations			
Storage Subsystem Performance: Constraint Violations			
Number of Rows: 2			
Subsystem	CPU Utilization Threshold	Write-cache Delay Percentage Threshold	Overall Backend Response Time Threshold
SVC01	0	84	35
SVC02	0	0	5

Figure 4-14 Constraint violation report example for SVC

Click the drill-down icon to get a detailed report for the selected subsystem. Figure 4-15 on page 153 shows a detailed report for subsystem SVC01 that presents a number of constraint violations for Overall Backend Response time and Write-cache Delay Percentage.

Note: The unusually high Write-cache Delay Percentage values shown in Figure 4-15 are manufactured by intentional SVC misconfiguration and an artificially skewed workload.























Selection Constraint Violations SVC01									
Storage Subsystem Performance: Constraint Violation Details									
Number of Rows: 123									
Time ▲	Component	Metric	Measured Value	Type	Critical Stress	Warning Stress	Critical Idle	Wa	
 Nov 17, 2006 11:36:40 AM	mdisk0	Overall Backend Response Time	25.2	Warning Stress	40	20			
 Nov 17, 2006 11:36:40 AM	mdisk1	Overall Backend Response Time	25.7	Warning Stress	40	20			
 Nov 17, 2006 12:01:45 PM	n1	Write-cache Delay Percentage	69.142	Critical Stress	10	3			
 Nov 17, 2006 12:01:45 PM	n2	Write-cache Delay Percentage	70.137	Critical Stress	10	3			
 Nov 17, 2006 12:06:46 PM	n1	Write-cache Delay Percentage	75.518	Critical Stress	10	3			
 Nov 17, 2006 12:06:46 PM	n2	Write-cache Delay Percentage	75.431	Critical Stress	10	3			
 Nov 17, 2006 12:11:47 PM	n1	Write-cache Delay Percentage	75.619	Critical Stress	10	3			
 Nov 17, 2006 12:11:47 PM	n2	Write-cache Delay Percentage	75.133	Critical Stress	10	3			
 Nov 17, 2006 12:16:48 PM	n1	Write-cache Delay Percentage	75.345	Critical Stress	10	3			
 Nov 17, 2006 12:16:48 PM	n2	Write-cache Delay Percentage	75.281	Critical Stress	10	3			
 Nov 17, 2006 12:21:49 PM	n1	Write-cache Delay Percentage	75.642	Critical Stress	10	3			
 Nov 17, 2006 12:21:49 PM	n2	Write-cache Delay Percentage	75.359	Critical Stress	10	3			
 Nov 17, 2006 12:26:50 PM	n1	Write-cache Delay Percentage	75.479	Critical Stress	10	3			
 Nov 17, 2006 12:26:50 PM	n2	Write-cache Delay Percentage	75.346	Critical Stress	10	3			
 Nov 17, 2006 12:31:51 PM	n1	Write-cache Delay Percentage	75.335	Critical Stress	10	3			
 Nov 17, 2006 12:31:51 PM	n2	Write-cache Delay Percentage	75.129	Critical Stress	10	3			
 Nov 17, 2006 12:36:52 PM	n1	Write-cache Delay Percentage	79.851	Critical Stress	10	3			
 Nov 17, 2006 12:36:52 PM	n2	Write-cache Delay Percentage	79.526	Critical Stress	10	3			
 Nov 17, 2006 12:41:53 PM	n1	Write-cache Delay Percentage	76.792	Critical Stress	10	3			
 Nov 17, 2006 12:41:53 PM	n2	Write-cache Delay Percentage	76.44	Critical Stress	10	3			
 Nov 17, 2006 12:46:54 PM	n1	Write-cache Delay Percentage	76.56	Critical Stress	10	3			
 Nov 17, 2006 12:46:54 PM	n2	Write-cache Delay Percentage	76.175	Critical Stress	10	3			

Figure 4-15 Subsystem details of a constraint violation report example

Automation in a Business Continuity solution with SVC

This chapter discusses the possibility of using automation in a Business Continuity environment, the considerations to be made, and gives some examples on how to use automation.

5.1 How and why to automate

When you design a total Business Continuity solution you need to think about many aspects other than IT. A total Business Continuity solution involves everything, not just IT installations, data, and applications. For a production company, it does not help that all data is secure if the production site is destroyed, and the company cannot continue. If all the logistics are not in place it is not always enough that the IT is running and all data is available.

You can find more information about general Business Continuity in the following IBM Redbooks publications:

- ▶ *IBM System Storage Business Continuity: Part 1 Planning Guide*, SG24-6547
- ▶ *IBM System Storage Business Continuity Solutions Overview*, SG24-6684

To ensure a robust Business Continuity solution, automation in some form needs to be used; otherwise, the demand for Business Continuity is not fulfilled. There are different levels where automation can be implemented, but often a fully automated solution is not a good solution. When we talk about a fully automated solution, we think of a solution where not only the first recovery is automated, like in a failover situation, but also that the procedure for going back to the original configuration is done automatically.

Before you decide to go back to the original configuration after a system failover, you need to verify that all is as expected, and verify what data is going to be used after going back to the original configuration. If the fallback situation is handled automatically, you might destroy the data you want to preserve unless you check the exact point that you restore to manually.

When ensuring a Business Continuity solution for an application, consider the likely situations you might encounter so that your applications keep running and are available for users when needed.

Another aspect is also to ensure that you only lose the amount of data, if any, that your business can recover from, so that the business can still continue without high costs. See more on this topic in section 1.3, “Recovery objectives” on page 6.

The biggest challenge is to decide what error situation the Business Continuity solution will cover and how long the application can tolerate being unavailable for. Error situations to consider are server error, network error, storage area network (SAN) error, storage error, a fire or flood situation, and so on.

When a Business Continuity solution is mentioned the impression is that no matter what the error, the application keeps running without data loss. A Business Continuity solution may not cover all kind of errors, but there are

different levels of situations a Business Continuity solution can be implemented to handle successfully.

When deciding to use automation, think about all aspects of what type of errors can happen and what the impact is before you decide what kind of automation you want to implement. Also think about what you want to benefit from the automation. Does the application need to be available 24 hours a day, 7 days a week, or less? Many internet solutions today demand this kind of availability; therefore, the Business Continuity solution in this kind of environment demands a lot of automation, including application and server cluster solutions.

5.1.1 Business Continuity tiers

Business Continuity solutions are available in seven tiers that define what functions and recover time objective (RTO) you have:

- ▶ Tier 1: Restore from tape
- ▶ Tier 2: Hot site-recover from tape
- ▶ Tier 3: Electronic vaulting
- ▶ Tier 4: Point in Time (PiT) disk copy
- ▶ Tier 5: Two phase commit (transaction integrity)
- ▶ Tier 6: Storage mirroring
- ▶ Tier 7: Storage mirroring with automated recovery

To get an understanding of the relationship between the Business Continuity tiers and the RTO look at Figure 5-1 on page 158.

Business Continuity Tiers and Technology

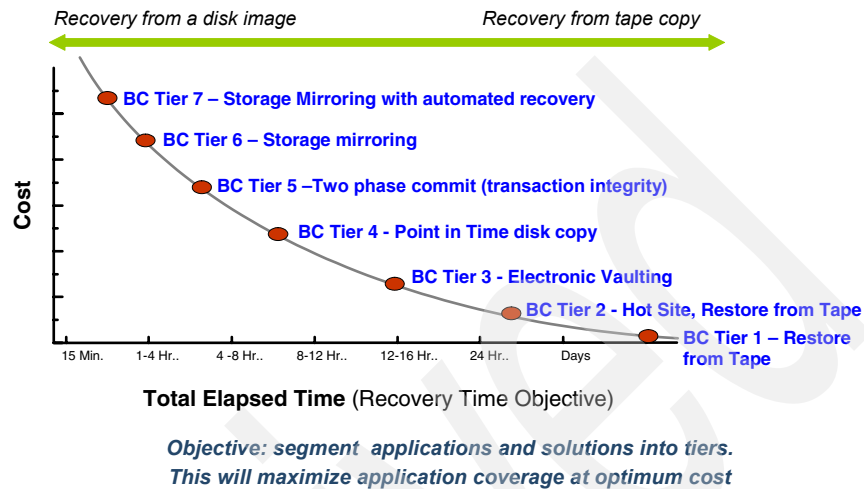


Figure 5-1 Business Tiers and Technology with Recovery Time Objective

Depending on the Business Continuity Tiers, the automation you use can vary. The most complex automation is used in Tier 7, and in Tier 1 the only automation may be the scheduler that handles your backup. The automation you need to implement not only depends on the Business Continuity Tier you implement, but also the Recovery Time Objective (RTO) and the Recovery Point Objective (RPO) that you accept. The RTO defines the time you expect it will take to get your system up and running after a failure. The RPO defines the amount of data/transaction you expect to have lost.

Following are some examples of Business Continuity solutions:

- ▶ Backup/Restore solutions, where the implementation ensures that the application is not affected when it needs to be active and available to users. It can be online backup solution or backup can be performed when the application need not be available.
- ▶ Backup/Restore solutions use tape or point in time (PiT) copy functions on disks, such as FlashCopy (FLC), and still keeps the application available. It can be used on a system to minimize the backup window where the application is not available.
- ▶ A manual failover solution between servers, if one server gets errors, like in a cluster solution, but controlled and triggered manually.

- ▶ Cluster implementations, either on a one site configuration where only server errors are handled or in a two-site solution where outages on server, data system, or the total site are handled.
- ▶ Data mirroring, where total site outages are handled, but does not include a cluster implementation.
- ▶ Total site mirroring solution, where all data is mirrored and applications and servers are protected by cluster solutions.

Even in a total site mirroring solution, not everything may need to be mirrored. An example can be that all production data and applications are mirrored and clustered, but test environments are not, and data that can be retrieved from tape, without causing a big impact on the business is not mirrored either.

It is possible and normal to have many different Tiers of Business Continuity solutions in the same company, where the dependency of the solution is an expression of the value to the business. The Business Continuity Tiers are not necessary based on the total IT environment, but can be implemented on an application and system level.

5.1.2 Considerations for Point in Time copy

There are different ways to implement automation in a Business Continuity environment. In a backup solution the backup can utilize point-in-time (PiT) copies, but it can still be implemented in many ways. It might be acceptable that the application is shut down to ensure data consistency. The PiT copy is performed, and then the application is restarted. This can be done by using scripts and using a backup client to access the data on the PiT target volume and backing it up to tape.

If a shut down of the application has a major impact on the business, another solution may be needed. This can be a backup solution that supports the application so that it keeps running while the backup is done. This can be an online backup. From a restore perspective, a total offline backup is often needed, so the solution also needs to cover that aspect. Many applications can be taken offline on weekends. This gives the possibility to perform a total offline backup, but if the application needs to be available during the weekends a third solution is needed.

Today there are many backup solutions that support and integrate into applications. So even when doing a PiT copy, the application can continue to operate. An example of this is the *IBM Tivoli Storage Manager for Mail - Data Protection for Microsoft Exchange* software, when implemented with *Microsoft Volume Shadows copy Service (VSS)* ensures the data consistency of a PiT backup, while the mail application continues to be available.

When using a PiT copy function to do the backup, a very important thing to know is how to restore. We must consider whether the data is usable or if the data is corrupted and the application fails to run. The backup solution might provide one or more of the following possibilities for restoring data from a PiT copy backup:

- ▶ Perform a restore at file level using the backup server to restore the data from the PiT copy volume.
- ▶ Perform a fast restore at a file level, whether the data is copied locally on the server or between the active volume and the PiT copy volume.
- ▶ Perform an instant restore, where files are restored by complete LUN-level copies from the local PiT copy volume.

Some PiT implementations give you the possibility to use space efficient PiT copy functions, where you do not need to have the same capacity on your PiT copy target volume, as you have on the active original volume. The DS4000, as an example, performs this function. This gives a better utilization of the capacity you have available in your storage environment. With this kind of solution you cannot use the instant restore technique, because you do not have all your data on the PiT copy volume. You only have the data that changed; therefore, you need to use the normal restore function.

5.1.3 Level of automation in Business Continuity environment

Automation can be implemented at different levels. Not everything has to be fully automatic, but the higher the Business Continuity Tiers the more automation is normally required.

Automation can be configured to automatically change the active server where the application is running, without moving the location of the data. This kind of solution is often implemented by using high-availability solutions such as Microsoft Cluster Service (MSCS) or IBM High Availability Cluster Multi-Processing (HACMP™) for AIX.

Some solutions support a server change, which is the location of the data from one disk array to another, so that the application keeps running on the same server, but the data in use is changed from the primary to the secondary storage. One solution to this is Geographically Dispersed Parallel Sysplex/Peer-to-Peer Remote Copy (GDPS/PPRC) with HyperSwap manager for System z™.

It is also possible to implement a solution where both application and data moves to another location, and performs a site failover.

The criteria for a change depends on what kind of error occurred, so it may not be necessary to always perform a total failover. Is the error that triggered a

change in the environment, an error on a server, communication error in the LAN, in the SAN, or in the storage subsystem?

Automation can be controlled by software products, such as the IBM TotalStorage Productivity Center (TPC), or it can be part of an integrated solution like Continuous Access for AIX (CAA) and Continuous Access for Windows (CAW).

TPC can be used to automate different functions of your SVC Business Continuity environment.

TPC for Replication can control the Flashcopy, Metro Mirror, and Global Mirror implementation.

5.2 TotalStorage Productivity Center

In this book we show how to install and configure TPC in general. We focus on how TPC can be used in a SAN Volume Controller (SVC) environment.

For installation and configuration of TPC and other relevant information see *IBM TotalStorage Productivity Center V3.1: The Next Generation*, SG24-7194.

TPC is a powerful management tool, and can be used to configure, monitor, analyze, and automate in a SVC environment.

TPC consists of four modules:

- ▶ TPC for Data
- ▶ TPC for Fabric
- ▶ TPC for Disk
- ▶ TPC for Replication (TPC-R).

5.2.1 TotalStorage Productivity Center for Data

TPC for Data performs the following functions:

- ▶ Discover and monitor disks, partitions, shared directories, and servers.
- ▶ Monitor and report on capacity and utilization across platforms to help you to identify trends and prevent problems.
- ▶ Monitor storage assets associated with enterprise-wide databases and issue notification on potential problems.
- ▶ Provides a wide variety of standardized reports about file systems, databases, and storage infrastructure to track usage and availability.

- ▶ Provide file analysis across platforms to help you to identify and reclaim space used by non-essential files.
- ▶ Provide policy based management and automated capacity provisioning for file systems when user-defined thresholds are reached.
- ▶ Generate invoices that charge back for storage usage on a departmental, group or user level.

5.2.2 TotalStorage Productivity Center for Disk

TotalStorage Productivity Center for Disk enables device configuration and management of SAN attached devices from a single console. In addition, it also includes performance capabilities to monitor and manage the performance of the disks. TotalStorage Productivity Center for Disk simplifies the complexity of managing multiple SAN attached storage devices.

It allows you to manage SANs and heterogeneous storage from a single console. TPC for Disk performs the following functions:

- ▶ Collect and store performance data and provide alerts.
- ▶ Provide graphical performance reports.
- ▶ Helps optimize storage allocation.
- ▶ Provide volume contention analysis.

5.2.3 TotalStorage Productivity Center for Fabric

TPC for Fabric performs the following functions:

- ▶ Fabric Manager

The manager performs the following functions:

- Discovers SAN components and devices.
- Gathers data from agents on managed hosts, such as descriptions of SANs and host information.
- 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.
- Monitors switch performance by port and by constraint violations

- ▶ Fabric agents on managed hosts

Each agent performs the following functions:

- Gathers information about the SAN by querying switches and devices for attribute and topology information.

- Gathers event information detected by host bus adapters (HBAs).

5.2.4 TotalStorage Productivity Center for Replication

TotalStorage Productivity Center for Replication simplifies copy services management for the SVC. IBM TotalStorage Productivity Center for Replication provides configuration and management of the FlashCopy and Metro Mirror capabilities of the SVC.

5.3 Automatic Configuration, single point-of-management

Not all kinds of automation are part of a Business Continuity solution. To be a Business Continuity solution the automation must improve the availability of the applications and data, and not just simplify management or improve other functions. TPC has some functions that enable such automatic features, but when the function interacts with other functions they can improve application availability, and thereby become part of a Business Continuity solution.

Often system administrators have requested a single point of management tool, to simplify management and improve control for applications, servers, and storage. With TPC, a storage administrator can add a host to the SVC cluster, and TPC creates SAN zoning and assigns storage capacity from the SVC to the host at the same time. It is not necessary to first define the host in a SAN zone using the switch management tool, then go to the SVC management console and define the host and assign storage capacity to the host, and lastly connect to the host and configure this to use the assigned storage capacity.

The TPC tool can thereby simplify a storage administrators daily tasks. The storage administrator need not handle many different storage management tools, even if there are different storage products in the environment because TPC uses the same interface independent of the storage subsystem.

TPC can also monitor the storage infrastructure. Therefore the errors can be captured fast, investigated, and the real reason for the error can be found and fixed. Often errors are noticed when something is not functioning anymore, for example, the stopping of an application. With TPC you may be able to solve the error situation before the application stops and the end user or other applications notice the error.

In a complex SAN environment TPC might be able to help you locate any performance issues that are in your SAN environment. TPC can show performance from a host and all the way down into the storage subsystem. This

makes it easier to locate where the performance issue is or if it is in the SAN, or at the server level. To get more information about how to analyze performance in a SVC environment using TPC see Chapter 4, “Performance considerations in SVC Business Continuity solutions” on page 119.

5.3.1 Automatic expansion of volumes and file systems

One of the situations that often occurs is that applications and systems stop running because they do not have enough space for data that they need. In other words, the systems ran out of free space. To check all hosts and systems daily is often a big task, and even if it is done, there is no guarantee that a system will not run out of capacity.

To solve this kind of problem, surveillance agents on the host are often used, and the agent can either send a message to the administrator or to an application that can take action so the system gets more capacity before it is impacted. When sending this kind of message to the system administrator, the administrator can verify the information and look into the reason as to why the system ran out of capacity. When the administrator does this, he can decide what action to take. This procedure takes some time, and if this happens in a time frame where the system administrator cannot get access to the system to verify and analyze the situation, the system may have used all the capacity that is available and may shutdown or fail.

In this case some form for automation can help, so a system automatically gets assigned capacity if it runs out of capacity. When implementing this kind of solution, you need to think about what situation could trigger this kind of need. Does the application always have extra capacity or is there a limit? The reason that the application needs more capacity can just be that more data was generated than expected. Therefore the extra capacity is normal behavior, but it can also be caused by an error in the application, so it generates a lot of unusable data, and therefore must be shutdown instead of getting the extra capacity. It is more difficult to remove assigned capacity than to add it.

What can be automated and done using TPC for Data together with TPC for Disk in an SVC environment? At the time of writing we can implement automatic expansion of a volume and the related file system by using scripts that can be activated by TPC for Data on the host. This script can then send commands to the SVC to expand the volume, and execute a command to the operating system to expand the file system when the volume has expanded.

5.4 TotalStorage Productivity Center for Replication

TotalStorage Productivity Center for Replication (TPC-R) is a tool that improves your copy service management including the recovery procedure. The initial set up and ongoing management of large copy services environments is complex and error prone. It is difficult to monitor progress and status of copy services tasks in large environments and installations. Copy services are implemented differently on different platforms. The TPC-R work platform is independent, so it can provide you with data consistency across platforms like zOS, UNIX, and Windows, despite the different management solutions for these operating systems.

Today many IBM Copy Service Management Tools are available.

Following are the control interfaces:

- ▶ TSO
- ▶ ICKDSF
- ▶ REXX™ Execs
- ▶ ANTRQST API
- ▶ DSCLI / Scripts
- ▶ ESS / DS Storage Manager Web GUI

Following are the automation solutions:

- ▶ GDPS
- ▶ eRCMF
- ▶ TotalStorage Productivity Center for Replication
- ▶ Continuous Availability for Windows, AIX, VERITAS Cluster.

With TPC-R you can define and manage your SVC Flashcopy and SVC Metro Mirror solutions using GUI or CLI.

The TPC-R exists in two versions:

- ▶ TPC for Replication
- ▶ TPC for Replication Two Site Business Continuity

The SVC supported only “TPC for Replication” when this book was written. The testing is not finished for “TPC for Replication Two Site Business Continuity”.

5.4.1 TPC for Replication

TPC for Replication provides wizard based and command line session and copy set definition. TPC-R provides administration and operation management for Advanced Copy Services for the SVC. This includes FlashCopy and Metro Mirror single direction.

TPC-R delivers the following functions and support:

- ▶ Multiple Storage Subsystems
 - ESS
 - DS6000
 - DS8000
 - SVC
- ▶ Multiple logical volume types
 - z/OS (CKD) volumes
 - Open system (FB) LUNs
- ▶ Multiple replication types
 - FlashCopy
 - Metro Mirror
 - Global Mirror
- ▶ Simplified replication management and monitoring
- ▶ GUI and CLI

TPC for Disk, Data, and Fabric are *not* required.

5.4.2 TPC for Replication Two Site Business Continuity

TPC for Replication Two Site Business Continuity provides the same functions as TPC for Replication enhanced with TPC for Replication Cluster implementation and failover/failback for Metro Mirror. The TPC-R Cluster implementation is a solution where you have two TPC-R servers in your environment. Therefore you do not have a single point-of-failure at the TPC-R level. TPC-R Cluster uses an active and a standby TPC-R server, but every configuration change you perform on the active TPC-R server is mirrored to the standby TPC-R server.

In this publication we concentrate only on the functions of TPC-R support for SVC.

Communication between TPC-R and SVC

The TPC-R server communicates with the SVC Master Console (MC), although the CIMOM agent runs on the MC. The TPC-R does not communicate directly with the SVC Cluster or any of the SVC Nodes in the SVC Cluster (See Figure 5-2 on page 167).

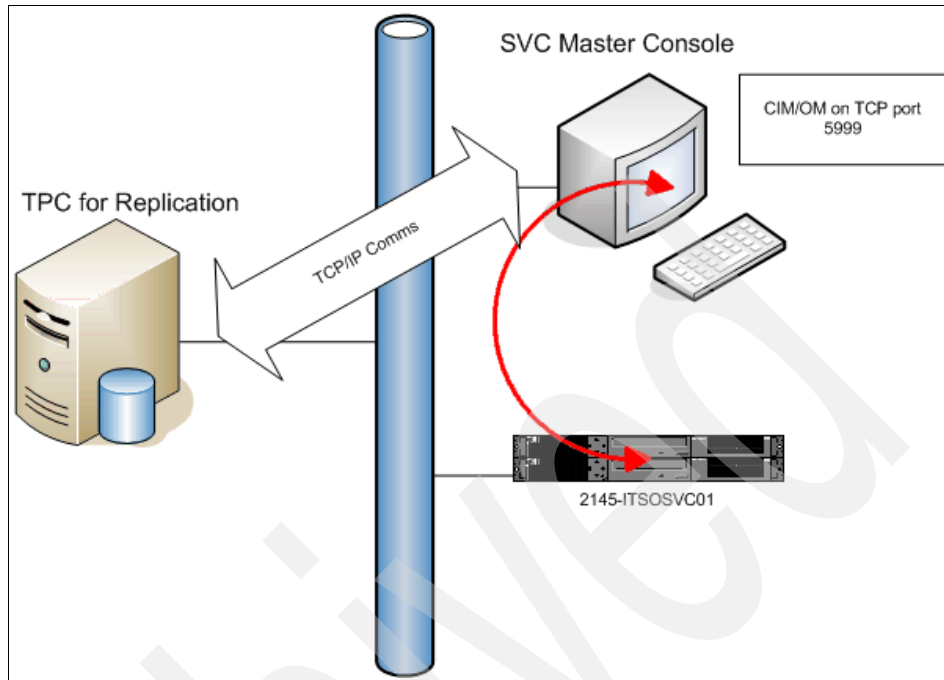


Figure 5-2 TPC-R to SVC MC communication

The TPC-R connects to the MC CIMOM using TPC/IP port 5999 as default, and the SVC MC uses port 6120-6140 to communicate with the TPC-R server. The first SVC MC gets to use port 6120, the second gets port 6121, and so on.

Session

A *session* is a container of multiple copy sets making a Consistency Group.

Primary Session State

TPC for Replication use the following states for a session:

- ▶ Preparing
- ▶ Prepared
- ▶ Suspend
- ▶ Target Available
- ▶ Defined

Copy set

A *copy set* is a set of volumes that contain copies of the same data. All the volumes in a copy set are of the same format (count key data [CKD] or fixed

block) and size. In a replication session, the number of volumes in a copy set and the role that each volume in the copy set plays are determined by the copy type.

Role

A volume's *role* is the function it assumes in the copy set and is composed of the intended use for Global Mirror and Metro Mirror, the volume's site location. Every volume in a copy set is assigned a role. A role can assume the functions of a host volume, journal volume, or target volume.

The role uses a two-digit identifier that indicates volume function and site.

Volume function

- ▶ H - Host (receives application I/O)
- ▶ T - Target (FlashCopy target)
- ▶ J - Journal (Global Mirror FlashCopy target, not applicable for SVC)

Site where volume is located

- ▶ 1 - Primary site
- ▶ 2 - Secondary site

You can see the role of a volume by selecting **Session Details** → **Role Pair**.

You see a window with information as shown in Figure 5-3 on page 169. Here you can see that the volumes SIAM_VD1 and SIAM_VD2 have the role H1, and that volumes SIAM_MM_VD1 and SIAM_MM_VD2 have the roles H2. Therefore the volumes SIAM_VD1 and SIAM_VD2 are the *Host volumes* (the one that receives application I/O) on Primary site, while the volumes SIAM_MM_VD1 and SIAM_MM_VD2 are *Host volumes on Secondary site*. Often these are referred to as Primary and Secondary volumes on disk storage subsystems. On SVC these are called *Master VDisk* and *Auxiliary VDisk*, when you define a Metro Mirror relationship.

MMSIAM H1-H2	
Error Count: 0	
Recoverable: 2	
Copying: 2	
Progress: 100 %	
Timestamp: n/a	
◇H1	◇H2
SIAM_VD1	SIAM_MM_VD1
SIAM_VD2	SIAM_MM_VD2

Figure 5-3 Volume Role

Role pair

A *Role Pair* is a pair of volumes/roles within a Copy Set. In the earlier version of TPC-R, TPC V2R3 Replication Manager this is called *sequence*.

In the following sessions, you see some examples of Session Role Pairs and what they mean. The H1-H2, H2-J2, H1-J2 is all for Global Mirror or Metro Mirror session types, where as the H1-T1 are FlashCopy Session types.

H1 - H2

- ▶ Storage subsystem hardware pair.
- ▶ For a Global Mirror session, H1-H2 are the Global Copy pair underlying Global Mirror.
- ▶ Alternatively can be a Metro Mirror relationship

H2 - J2

- ▶ Storage subsystem hardware pair
- ▶ Point in Time Copy underlying Global Mirror

H1 - J2

- ▶ TPC-R logical pair
- ▶ Global Mirror A and C volume

H1 - T1

- ▶ Hardware pair
- ▶ Point in Time Copy (FlashCopy) relationship

At the TPC-R GUI you can see the role pair when you select **Session**. Here you see the role pair listed in two places. You see H1-H2 under Role Pair (left), for a Metro Mirror relationship. On the right you see the Role Pair - H1 and H2. See Figure 5-4.

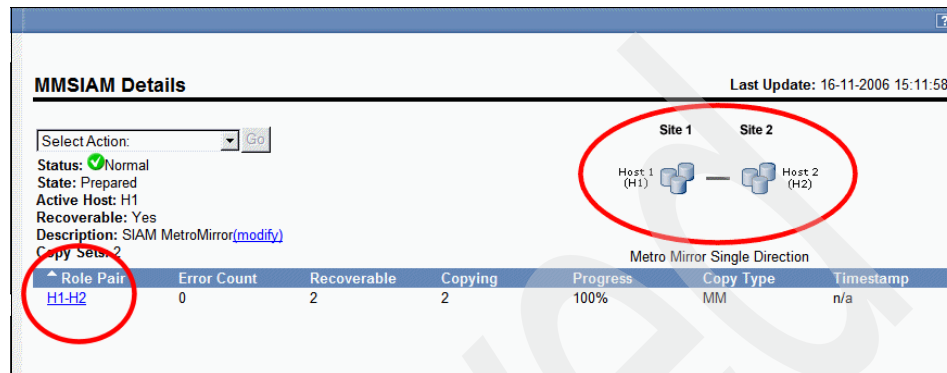


Figure 5-4 Session Role Pair under Session details

TPC-R versus SVC-Metro Mirror naming

When you define a Metro Mirror session using the TPC-R, TPC-R initiates commands on the SVC through MC. When TPC-R defines a Metro Mirror Consistency Group on the SVC, TPC-R use standard names. TPC-R uses the name *rccstgrp* for a Metro Mirror Consistency Group. *rccstgrp* stands for RemoteCopyConSisTencyGRouP. On SVC you see *ex.rccstgrp1*, which is followed by a number. You can see the name for the Metro Mirror Session on the SVC using the GUI or by executing the following command using the SVC CLI tool

```
svcinfolsrconsistgrp
```

When you define a Metro Mirror session, the TPC-R also initiates commands for the Copy Set you create. For the Metro Mirror Copy Sets, the TPC-R uses the name convention *rcrel* for a Metro Mirror Relationship. *rcrel* stands for RemoteCopyRELationship. On the SVC you see this followed by a number, *ex. rcrel1*. You can also see the name for the Metro Mirror Copy Sets on the SVC using the GUI or by executing the following command using the SVC CLI:

```
svcinfolsrcrelationship
```

TPC-R versus SVC - FlashCopy naming

When you define a FlashCopy session using the TPC-R, TPC-R initiates commands on the SVC through the MC. When TPC-R defines a FlashCopy Consistency Group on the SVC, TPC-R uses standard names. TPC-R uses the naming convention *fccstgrp* for a FlashCopy Consistency Group. *fccstgrp*

stands for FlashCopyConsistencyGroup. On the SVC you see this followed by a number, *ex. fccstgrp1*. You can see the name for the FlashCopy Session on the SVC using the GUI or by executing the following command using the SVC CLI:

```
svcinfolsfccconsistgrp
```

When you define a FlashCopy session, the TPC-R also initiates commands for the Copy Set you create. For the FlashCopy Copy Sets, the TPC-R uses the name convention *fcmap* for a FlashCopy Mapping. *fcmap* stands for FlashCopyMapping. On the SVC you see this followed by a number, *ex. fcmap1*. You can also see the name for the FlashCopy Session, after you have created it, on the SVC using the GUI or by executing the following command using the SVC CLI:

```
svcinfolsfcmmap
```

5.4.3 TPC for Replication installation

For information about how to install TPC for Replication see *TotalStorage Productivity Center for Replication: Installation and Configuration Guide*, SC32-0102.

In this section we go through the TPC-R installation on a Windows 2003 Server. The installation can also be done on AIX V5.3 ML3 or Linux® (RedHat EL4 AS1 and SUSE LES9 SP2).

Before installing TPC-R you must install the IBM DB2 UDB Express V8.2 database. This database is part of the TPC-R package. If you have not installed the database the TPC-R installation fails with the following error message as shown in Figure 5-5.

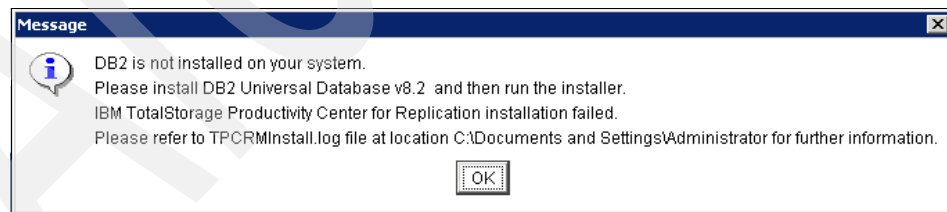


Figure 5-5 TPC-R DB install error message

When you install the TPC-R you need a DB2 user ID and password. The ports number allows the TPC-R to connect to the database the name for the TPC-R database to be created. You need to define a TPC-R user ID, password, and a group this user is to be part of.

Reference the *TotalStorage Productivity Center for Replication: Installation and Configuration Guide*, SC32-0102, for details on how to install the DB2 database.

The DB2 user ID and password needs to be the same as you used when installing the IBM DB2 UDB Express V8.2 database.

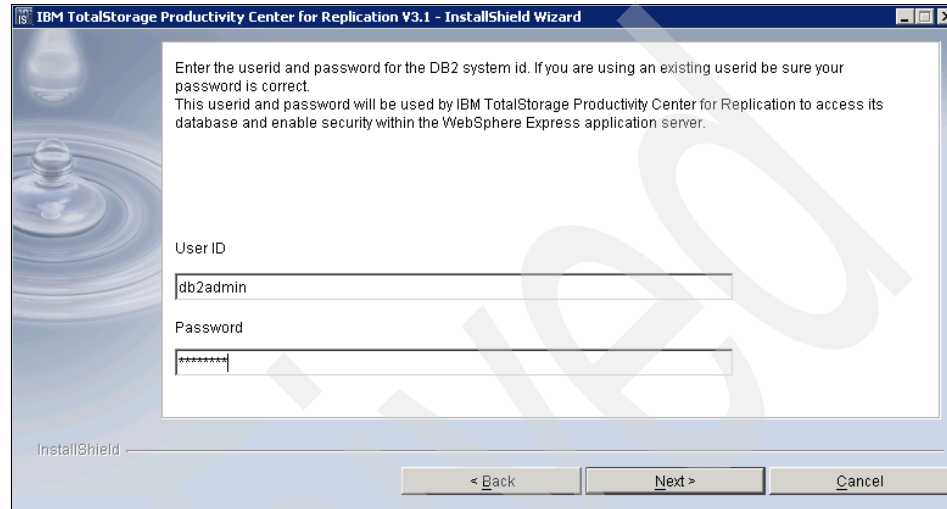


Figure 5-6 DB2 user id and password for TPC-R installation

The port numbers that TPC-R uses can be changed—if you are already using the port numbers for other purposes, or if you do not want to use standard TCP/IP ports—because this might make it easier for hackers to get access to your systems. The default ports the TPC-R uses are 9080 and 9443. But you can choose whatever you want, as long as the port numbers are not in use and fulfil the TCP/IP requirements. Port 9080 is used for normal HTTP and port 9443 is used for secure HTTPs. See Figure 5-7 on page 173 to see where to define the uses of ports TPC-R.

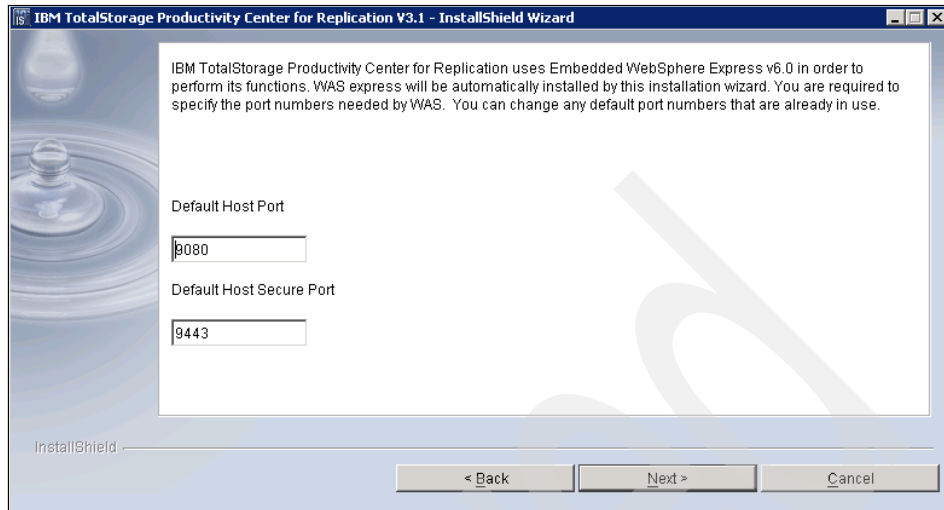


Figure 5-7 Define TPC-R host ports used

You can define the name of the TPC-R database as you like. The TPC-R installation program checks if the database name is already in use by the IBM DB2 UDB Express V8.2 database. In our example we define the database name as TPCRM. See Figure 5-8.

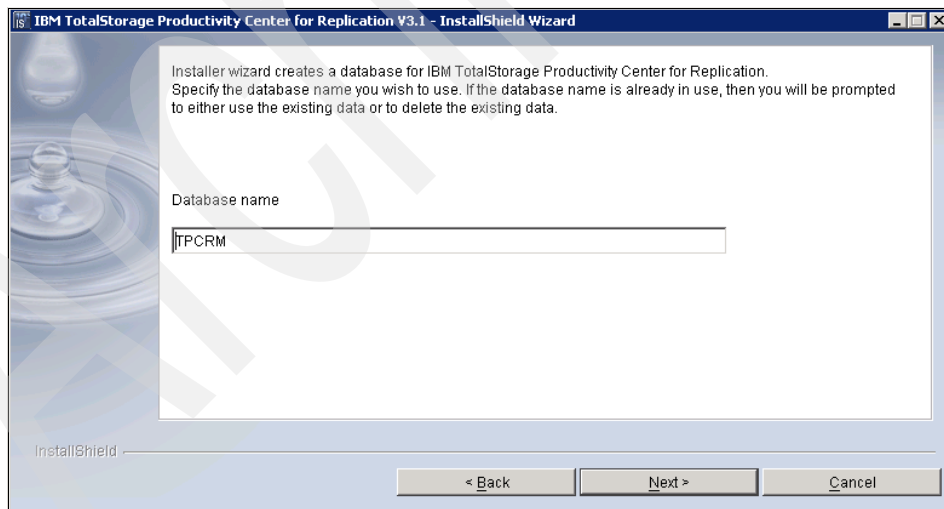


Figure 5-8 Define TPC-R database name

The TPC-R user ID and password, called *Copy Services Manager (CSM)* user, is defined under the TPC-R installation. This is the user ID and password you use

to connect to TPC-R when the installation is finished. We expect the CSM user ID to be changed to TPC-R Administrator ID in the next version of the TPC-R installation tool.

You can use a user ID that is already defined on the server or in the Active Directory® if it is installed. If you use an existent user ID, be careful to get both user ID and password right. If the user is not defined on the server, the TPC-R installation creates the user on the server.

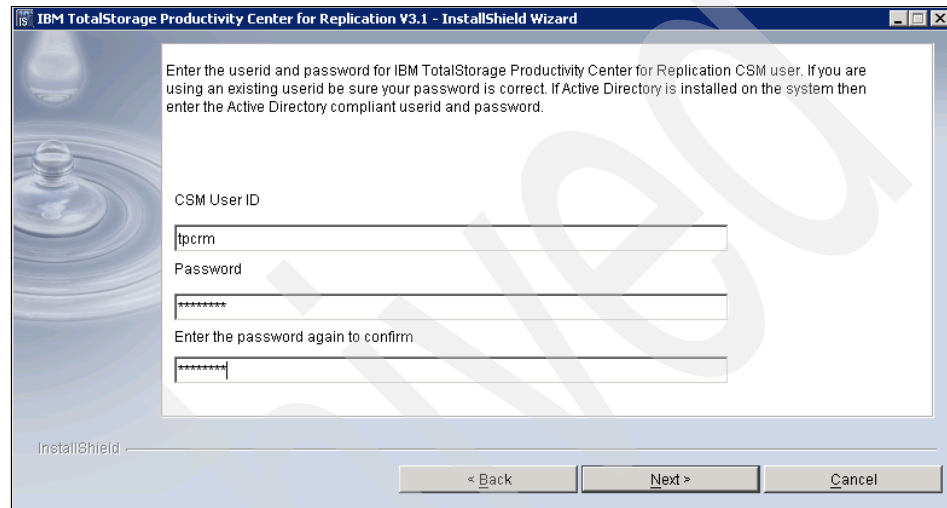


Figure 5-9 Define the user ID and password for TPC-R

The last thing you need to define in the TPC-R installation is a group name. It must contain all the user IDs that are authorized to access the TPC-R server. If you want more defined users under the installation to access the TPC-R, you just need to make sure they are members of the TPC-R group.

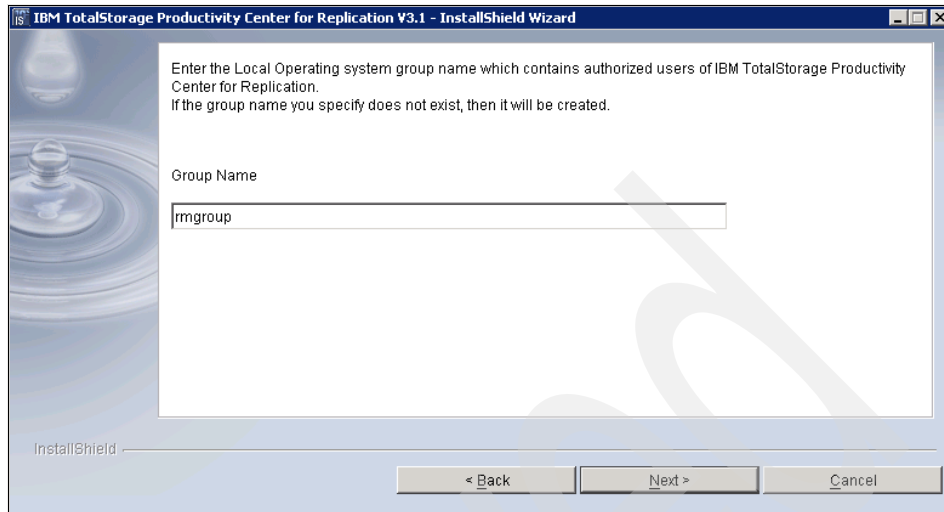


Figure 5-10 Define TPC-R group name

After you give the TPC-R installation program the information it needs, you are prompted to review the location of the TPC-R. It shows you how much space it needs. If you accept the parameters the installation begins.

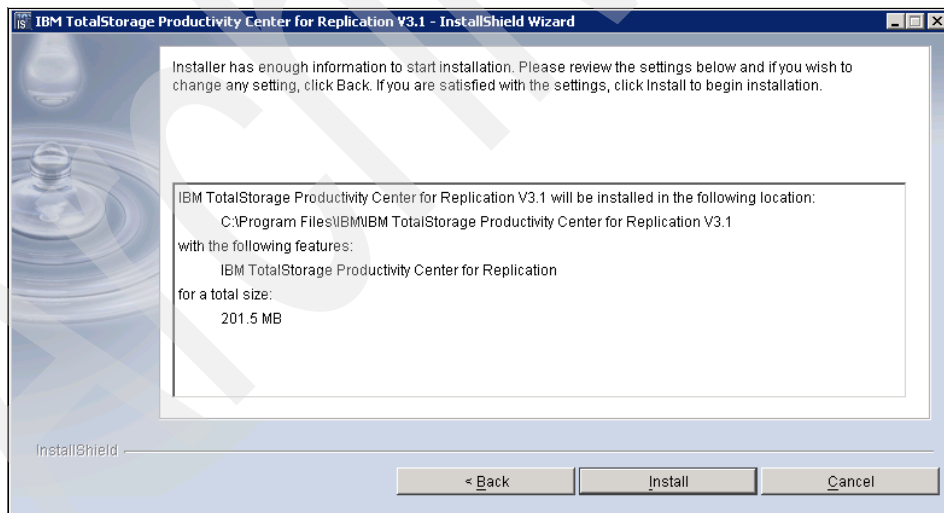


Figure 5-11 Verify the installation information and start the installation

When the installation starts, it prompts you for information that it needs to change some database parameters. Select **Yes** to increase the memory heap size allocated to the DB2 database, so that TPC-R can handle large numbers of copy

sets (100 or more). If you do not set the DB2 parameters, a resulting crash of the DB2 server can occur (when there are 100 or more copy sets in a session), and the easiest way to recover from that is by restarting the TPC-R server.

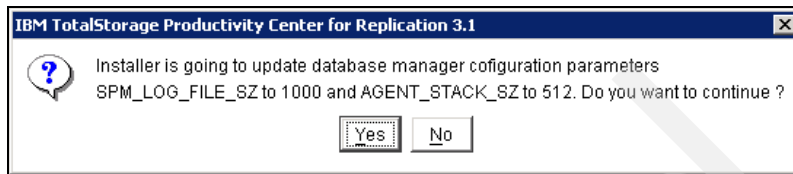


Figure 5-12 Change of database parameters

When you have done all this, you get a confirmation that the change on the DB2 database is successful. See Figure 5-13 and Figure 5-14.

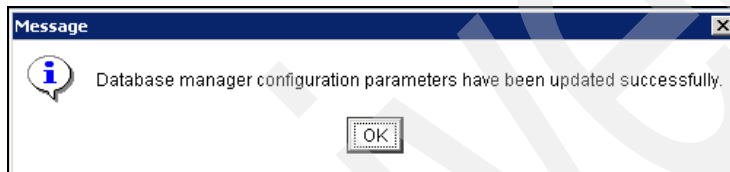


Figure 5-13 Successfully DB change confirmation

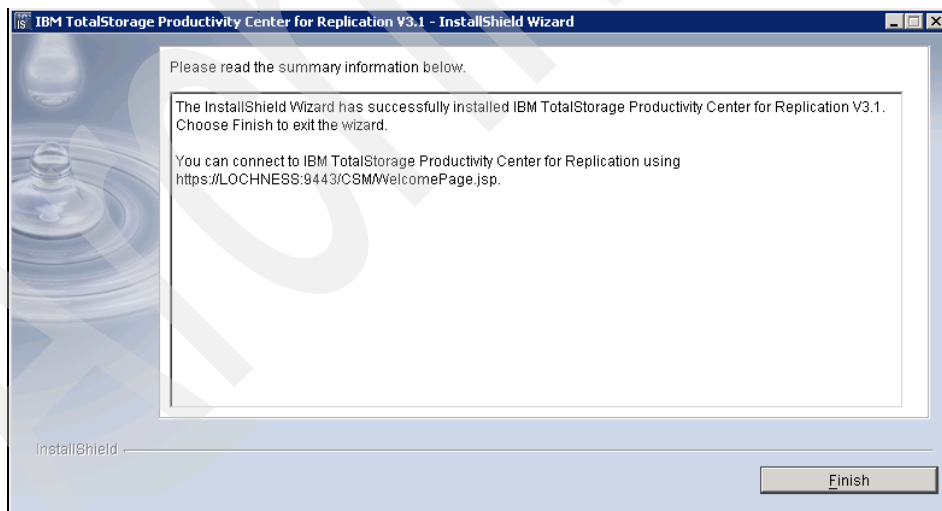


Figure 5-14 Successfully installation confirmation

5.4.4 Overview of TPC-R

After you install TPC-R, configure it by adding storage systems, defining sessions, adding copy sets to the sessions, and starting the sessions.

When you connect to the TPC-R server, use the URL including the port number to be used, and the CSM directory. In our implementation this is the following:
<https://9.43.86.84:9443/CSM/>.

The 9443 is the secure port number defined under the TPC-R installation (See Figure 5-7 on page 173). You can also connect to the TPC-R server using the unsecure port. Define this port number in your URL, and use HTTP instead of HTTPS as the connection type. The URL in our example is as follows
<http://9.43.86.84:9080/CSM/>.

After pointing your browser to the right URL you see a login window as shown in Figure 5-15 on page 178. To login you need to use the user ID and password you defined under the installation (See Figure 5-9 on page 174).

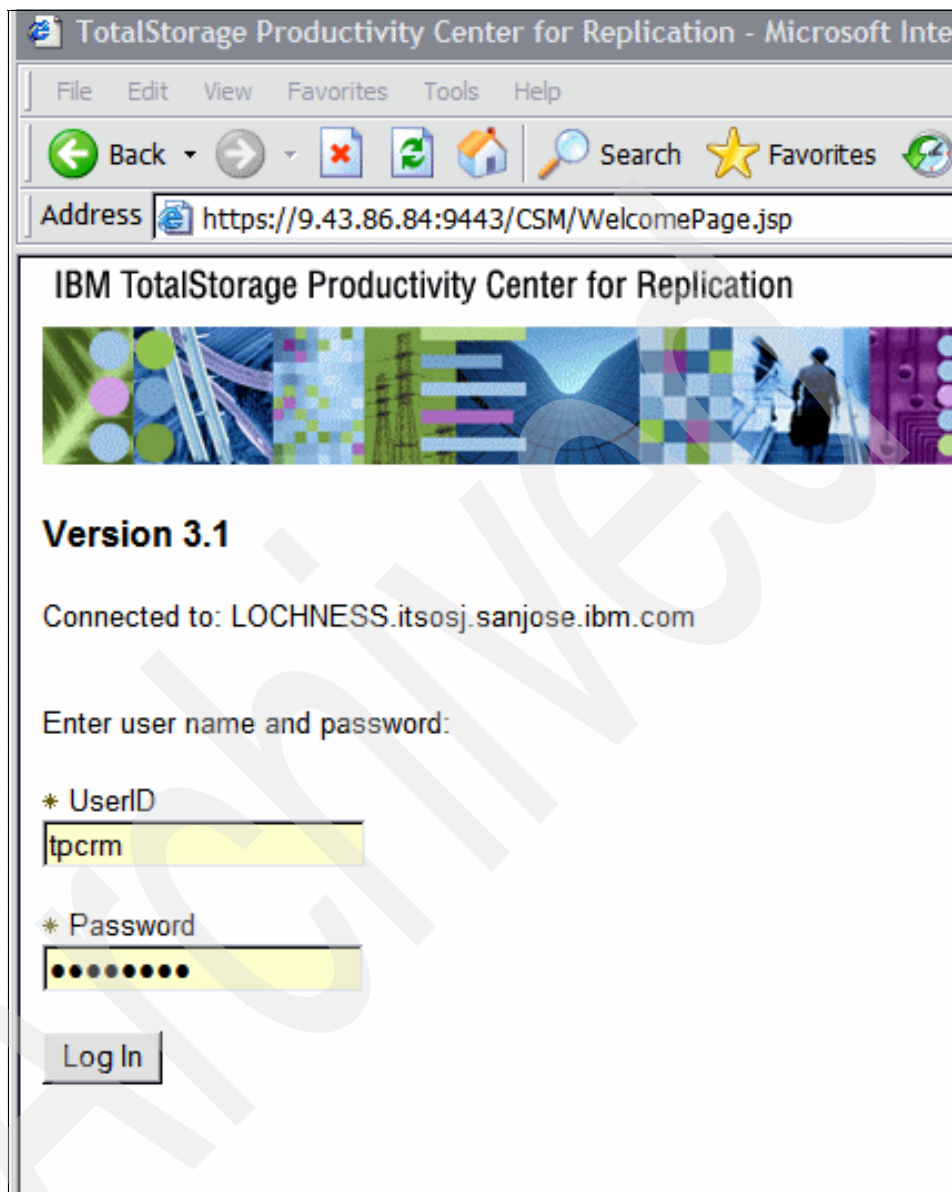


Figure 5-15 Login windows from TPC-R server

The first window you see after logging in looks similar to Figure 5-16 on page 180. The window is divided into three sections.

- **My Work** is a menu with the tasks you select, depending on the task you perform.

- ▶ **Health overview**
 - Gives you the same overview window when you logged in.
- ▶ **Session**
 - Session is where you define and check status information about the sessions you define. You also use the session to start/stop/suspend your copy tasks.
- ▶ **Storage subsystem**
 - Storage subsystem is where you manage the storage subsystems you want TPC-R to use.
- ▶ **ESS/DS Paths**
 - It is here that you define the paths used for remote copy solution for the ESS, DS6000, and DS8000. For the SVC you do not need to configure dedicated paths to handle remote copy solutions. Hence this function is not covered in this book.
- ▶ **Management servers**
 - This is for configuring and managing a cluster TPC-R environment, with an active and a passive TPC-R server. This is not supported for the SVC. Therefore this is not covered in this book.
- ▶ **Advance tools**
 - This is used to generate log files to be used in troubleshooting and to define the refresh rate of the TPC-R management tool. Here you enable and disable the heartbeat function in a cluster TPC-R installation.
- ▶ **Console**
 - This gives you an overview of all the commands and the status on the commands of the logged in user.

At the bottom left side, you see the **Health Overview**. This overview picture is always there, no matter what task you are working on. So you will always be able to see the status on the **Sessions, Storage Subsystems, and Management Servers**.

The third section on the right side is the **Active Area**. This depends on the task you have selected under **My Work**. When you login, this shows you the **Health Overview** that gives you the same information as in the second section at the bottom left side.

Because this is the first time we are logging into the TPC-R and we have not defined anything, the **Health Overview** shows you gray status icons for **Sessions, Storage Subsystems and Management Servers**. See Figure 5-16 on page 180.

We define Storage Subsystems and Sessions later. The Management Server part is for the TPC-R Two Site Business Continuity installation and shows you the status on the active and the standby TPC-R server. The TPC-R Two Site Business Continuity version is not supported by the SVC at this time of writing; therefore, we are not defining a secondary TPC-R server.

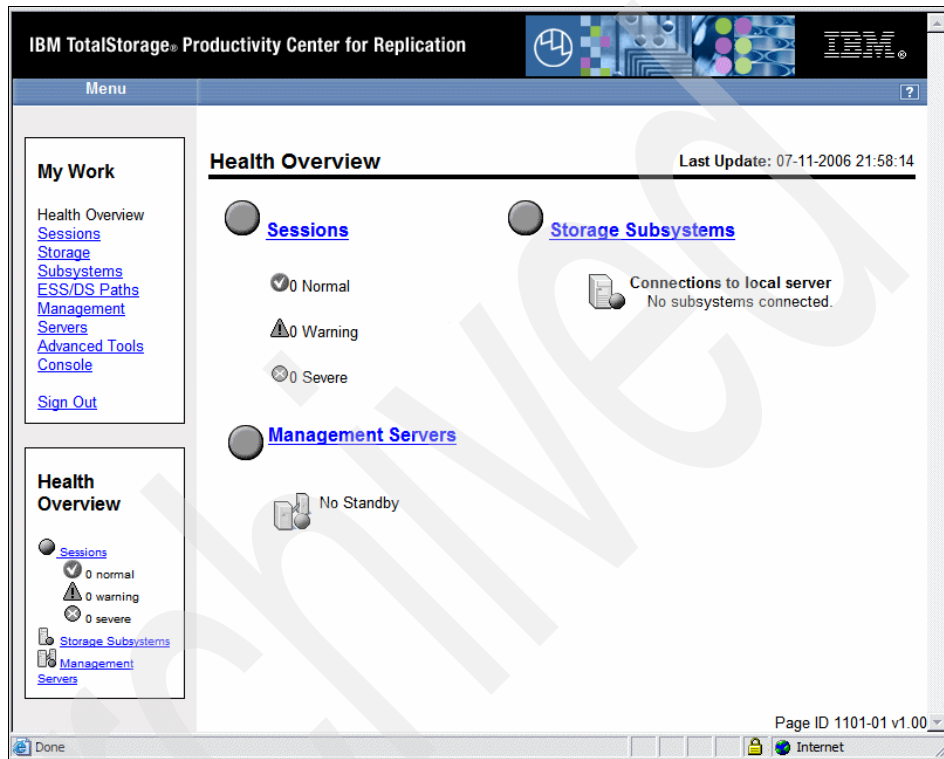











Figure 5-16 Startup windows for TPC-R

The TPC-R uses the following nine different icons to show the actual status:

-  Icon for normal state
-  Icon for warning state
-  Icon for error state
-  Icon for all Storage Subsystem can communicate with TPC-R
-  Icon for no Storage Subsystem is defined
-  Icon for Storage Subsystem error

-  Icon for Standby Server is active and synchronized
-  Icon for there is not defined any Standby Server
-  Icon for Standby Server is synchronizing

5.4.5 Adding the SVC to TPC for Replication

When you add the SVC to the TPC-R you will need the following information:

- ▶ CIMON Server IP / Domain name
- ▶ CIMON communication port
- ▶ CIMON username
- ▶ CIMON password

The CIMON server is the SVC Master Console, and the standard communication port is port 5999, but you can choose another CIMON port in your environment.

The first time you select **Storage Subsystem** in **My Work** or use the active link at the **Health Overview** page, you see a window where you can add a storage subsystem. It also tells you that there is no storage subsystem defined. After you click the bottom **Add Subsystem**, you are guided to the next window as shown in Figure 5-17 on page 182. Select **SVC** and select **Ok**.

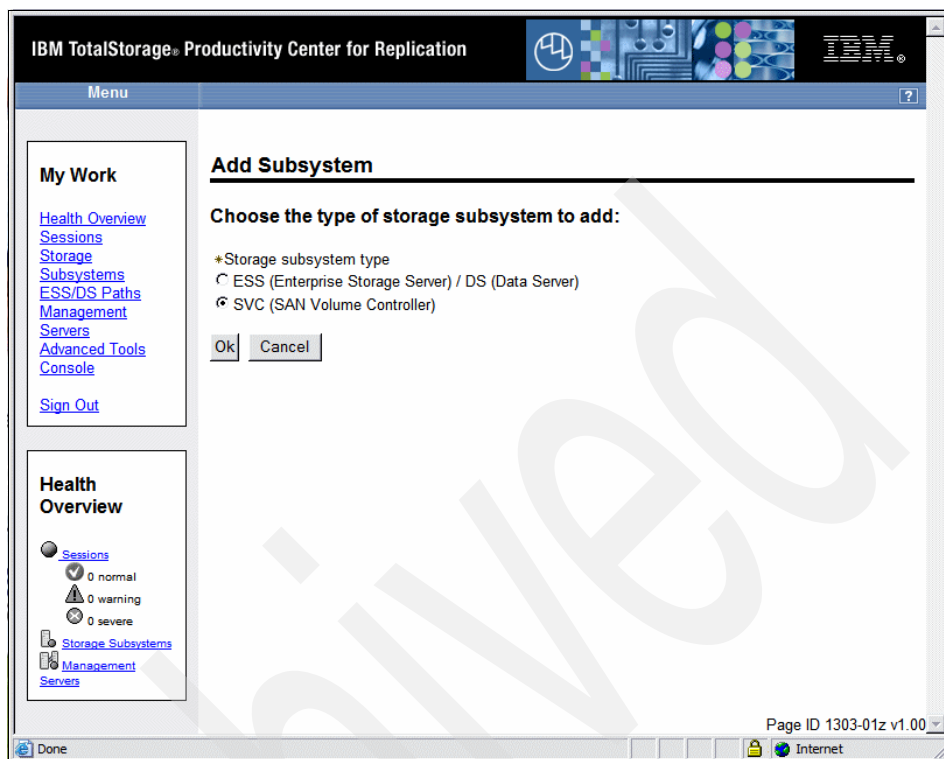


Figure 5-17 Add Subsystem window

You get the following window, where you need to define the SVC to the TPC-R, see Figure 5-18 on page 183.

IBM TotalStorage® Productivity Center for Replication

Menu

My Work

- [Health Overview](#)
- [Sessions](#)
- [Storage](#)
- [Subsystems](#)
- [ESS/DS Paths](#)
- [Management](#)
- [Servers](#)
- [Advanced Tools](#)
- [Console](#)
- [Sign Out](#)

Health Overview

- [Sessions](#)
 - 0 normal
 - 0 warning
 - 0 severe
- [Storage Subsystems](#)
- [Management](#)
- [Servers](#)

Add Subsystem (SVC)

Please enter the information below to add a new subsystem.

Defining a CIMOM server will add all the SVC clusters managed by the CIMOM server.

*CIMOM Server IP/Domain Name
9.43.85.141

*Port
5999

* Username
superuser

* Password
••••••

Ok Apply Cancel

Page ID 1303-01b v1.00

Done Internet

Figure 5-18 Add SVC Subsystem information

The CIMOM Server IP/Domain Name is not the SVC cluster IP address, but the Master Console IP address or domain name. The Username is the name you use to access the SVC Cluster Console on the Master Console. It is not the TPC-R username and password or the username you use to login to the Master Console.

After you define the Master Console as the Storage Subsystem, the TPC-R tries to connect to the CIMOM service on the SVC Master Console. If the connection is successful you will get a message that the TPC-R is connected to the SVC Master Console. We added two SVC Master Consoles and have shown how to define a Metro Mirror solution using TPC-R in a two SVC Cluster environment. Figure 5-19 on page 184 shows you that the TPC-R configuration is a success.

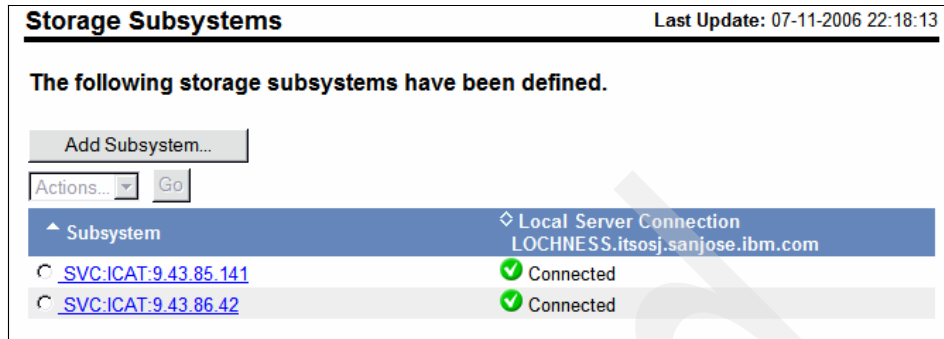


Figure 5-19 Connection to Storage Subsystem is okay

In Figure 5-19 you can see that we defined two SVC Master Consoles as our Storage Subsystems. The Subsystem shows you the type, IP address, and the status in this case as **Connected**.

5.4.6 Define the SVC Metro Mirror session

After we define the Storage Subsystems, we can define some sessions to hold and control the tasks we want TPC-R to handle. First we define a Metro Mirror relationship between the two SVC Clusters.

1. In the TPC-R windows we select the **Session**.
2. To create the session we need to select the type of session we want to create. In Figure 5-20 you can see the different type of sessions TPC-R can manage.

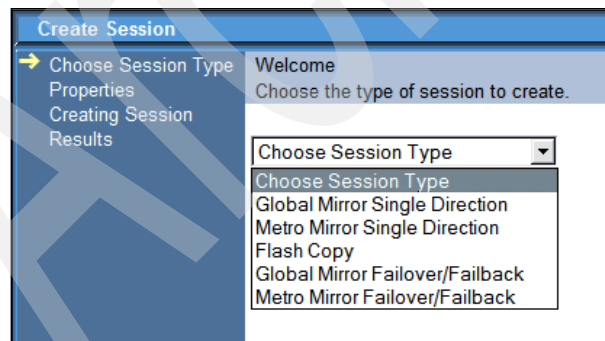


Figure 5-20 Sessions type in TPC-R

At the time of writing not all the session types were supported by the SVC. Metro Mirror Single Direction and FlashCopy are the only functions supported by SVC at the time of writing.

3. To define our Metro Mirror session, we select **Metro Mirror Single Direction** as the Session type.
4. Assign a name and a description for the session. The description is optional but recommended because TPC-R can manage many sessions, even for the same host. In Figure 5-21, we define a session with the name SIAM Metro Mirror and a short description.

https://9.43.86.84:9443/CSM/WizardFactory.jsp?wizardname=CreateSessionWizard

Create Session

✓ Choose Session Type
→ Properties
Creating Session
Results

Properties
Name and describe the session.

*Session name
SIAM MM

Description
SIAM Metro Mirror, from SVC1 to SVC2

Click 'Next >' to create the session.

Figure 5-21 Define Metro Mirror Session

5. After defining the Session, look at the session windows again. Observe the status of the new session. In Figure 5-22 on page 186 you can see the status of the session SIAM Metro Mirror we just created, including information about the number of Copy Sets we defined for the session (in this case zero). The **State of the Session** is defined, since we just created it but have not assigned any tasks for the Session.

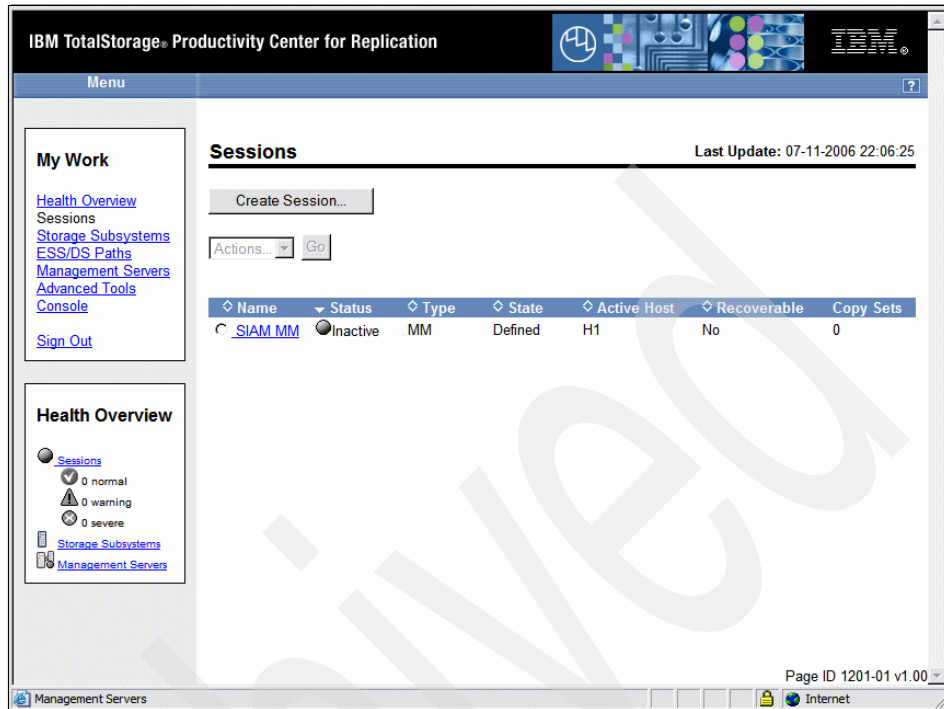


Figure 5-22 Session status

The status and the state of a session depends on what you defined and what action you took for the session. In Figure 5-23 on page 187 you can see the dependencies between the state and the status for a Metro Mirror session. We only created the session SIAM Metro Mirror and have not added any Copy Set to it yet, so the status is **Inactive** and the State Define is as expected.

When we add Copy Set to the Session, it does not change state or status. The state and status will change after we start the session, as we show later.

Sessions in TPC-R are like Consistency Groups on the SVC, and when we define a session and add copy sets to it in TPC-R, this creates a Consistency Group for the session and adds the copy sets as relationships in the Consistency Group.

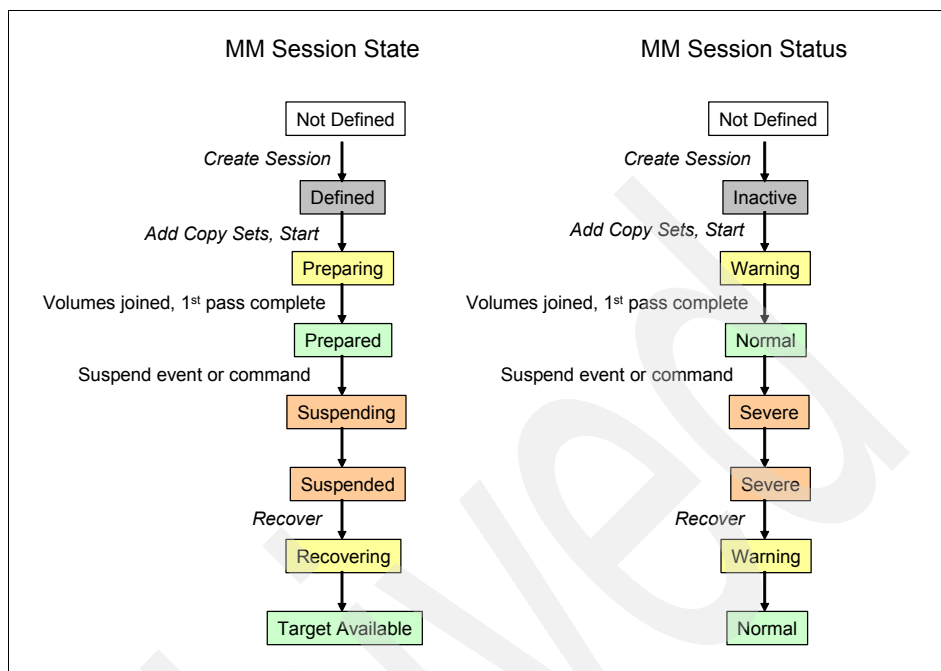


Figure 5-23 Relation between State and Status of a Metro Mirror session

Before we can start the session we created, we need to add the Copy Set to it.

1. Under the same task as where we define the Sessions, we select **Add Copy Sets** from the pull-down menu.
2. In the first window you see after you start to define a copy set, you have to select **Host 1 storage subsystem**. By Host storage subsystem TPC-R means the storage subsystem that receives host I/O. Host 1 means the primary storage subsystem that receives host I/O. So Host 1 and Host 2 can be translated into primary and secondary storage subsystems.

As you can see in Figure 5-24 on page 188, we selected one of the SVCs in our environment as the Host 1 storage subsystem. Notice that now we do not get the SVC Master Console as the storage subsystem. This is what we used when we defined the storage subsystem. Now you get the IP address of the SVC Clusters that the SVC Master Console is managing. After you select the SVC Cluster as the Host 1 storage subsystem, you can select in what I/O group the VDisk you want to make a Copy Set for is located. After this you will get a list of all the VDisks in the I/O group that you can select as the primary volume in the copy set.

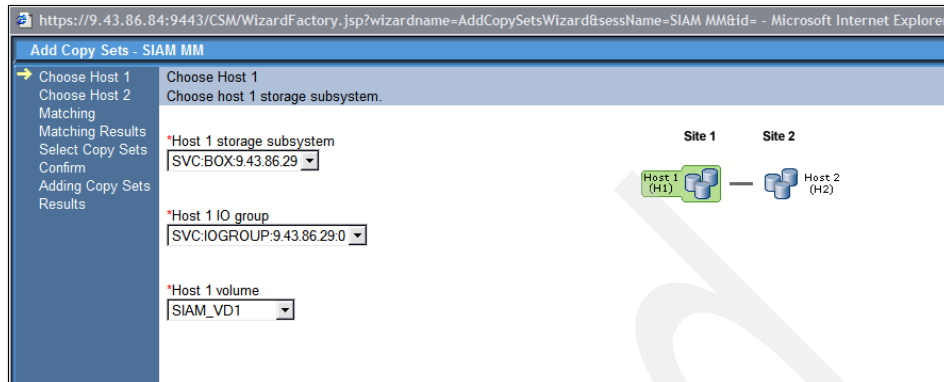


Figure 5-24 Copy Set definition for Host1

3. In this example we chose SVC 1 as the Host 1 storage subsystem and I/O group 0 as the I/O group, where the volume we want to make a metro mirror relationship for is located. Last we select the specific volume, under the Host 1 volume menu, we want to use. Under Host 1 volume, we get all the VDisks that are defined to be managed by the I/O group in the SVC Cluster we chose. You can see the VDisk name in the Host 1 volume pull-down menu.

If you do not define a specific I/O group, the TPC-R goes out and tries to get information about all I/O groups, from 0 to 3. It automatically tries to match volumes from your primary SVC Cluster to your secondary SVC Cluster. You cannot select any volumes under the Host 1 volume menu. The TCP-R tries to match all the VDisks on your primary SVC Cluster with VDisks on your secondary SVC Cluster.

4. After you define the Host 1 storage subsystem, I/O group, and volume, define the Host 2 storage subsystem you want to use. By the Host 2 storage subsystem we may use the name of the secondary storage subsystem.

In this example we choose SVC2, and again this is the IP address of the SVC Cluster and not the SVC Master Console. We select I/O group 0 and the volumes we want to use as the target volumes for the Metro Mirror (see Figure 5-25 on page 189). The TPC-R only lists volumes that can be selected as target volumes, which means volumes that have the same size as the primary volume.

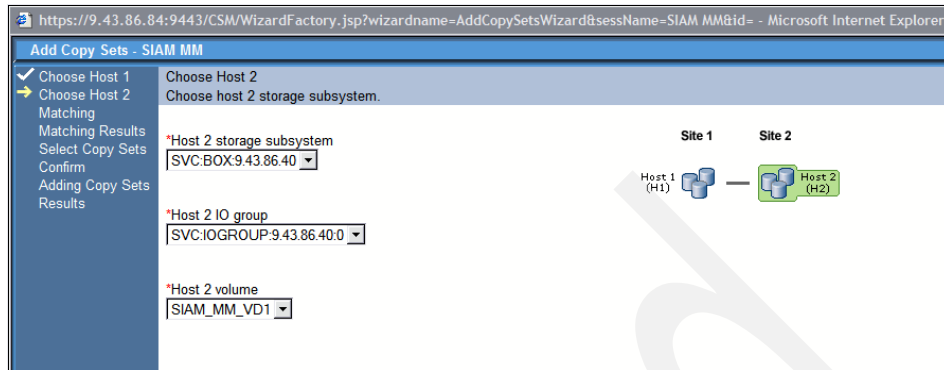


Figure 5-25 Copy Set definition for secondary volume

- After you add the Copy Set to the session, and confirm the action, return to the session window, where you see the Session including information about the Copy Sets it holds. Select **Session** and **View Detail**, to get all the information as shown in Figure 5-26.

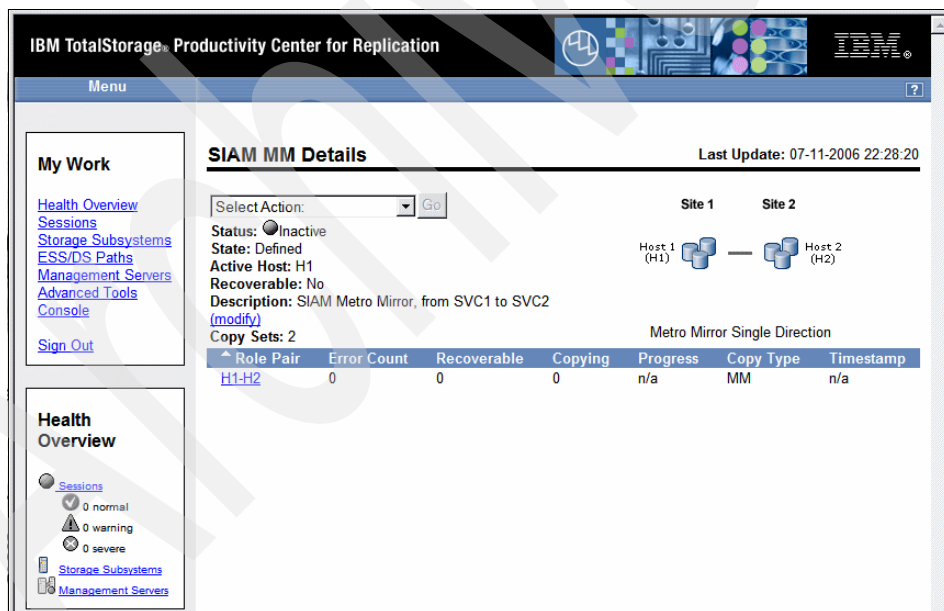


Figure 5-26 Status information about Session just defined

Because you have just defined the Session, the Status is **Inactive** and the State is **Defined**. The Active Host is **H1**, which means the primary storage subsystem is **active**. The status Recoverable is **No** because the session has not started and the data is not mirrored to the secondary volume. You cannot recover from a

volume that does not have any data on it. When the Session is started and the Metro Mirror is in a synchronized state, the status for Recoverable changes to **Yes**.

After you start the Session, all the Copy Sets defined in the session also start. The status changes to **Warning**, and the state changes to **Preparing**.

While the status is preparing, the data is synchronized between the volumes in the Copy Sets. In our example all data on volume SIAM_VD1 is mirrored to SIAM_MM_VD1 and all data on volume SIAM_VD2 is mirrored to volume SIAM_MM_VD2. When all the volumes are in synchronized mode, the status for the Session changes from **Warning** to **Normal**. The State changes from **Preparing** to **Prepared**.

When the Session is in this state, it is possible to use it for recovering.

5.4.7 Define the SVC FlashCopy session

When you define FlashCopy sessions for the SVC in TPC-R, the process is the same as for making a Metro Mirror session. See section 5.4.6, “Define the SVC Metro Mirror session” on page 184.

After you define the session for the SVC FlashCopy, add Copy Sets to the session. The procedure is almost the same as adding Copy Sets for a Metro Mirror session.

Sessions in TPC-R are like Consistency Groups on the SVC, and when we define a session and add copy sets to it in TPC-R, this creates a Consistency Group for the session and adds the copy sets as FlashCopy mappings in the Consistency Group.

1. When you add the FlashCopy Copy Set you first define the “Host 1 Storage Subsystem”. This is the SVC Cluster IP address, on which you want to perform the FlashCopy. After selecting the SVC Cluster, select the SVC I/O Group and the Volume on which you want to perform the FlashCopy. In Figure 5-27 on page 191 we selected volume “SIAM_FC_VD1” as the source volume.

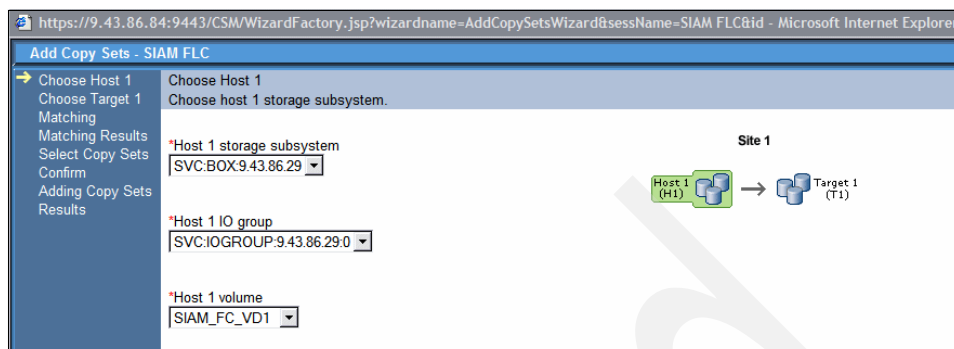


Figure 5-27 Define FlashCopy Source

In the window you can also see that there is only one site shown, Site 1. Host 1 and Target 1 are the names for Source and Target volumes.

The list of volumes that we can select as Target 1 volume contains only those volumes not in a FlashCopy relationship.

After you define the Target 1 volume, the volume list only shows you volumes that have the same size as the volume you selected as your Host 1 volume.

2. After you define both Host 1 volume and Target 1 volume, you get a message that the matching of the FlashCopy relationship is a success. Select **Next**. You will get a window where you have to select the Copy Set you added to the Session. You have to confirm that you added this Copy Set to the Session before you get a window telling you that it has successfully completed.

After you add the Copy Set to the Session, a window appears with information similar to Figure 5-28.

Sessions							Last Update: 14-11-2006 11:53:41
Create Session...							
Actions... Go							
◇ Name	▼ Status	◇ Type	◇ State	◇ Active Host	◇ Recoverable	Copy Sets	
◇ SIAM MM	✓ Normal	MM	Prepared	H1	Yes	2	
◇ SIAM FLC	● Inactive	FC	Defined	H1	No	1	

Figure 5-28 FlashCopy session status

Even though you defined a FlashCopy Session and added a Copy Set to it, you still have to start the session before the actual FlashCopy is performed.

3. When you want to perform a FlashCopy for the first time, select either **Start** or **Flash** as shown in Figure 5-29.

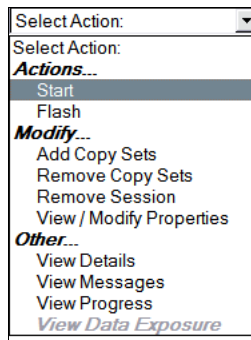


Figure 5-29 FlashCopy Session commands

4. If you select **Start**, the state of the Session goes from **Defined** to **Preparing** and then ends up with the state **Prepared**. It does not trigger the actual copying of the FlashCopy, but it is prepared to perform the FlashCopy. This puts the source volume in write-through mode, and this increases the latency since no write data is held in cache for that disk. The status of the volume stays like this until you trigger the FlashCopy process or stop the FlashCopy relationship. This is similar to the SVC FlashCopy GUI command **Prepare a Consistency Group** or the CLI command:

```
svctask prestartfcconsistgrp
```

5. If you select **Flash**, the session goes from **Defined** to **Preparing**, from **Prepared** to **Target Available** (Copying/Copied). Depending on how many FlashCopy Copy Sets you defined, the time you initiate the Flash command until the actual FlashCopying is performed can vary. If you want to make sure you know when the copy is started, use the Start command the first time. The Flash command is similar to the SVC FlashCopy GUI command **Start a Consistency Group** or the CLI command:

```
svctask startfcconsistgrp -prep
```

6. After you start the FlashCopy you can follow the progress of the copying by selecting **View progress** action under **Sessions**. The TPC-R only shows you zero percent until the copy is done for the Session. This changes to 100 percent. The TPC-R does not go and ask the SVC CIMOM agent for status on the copy progress (See Figure 5-30 on page 193 and Figure 5-31 on page 193).

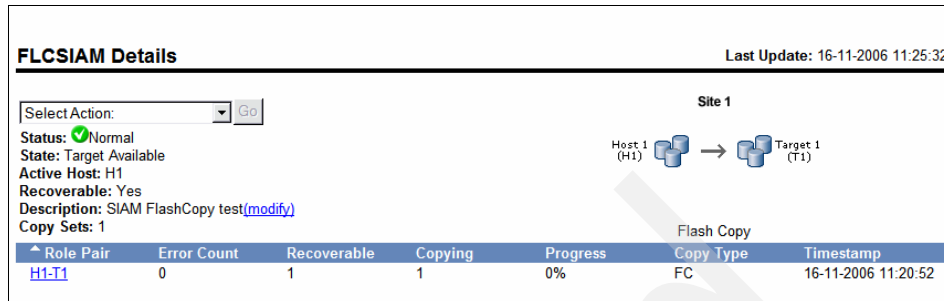


Figure 5-30 FlashCopy Session zero percent progress

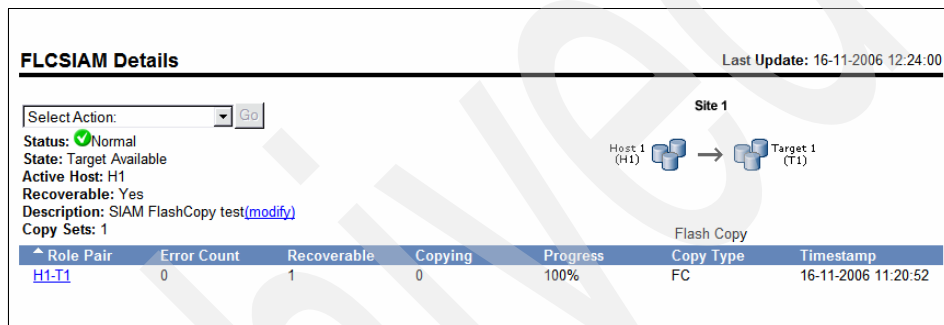


Figure 5-31 Flashcube Session 100 percent progress

5.4.8 Command Line Interface

You can use the *Command Line Interface* (CLI) to make scripts or to perform the same tasks as you perform when using the GUI.

The CLI is intended to be used only on the TPC-R server, while the GUI can be accessed from any computer that has a supported browser. It is possible to implement the CLI software on another computer. To do this you need to copy the files in the TPC-R installation directory to the other computer. After you have done this you might need to change the information for the server that the CLI uses. The information is located in the file repcli.properties in the directory CLI, under the TPC_R installations directory.

1. To activate the CLI, start the csmcli.bat program under the TPC-R installation directory. If you installed TPC-R on a Windows server, you get a prompt starting with *csmcli>*.
2. Type *help* at the command prompt, to get a list of the available commands you can execute (See Figure 5-32 on page 194).

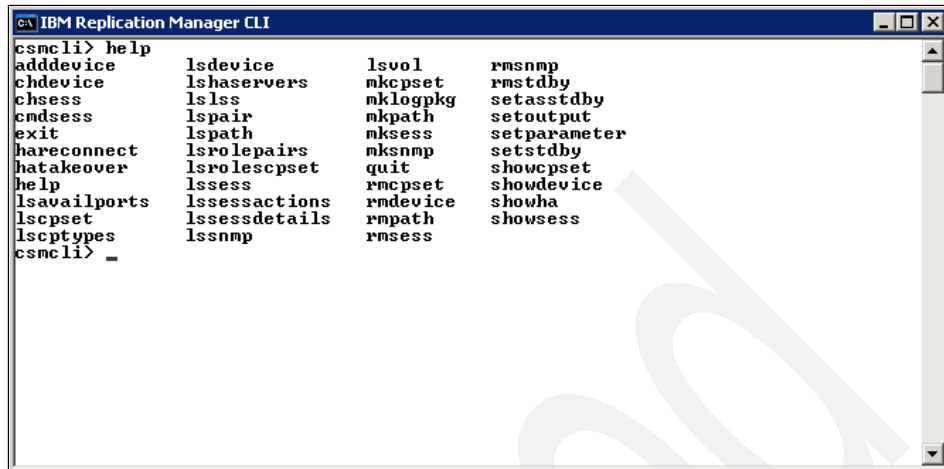


Figure 5-32 List of TPC-R CLI commands

5.5 Volume Shadow Copy Service

Customers rely on applications like Microsoft Exchange, Microsoft SQL Server™, and various internet/intranet applications as key business applications. Even short periods of unexpected downtime have a serious impact. The capability of backing up and restoring data quickly and consistently is essential. The problems in backing up data include the fact that backup jobs frequently overflow their backup time window, open files, and open application issues, which ultimately lead to development of *snapshots*.

The capability to take snapshots has been around for several years. Snapshots allow administrators to *snap* a copy of data while allowing applications to continue running. Applications may pause just long enough to allow the disk system to create the snap copy. These volumes can then be mounted to another server for backing up to various storage devices. Snapshots, however, depend on hardware and software compatibility, and that is in itself problematic, possibly introducing errors that are difficult to track and resolve. In addition, the vendors need to support various versions of SQL Server, Exchange, Windows files, and applications.

Microsoft developed *Volume Shadow copy Service* (VSS), a common framework to resolve these issues. Available in Windows Server® 2003, software and hardware vendors now have a common interface model for generating snapshots. The VSS framework specifies how three distinct components should interact. The three different components are the requestor, writer, and provider. The requestor is typically the backup software. The writer is the application

software such as SQL server or Exchange, that pause to allow the snapshot to be taken. The provider is the specific hardware/software combination that generates the snapshot volume.

5.5.1 Considerations for basic SVC and VSS

To implement VSS support for the SVC the following is needed:

- ▶ SAN Volume Controller
- ▶ SAN Volume Controller Master Console
- ▶ IBM VSS Hardware provider
- ▶ Microsoft Volume Shadow copy Service

The benefits of using the VSS function together with the SVC from an automation point-of-view, is that you improve your applications availability when performing backup/restore using the SVC FlashCopy function (the PiT copy). Not only does this increase the function of your application availability but you can also use it to improve the security of the data you backup. When you use normal internal PiT copy functions in a disk subsystem you are limited to this disk subsystems' possibilities and costs. With the SVC you can utilize another and maybe cheaper storage to hold your copied data, and also by doing this, ensure that your data is safe even if you encounter a total disk array error. Your data is available in the second disk storage system.

In section 6.4, "Microsoft Volume Shadow Copy Service" on page 205 there is an example of how to configure VSS and SVC to improve the backup/restore possibilities.

5.6 Logical Volume Manager versus Global Mirror, Metro Mirror, and automation

Logical Volume Manager (LVM) functions are often used to mirror data between disks under server control.

When you use LVM to perform and control the mirroring, you do not have to implement any kind of automation. The LVM on the server should automatically use both volumes, if they are available. Ensure that the mirroring is synchronized, and if you lose access to one of the disks, LVM makes sure the application does not notice it and keeps running. When the volume comes back online, LVM automatically re-synchronizes the volumes. There is no need for configuring any automation when using LVM, everything is controlled by LVM.

In a very unstable environment you may lose data or get invalid data. One kind of error situation that can cause this is when the server keeps losing a volume and it comes back online. A reason for this can be that the storage subsystem

reboots many times, caused by a microcode error or an error in the SAN, which from a server point-of-view looks like it has lost a volume. The LVM can end up in a situation where it cannot obtain the volume that has the data to mirror and does not know what data to overwrite when it synchronizes.

When you use Metro Mirror or Global Mirror solutions, where the mirroring is controlled by the disk subsystem, you need to configure automation because the servers do not know anything about what is happening at the storage level. The application and servers are not aware that the data is being mirrored, and often the server does not know about the secondary disk subsystem. So when something goes wrong at the storage level, you need to inform the server about this, and tell it how to react.

From a server point-of-view it is an easy solution to implement and control, but you also have to consider the need for server CPU cycles needed to perform and control the mirroring. The bandwidth between the two disk arrays has to be larger than that required for Metro Mirror or Global Mirror. You do not have the option to do a synchronized mirroring over long distances or include FlashCopy at the secondary array to make a consistent copy to recover from in case of a logical error.

With the version of SVC at the time of writing, the Global Mirroring solution does not include the FlashCopy option for SVC, which is only available for DS6000 and DS8000. A disk subsystem is designed to handle data, and solutions also have an advanced remote copy function. The SVC controls the data that is being mirrored, both for Metro Mirror and for Global Mirror, and it uses bit map tables to keep track of data that is not mirrored or that needs to be re-synchronized due to an incident. The SVC's method is more reliable than the LVM that does not have a table to keep track of the data that is not synchronized. LVM needs to analyze every time a disk has an error and when an I/O was not completed.

For more information about LVM versus Metro Mirror and Global Mirror and the considerations to be aware of when selecting the solution to secure data, see 6.6, “LVM versus Metro Mirror, Global Mirror, and DR” on page 222.

5.7 Scripting

Depending on the demands, scripts might fulfil the needs for how to automate functions and tasks. Scripts to improve the automation in your Business Continuity environment are easy to do, but may be more difficult to manage.

You need to have well documented procedures for all changes you perform in your environment; otherwise, you could end up in a situation where the scripts you configured for automation are the reason your data is destroyed or lost.

Depending on the script, you need to change it every time you add, delete, or move a volume host or SVC component.

In the environment in this book, you can make scripts using TPC, TPC-R, and SVC functions, that can improve your automation.

As an example, you can define a script that is triggered when TPC for Data finds out that the server runs out of space. The script can then send a command to the SVC to expand the volume. When the command returns with the status completed without errors, the script will execute the command to expand your file system on the volume just being expanded.

If you install the CLI tool for the SVC, you can execute scripts using the SVC commands directly from the server.

5.7.1 Scripting tools for SAN Volume Controller

IBM has developed a SVC tool that not just utilizes the SVC CLI commands, but gives you the possibility to use Perl scripts to manage your SVC Cluster. The tool is free and can be downloaded at the following Web location:

<http://www.alphaworks.ibm.com/tech/svctools>

Scripting Tools for SVC extends Perl to provide a powerful and flexible scripting environment and simplifies the automation of complex tasks within SVC. An example script is included along with the core module to showcase one such *solution*, the *rebalancing* of the virtualized storage across the underlying storage. This solution allows customers to maximize the use of their underlying storage.

Recovery solutions

There are different types of errors that can occur in an IT installation. In this chapter we show some of the possibilities as to how you can use the SAN Volume Controller to improve your Business Continuity environment. We discuss different ways to protect your data and fulfil your Business Continuity demands.

6.1 What is Disaster Recovery

Disaster Recovery (DR) is the ability to recover from unplanned outages at a different site, usually on different hardware. There are different levels of errors that result in some kind of recovery. Not all errors need to involve DR, but general IT recovery does not always include a site failover.

A DR solution involves a whole company, not only the IT environment, but also human resources, production facilities, physical location, communication lines and more. We only cover *IT Disaster Recovery* (ITDR). ITDR only involves the IT environment, such as the infrastructure (LAN, SAN, WAN), data, servers, and applications.

Many implement an IT environment with a DR solution based on a two-site solution, where data is mirrored between the two sites and where the application can move from one site to the other, in case of a total site failure. If a total site failure occurs, caused by a fire, flood, earthquake, or similar, it does not only affect your IT environment, but also all other facilities at the site. This could be offices, production machines, labs, communications lines and more. So a DR solution where only the IT is protected, might not be enough to protect your company. You may have your data and applications up and running, but you may not have any place where your employees can work with the data or keep production running.

An ITDR can occur on different levels. It may only be necessary that one application with corresponding data performs an ITDR, while other systems, servers, and storage keep running as normal. We are not talking about a total site failure here because that demands a total DR procedure to be activated.

Even if the computer room on the primary site is destroyed by fire, water, and so forth, it might only be the IT environment that performs an ITDR, while human resources, production, and more keeps running as normal.

6.2 Backup/restore

One of the most common tasks in an IT environment is to secure data by taking backups. When you try to find the best backup solution that also fulfills your Business Continuity demands, you also need to make sure that the recovery fulfills your Business Continuity, Recovery Time Objective (RTO), Recovery Point Objective (RPO), or ITDR demands. Often people focus only on the backup solution, such as the time it takes to perform a backup and what the impact is on the application.

Another thing to consider is the recovery time in case a restore is needed. Do you need to restore all data or just some files or database tables? Is it a total disk error, and where do you have the data? The restore time depends on where the backup data is located, how fast you can get access to it, and if you can use the data in your backup from which to restore.

Backups are taken to avoid losing any data in the case of an error. Depending on your environment, the need to perform a restore of data can vary. If you have an IT installation with only one location, a server or disk subsystem error often results in a restore of data. But if you have an IT environment spread over two locations and your data is replicated between the two locations, then you might not need to restore your data if a disk subsystem fails. Because you already have the data located on the secondary site. If the error you have is a logical error, such as a corrupted database, you might need to restore the data from your backup even if you have the data mirrored between two locations.

For example, depending on the method you use to mirror the data between the two locations, the corrupted database may also be on the mirrored site, and therefore the data is not usable from there anymore. If you use an asynchronous mirroring solution, there might be a delay before the error is copied to the mirrored site.

6.2.1 Tape or disk backup

A backup solution that was used for many years is based on tape. Tape is a very cheap medium to store data on—if you look at the price per GB. Others think that tape today is outdated, and they only rely on disk backup solutions. One of the benefits with tape, other than price, is that it is moveable and independent of hardware error. By independent of hardware we mean an error in the microcode/firmware or something else that destroys the configuration or function of the tape library. You can still move your tape to another tape library without losing any data that is on the tape.

If the same kind of error occurs in a disk-based backup solution, you might lose all the data stored on the disks. You may not be able to restore this data anymore. If you have archived data you might have lost that data forever. This risk can be eliminated by having a second disk-based backup solution, where the two mirror data across disk subsystems.

A benefit of using disk for backup is that the recovery time might be shorter than using tape. The speed of most disk today outperforms a tape.

An approach to solve both issues is to backup data to disk, and copy or move it to tape. This way you have your data secure and available, while you also have a quick restore solution. In order not to have too much duplicated data on many

medium, you can, for example, have all your daily or weekly backup data on disk. On tape you can have older backup data. This approach makes sure that the most recently used data can be restored fast, while older data might take a longer time, but the probability that you need to restore the older data is not as likely.

If you only use tape for your backup solution, and do not keep any backup data on disk, the impact on your Business Continuity solution is greater in a restore situation than if you had the backup data on disk. It normally takes longer to restore data from tape than from disk, even though you may have an acceptable Business Continuity solution when performing your backup.

If you use a point-in-time (PiT) copy function like FlashCopy, you might improve the Business Continuity of your applications compared to tape or even the disk-based backup solution. The FlashCopy implementation, depending on the solutions, can improve the Business Continuity when performing backup and restore.

6.3 Point-in-time copy

One of the biggest challenges when using point-in-time (PiT) copy functions is to ensure data integrity. Some applications can suspend their writes and ensure that their data is consistent, but this is only from an application point-of-view. Data that the application believes is written to disk, could still be located in the operating system's cache on the server, and not in the disk subsystem. When performing a PiT at the disk subsystem, it might not be able to restore and use the data directly; therefore, you are depending on the application's recovery functions. For more information see section 5.1.2, "Considerations for Point in Time copy" on page 159.

You can use the PiT copy function in a Business Continuity solution to bring the backup time down to a minimum, so the application can be available as much as possible. If you use a disk subsystem that performs the PiT copy for you, and this PiT copy does not make a copy of all the data, your recovery possibilities might not be the same. Some disk subsystems claim that they do not need to copy all the data, just keep track of pointers. When performing a PiT copy, the subsystem freezes the pointers and locks all blocks from being over written. So new or changed data is being written at a new place on the subsystem, preserving the original data. In this case you can restore to the time you did the PiT copy, as long as your error is not related to an error in your disk subsystem. If the error you get is caused by failing physical disks in the subsystem or that your subsystem for some reason dies, you do not have a copy of your data anymore, except if you took a copy using a backup software solution.

6.3.1 FlashCopy

The FlashCopy function in the SVC enables you to perform a full copy of your volume. The SVC needs to have a target FlashCopy volume that has the same size as the source volume that is going to be copied. The SVC can run in two different FlashCopy modes: A mode where all data is copied from source volume to the target volume, or a mode where only the data that changes is copied, and the data copied is the original data that was valid when the FlashCopy was initiated, not the newly changed data. See Figure 6-1.

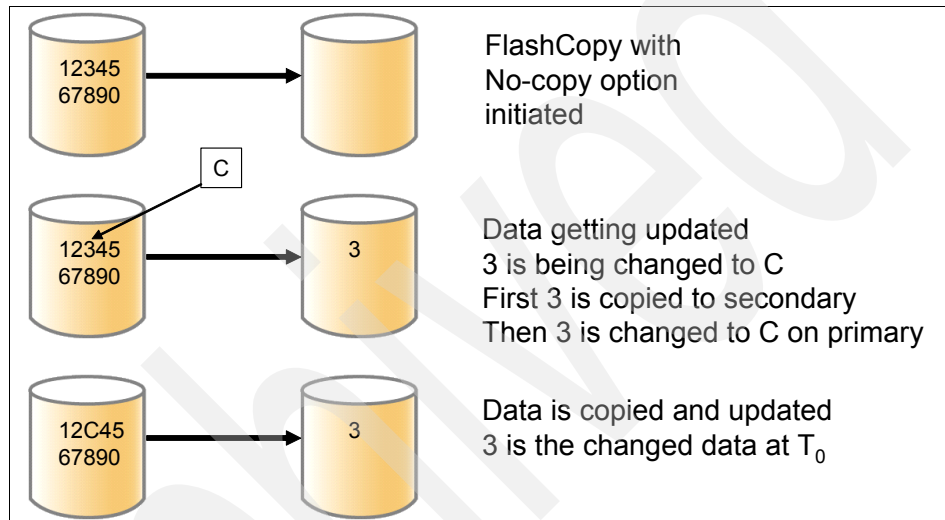


Figure 6-1 FlashCopy NoCopy

No matter what type of FlashCopy modes you use on the SVC, you have access to the target volume as soon as the FlashCopy finishes the initiation. Often this takes a few seconds. At the target volume you see all the original T_0 data, if it is copied or not. The server attached to this volume does not know whether the data is copied or not because it has access to all the data. You can then perform the backup using this volume as it can be mounted on your backup server or another server responsible for performing the backup.

Can FlashCopy replace tape backup

If you do not take a backup to tape, you might have some situations where you lose data because you cannot perform a restore. Because the purpose of taking a backup is to ensure that you do not lose data, we can only recommend that you still backup to tape.

If you perform the same PiT copy every day to the same target volume, then you might risk losing data, and you might not be able to recover if you get a major

disk error while you perform your PiT copy. From the time you issue your PiT copy until all data is copied to the target VDisk, and your backup is completed and secured, you do not have a copy of your data. In this time frame your data is unprotected since you do not have a copy of your data. You could use many PiT copy targets, but even this might not help you in securing your data.

If your disk subsystem gets a microcode error or loses the configuration, you not only lose your data, but also your backup data, even if the data is saved on many PiT volumes. In a tape library, if you get a microcode error, or the library loses the configuration, you might not be able to access your backup data until you have recovered from the error. But the data on your tape is not lost, and you therefore have not lost your backup data. Some disk subsystems today use disk technology to secure your data, while it emulates a tape library. Again, if the disk subsystem gets an error, as mentioned previously, you can end up in a situation where you have lost all your backup data.

The reason for using FlashCopy with your backup solution is to minimize the backup time from an application point-of-view and to have the ability to perform a quick, total recovery from disks, if possible, to minimize the recovery time.

An important element is to ensure that you have the ability to restore—a restore that contains valid and useful data. The most important thing to ensure when using FlashCopy is that the data on the volumes is in a consistent state, and can be used directly, and as fast as possible. If you can ensure that the data in the SVC is in a consistent state, you can use the FlashCopy function as part of a backup and recovery solution.

Restoring from a FlashCopy volume

When you perform a FlashCopy you can restore your data using different techniques.

You can perform a *flash back*, where you stop your server from accessing the volume, and you make another FlashCopy in the opposite direction. As a result, your original FlashCopy target volume is now your source volume. Your server can remount the same volumes as it had before, but now with the backup data on. If you use this restore function, the FlashCopy process needs to have copied all the data from the original T_0 source disk to the target disk. Using this technique you do not need to change any configuration on your server or on the SVC.

If you use a FlashCopy target to restore from, and you use the FlashCopy function to perform the restore, you restore the whole volume or volumes to the point when you issued the FlashCopy backup. You cannot restore single files, and you lose data that was added or changed on these volumes when you performed the restore using the FlashCopy function.

You can also choose to unmount the volume at the server and mount the target volume from your FlashCopy. Change the host volume mapping on the SVC to perform this. You can use this method even if all the data is not copied to the target FlashCopy volume yet. You need to change the configuration on your server and on the SVC, and make a new FlashCopy configuration on the SVC to perform the backup copy for the next time. You still cannot restore single files, and you still lose data that was added or changed on the volume when you perform the restore using the FlashCopy function.

On some operating systems, such as Windows and AIX, you can mount both the FlashCopy source and the target volume on the same host, and use file copy functions to restore only the files you need to restore. For information about how to mount the FlashCopy target volume on the same server where the source FlashCopy volume is located, read *IBM System Storage SAN Volume Controller*, SG24-6423.

6.4 Microsoft Volume Shadow Copy Service

One of the biggest challenges when using PiT copy functions such as SVC FlashCopy is to ensure data integrity. Some applications can suspend their writes and ensure that their data is consistent, but this is only from an application point-of-view. Data that the application believes is written to disk can still be located in the operating system's cache on the server and not in the disk subsystem. Performing a PiT copy without all the data written to the disk results in corrupted data, or lost data transactions. The Microsoft Volume Shadow Copy Service (VSS) solves this problem by being the interface controlling the applications that are aware of the PiT function. They suspend their writes, while VSS makes sure that the Windows operating system does not keep data in the cache on the server. When this is done VSS initiates the FlashCopy function on the SVC.

This ensures data consistency, and automates the backup and restore procedure. It also ensures that the application is available all the time, even when performing backup.

The benefits of using the VSS function together with the SVC is not only that you improve your application's availability when performing backup and restore, but also that the SVC supports different disk subsystems as the target for the FlashCopy (the PiT copy). When you use normal internal PiT copy functions in a disk subsystem you are limited to this disk subsystems possibilities and costs. With the SVC you can utilize another and maybe cheaper storage subsystem to hold your copied data. Also doing this ensures that you data is safe even if you encounter a total disk array error. Your FlashCopy data is available in the second disk storage system.

The SVC provides support for the Microsoft Volume Shadow Copy Service.

To enable support for VSS you must install IBM TotalStorage support for Microsoft Volume Shadow Copy Service (IBM VSS Hardware Provider).

For more information about Microsoft Volume Shadow Copy service, refer to the following Web site:

<http://www-1.ibm.com/support/docview.wss?uid=ssg1S4000342>

6.4.1 How it works

The IBM VSS Provider interfaces with the Microsoft Volume Shadow Copy Service and with the CIM Agent on the master console, seamlessly integrating FlashCopy on the SVC with an initiator such as a backup application issuing a snapshot command and subsequently performing the backup from the FlashCopy target VDisk to create a PIT data consistent backup.

The Microsoft Volume Shadow Copy service handles quiesce of the applications and OS buffer flush prior to triggering the FlashCopy on the SVC, and when done it thaws the quiesce and the application I/O is resumed.

To enable SVC support for VSS, we install the IBM TotalStorage hardware provider on the Windows host.

At the time of this writing IBM Tivoli Storage Manager (TSM) utilizes the SVC FlashCopy function and the VSS function to enhance the backup possibilities for Microsoft Exchange, by performing an automated and integrated backup. To utilize this possibility you have to install the IBM Tivoli Storage Manager for Mail - Data Protection for Exchange.

Data Protection for Exchange performs online backups and restores of Microsoft Exchange Server storage groups to Tivoli Storage Manager storage and local shadow volumes. You can perform backups and restores using a command-line or graphical user interface (GUI). VSS operations require Windows 2003. You must install Data Protection for Exchange on the same machine as the Exchange Server, and it must be able to connect to a Tivoli Storage Manager server. Data Protection for Exchange also supports operations in an MSCS environment.

Data Protection for Exchange VSS operations use the Tivoli Storage Manager Backup/Archive Client and Microsoft Volume Shadow Copy Service technology to produce an online snapshot (point-in-time consistent copy) of Exchange data that can be stored on local shadow volumes or on Tivoli Storage Manager server storage. Data Protection for Exchange also supports traditional backup operations, which can be used in conjunction with the VSS backups. All TSM backup operations can be completely automated.

Data Protection for Exchange provides the following key advantages:

- ▶ **Fast Recovery**
Perform very fast restores from Exchange Server backups.
- ▶ **Fast Backup**
Performs fast, online backups of the Exchange Server.
- ▶ **Off-loaded Backups**
Off-load the backup of data to the TSM Server storage pools to another machine.
- ▶ **TSM Management of Snapshot™ Backups**
Have TSM manage the snapshot backups, including complete automation.
- ▶ **Integrated User Interface**
Integrate the VSS snapshot operations and heritage backup API operations into the same user interface.

In this book we show you an example on how to use the SVC FlashCopy function and the VSS function to perform a backup. This is not shown for a Microsoft Exchange installation, but for using the native Windows backup program NTBackup.

6.4.2 Installing the IBM TotalStorage hardware provider

IBM TotalStorage support for Microsoft Volume Shadow Copy Service can be downloaded from:

http://www-1.ibm.com/support/docview.wss?rs=591&context=STCUFBE&context=STCUF97&dc=D400&uid=ssg1S4000531&loc=en_US&cs=utf-8&lang=en

When installing IBM TotalStorage support for Microsoft Volume Shadow Copy Service, follow the instructions in the configuration guide.

In the following section, we go through the steps that you must perform before you can use the IBM VSS Provider:

1. On the master console, do the following:
 - a. Create a Superuser for the IBM VSS provider—if no superuser is created, the default superuser for the master console must be used.
2. Installing the IBM VSS Provider:
 - a. Define the IP address of the master console and the superuser and password to be used by the IBM VSS provider.

- b. Enable access from the Windows host to the truststore file from the master console (we recommend copying it to the Windows host).
3. On the SVC, do the following:
 - a. Create two hosts (VSS_FREE and VSS_Reserved) on the SVC, to be used by the IBM VSS Provider.
 - b. Create VDisk(s) equal in size to the potential FlashCopy source VDisks and map them to the host VSS_FREE.

On the master console

During the install, you are prompted for the IP address to the master console (CIM host) and a valid username with *superuser* credentials.

In our setup we create a user on the master console for the IBM VSS Provider called *vssuser*, as shown in Figure 6-2. Select **Add a Superuser**, and click **Go**.

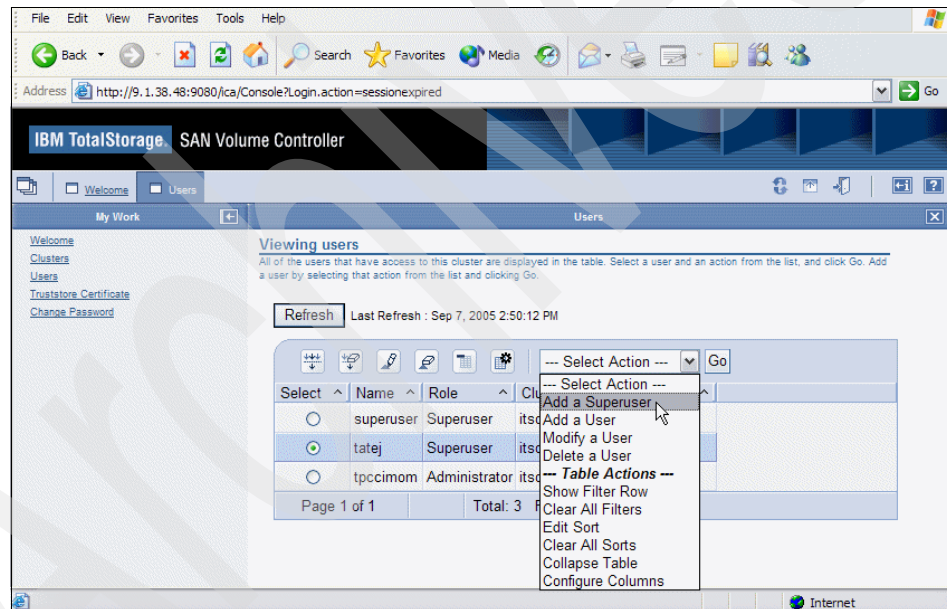


Figure 6-2 Add a superuser

As shown in Figure 6-3 on page 209, we define the username *vssuser* and the password. Click **OK** to create the user.

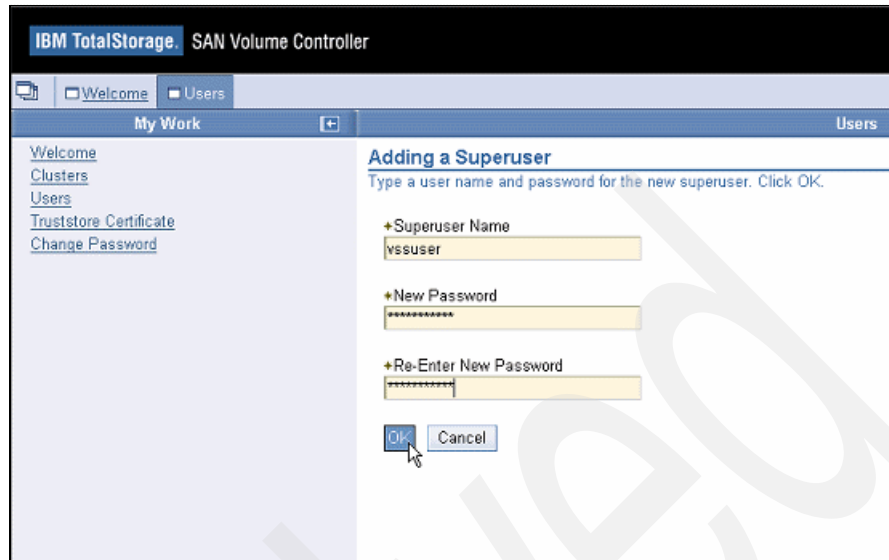


Figure 6-3 Defining the Superuser name and password

When the user is created, we return to the users list as shown in Figure 6-4, and the vssuser is ready to be used by the IBM VSS Provider.

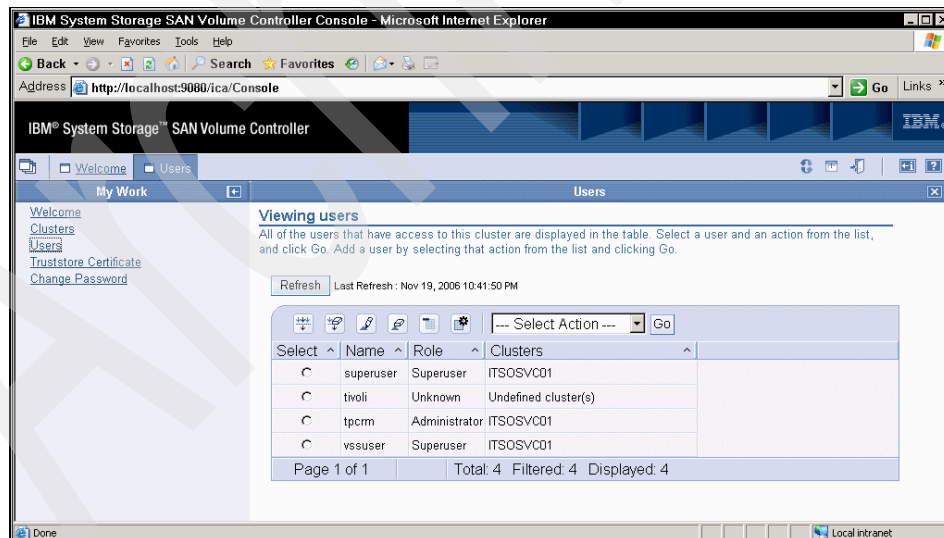
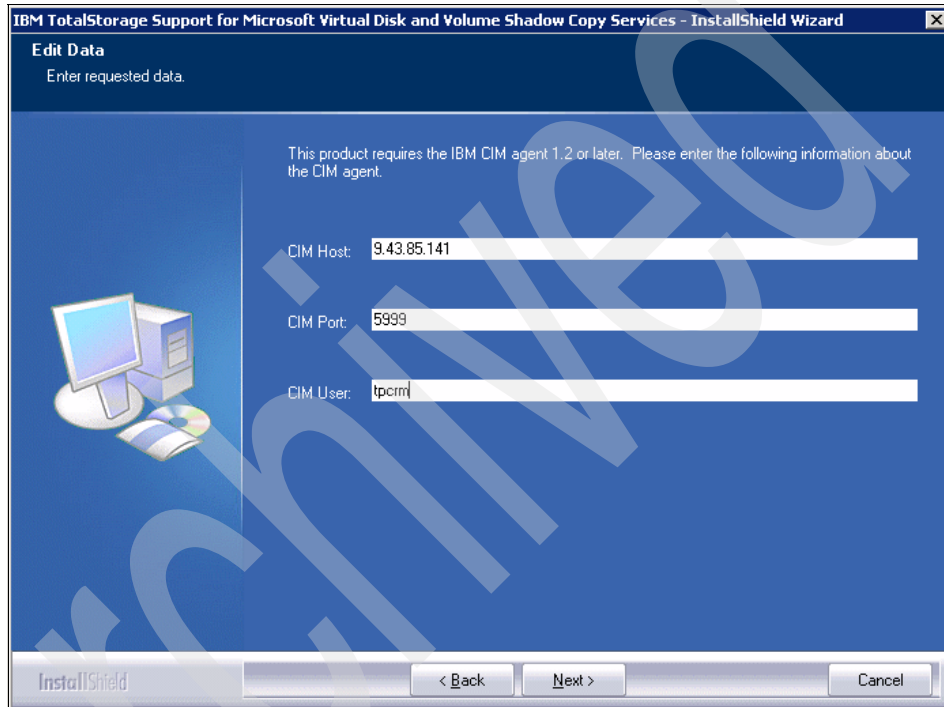


Figure 6-4 Viewing SVC users

On the Windows host

Following is a subset of the Windows displayed when installing the IBM VSS Provider.

During the install of the IBM VSS Provider, we are prompted for the IP address of the SVC master console (CIM Host) and the previously created *vssuser* (CIM User). Click **Next** as shown in Figure 6-5.



IBM TotalStorage Support for Microsoft Virtual Disk and Volume Shadow Copy Services - InstallShield Wizard

Edit Data
Enter requested data.

This product requires the IBM CIM agent 1.2 or later. Please enter the following information about the CIM agent.

CIM Host: 9.43.85.141

CIM Port: 5999

CIM User: tpcirm

InstallShield < Back Next > Cancel

Figure 6-5 Defining the CIMOM host and user to be used by the IBM VSS Provider

As shown in Figure 6-6, we enter the password for the CIM agent user vssuser, and click **Next** to proceed.

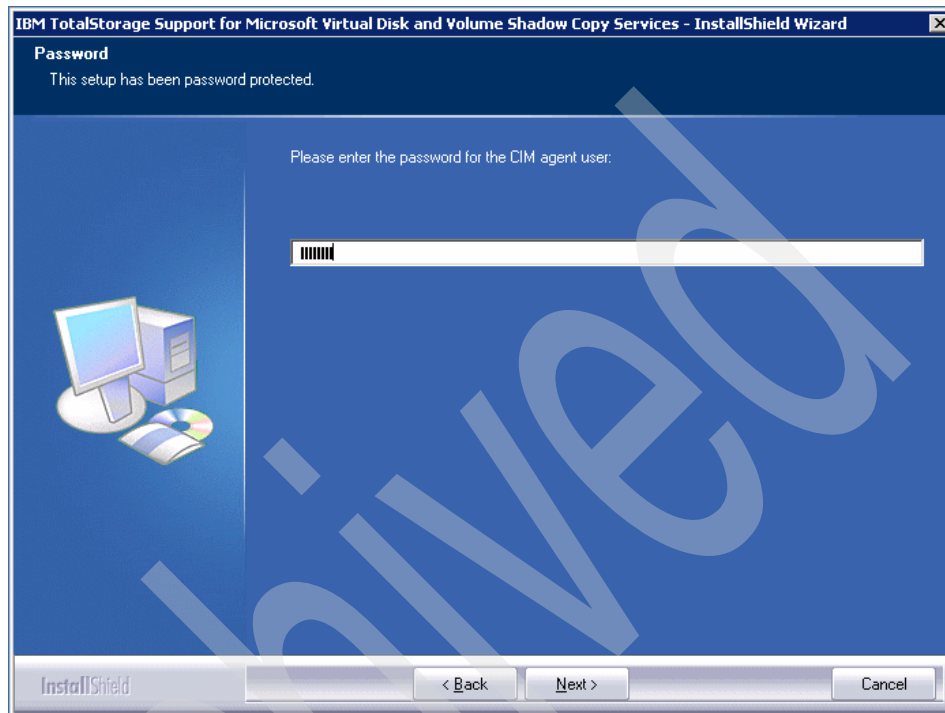


Figure 6-6 Entering the password for the CIM agent user (vssuser)

After selecting whether to use SSL or not when connecting to the CIM agent (this window is not shown), we must enter the location where we copied the truststore file and the password as shown in Figure 6-7.

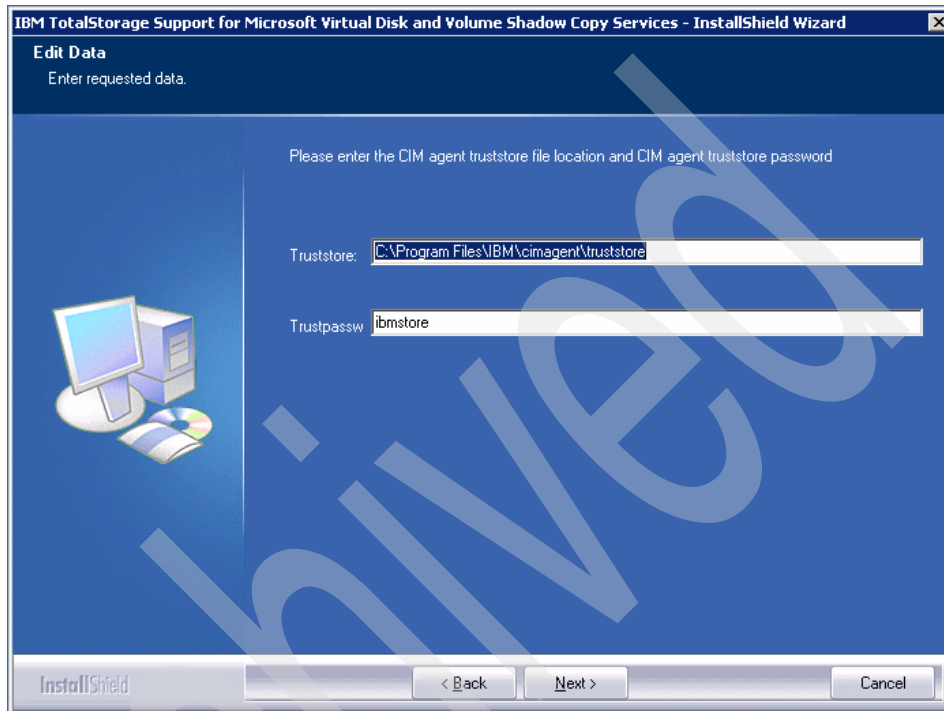


Figure 6-7 Entering the path and password to the truststore file

On the SVC

As shown in Example 6-1, we create the hosts to which we are going to map the VDisks to be used by the IBM VSS Provider using the CLI (this can also be done using the GUI).

Example 6-1 Creating the hosts VSS_FREE and VSS_RESERVED

```
IBM_2145:itsosvc1:admin>svctask mkhost -hbawwpn 5000000000000000 -force -name  
VSS_FREE  
Host id [1] successfully created  
IBM_2145:itsosvc1:admin>svctask mkhost -hbawwpn 5000000000000001 -force -name  
VSS_RESERVED  
Host id [2] successfully created
```

In Example 6-2 on page 213, we map the target VDisks to the VSS_FREE.

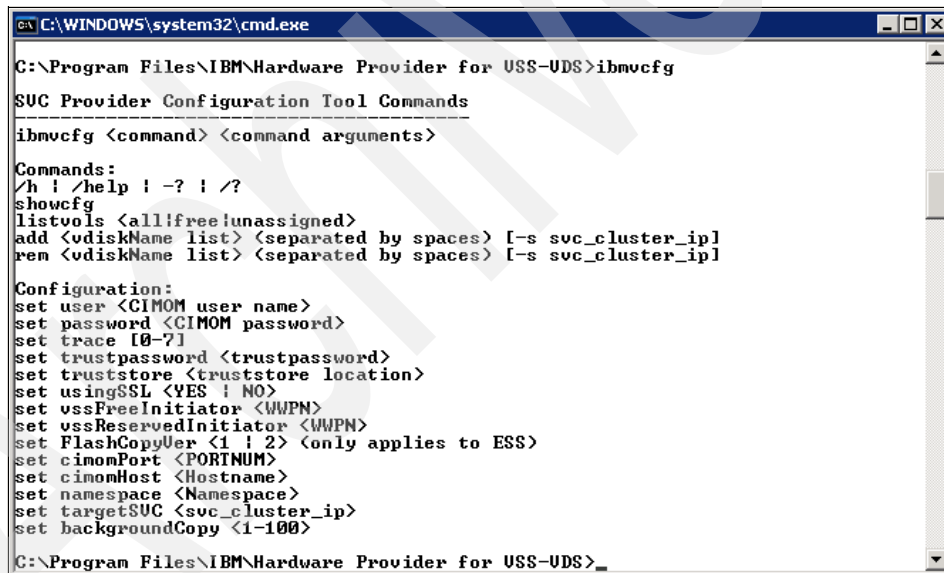
Example 6-2 Mapping the target VDisks

```
IBM_2145:itsosvc1:admin>svctask mkvdiskhostmap -host VSS_FREE VDiskT1
Virtual Disk to Host map, id [5], successfully created
IBM_2145:itsosvc1:admin>svctask mkvdiskhostmap -host VSS_FREE VDiskT2
Virtual Disk to Host map, id [6], successfully created
```

6.4.3 Verifying the VSS configuration

When we install the IBM Hardware Provider, the installation procedure creates the directory C:\Program Files\IBM\Hardware Provider for VSS-VDS. In this directory the SVC Provider Configuration Tool is installed, and it gives you the ability to verify and change the parameters you defined when you first installed the IBM VSS Provider.

You can get a list of possible commands for the SVC Provider Configuration Tool, by executing the command **ibmvcfg**. This gives you the list as shown in Figure 6-8.



```
C:\WINDOWS\system32\cmd.exe

C:\Program Files\IBM\Hardware Provider for VSS-VDS>ibmvcfg

SVC Provider Configuration Tool Commands

ibmvcfg <command> <command arguments>

Commands:
/h /help /-? / ?
showcfg
listvols <all|free|unassigned>
add <vdiskName list> <separated by spaces> [-s svc_cluster_ip]
rem <vdiskName list> <separated by spaces> [-s svc_cluster_ip]

Configuration:
set user <CIMOM user name>
set password <CIMOM password>
set trace [0-7]
set trustpassword <trustpassword>
set truststore <truststore location>
set usingSSL <YES | NO>
set vssFreeInitiator <WWPN>
set vssReservedInitiator <WWPN>
set FlashCopyVer <1 | 2> <only applies to ESS>
set cimomPort <PORTNUM>
set cimomHost <Hostname>
set namespace <Namespace>
set targetSVC <svc_cluster_ip>
set backgroundCopy <1-100>

C:\Program Files\IBM\Hardware Provider for VSS-VDS>
```

Figure 6-8 List of IBM VSS Provider commands

To verify the IBM Hardware Provider configuration execute the command **ibmvcfg showcfg**.

```
C:\Program Files\IBM\Hardware Provider for USS-UDS>ibmvcfg showcfg
cinonHost: 9.43.85.141
username (cinon): vssuser
namespace: \root\ibm
cinonPort: 5999
truststore: C:\Program Files\IBM\cinagent\truststore
trustpassword: ibmstore
usingSSL: true
FlashCopyUser:(only applies to ESS) 2
vssFreeInitiator: 5000000000000000
vssReservedInitiator: 5000000000000001
C:\Program Files\IBM\Hardware Provider for USS-UDS>_
```

Figure 6-9 Verification of IBM Hardware Provider configuration

You can also check that the Microsoft Windows VSS service can communicate with the IBM VSS Provider. In Windows 2003 Server you can get an overview of the VSS commands by typing `assadmin`. This replies with an error, since it needs some parameters, but it lists the commands supported. To verify that VSS can communicate with IBM Hardware Provider, you can execute the command `vssadmin list providers`.

In Figure 6-10 you can see the commands supported by the VSS and that the VSS can communicate with the IBM Provider, after we executed the `vssadmin list providers` command.

```
----- Commands Supported -----
Add ShadowStorage - Add a new volume shadow copy storage association
Create Shadow - Create a new volume shadow copy
Delete Shadows - Delete volume shadow copies
Delete ShadowStorage - Delete volume shadow copy storage associations
List Providers - List registered volume shadow copy providers
List Shadows - List existing volume shadow copies
List ShadowStorage - List volume shadow copy storage associations
List Volumes - List volumes eligible for shadow copies
List Writers - List subscribed volume shadow copy writers
Resize ShadowStorage - Resize a volume shadow copy storage association
Revert Shadow - Revert a volume to a shadow copy
Query Reverts - Query the progress of in-progress revert operations.

C:\Documents and Settings\Administrator>vssadmin list providers
vssadmin 1.1 - Volume Shadow Copy Service administrative command-line tool
(C) Copyright 2001 Microsoft Corp.

Provider name: 'Microsoft Software Shadow Copy provider 1.0'
Provider type: System
Provider Id: {b5946137-7b9f-4925-af80-51abd60b20d5}
Version: 1.0.0.7

Provider name: 'IBM TotalStorage Hardware Provider for USS'
Provider type: Hardware
Provider Id: {d90dd826-87cf-42ce-a88d-b32caa82025b}
Version: 2.5.0.1027

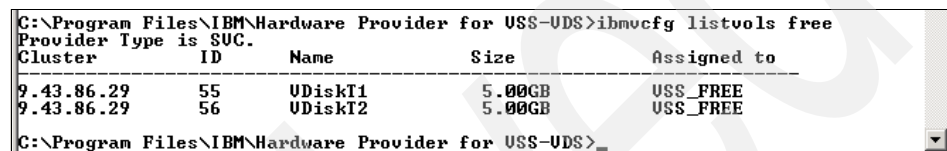
C:\Documents and Settings\Administrator>_
```

Figure 6-10 VSS and IBM verification

To be able to use the VSS function and integrate the SVC FlashCopy function, following are some services that need to be run:

- ▶ Service Location Protocol
- ▶ IBM TotalStorage Hardware Provider

By executing the command **ibmvcfg listvols free** you can verify that the server can see the VDisks we created in Example 6-2 on page 213. Even the VDisks is assigned to the host VSS_FREE. The server can see and use the VDisks because the servers VSS_FREE and VSS_RESERVED are virtual servers, created for the IBM Hardware Provider and for VSS to use. The output of the command we executed to verify the available VDisks can be seen in Figure 6-11. Here you can see the two VDisks that are mapped to the virtual host VSS_FREE.



```
C:\Program Files\IBM\Hardware Provider for USS-UDS>ibmvcfg listvols free
Provider Type is SVC.
Cluster      ID      Name      Size      Assigned to
-----
9.43.86.29   55      VDisk11   5.00GB    VSS_FREE
9.43.86.29   56      VDisk12   5.00GB    VSS_FREE
C:\Program Files\IBM\Hardware Provider for USS-UDS>
```

Figure 6-11 Verify assigned VDisks

6.4.4 Using VSS with NTBackup

In the following scenario, we use NTBackup to initiate a FlashCopy.

Since VSS activation is the default for NTBackup, we select the object we want to back up, and click **Start Backup**. As shown in Figure 6-12, the Backup Job Information box pops up.

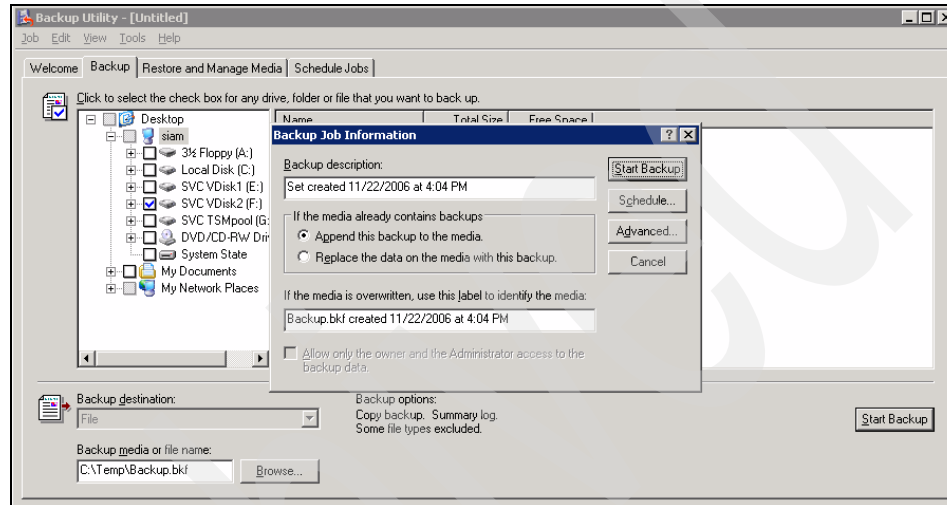


Figure 6-12 Selecting the object to backup and reviewing the Backup Job Information

After verifying the backup job, click **Start Backup** to start the backup job, which subsequently prepares activation of VSS (see Figure 6-13).

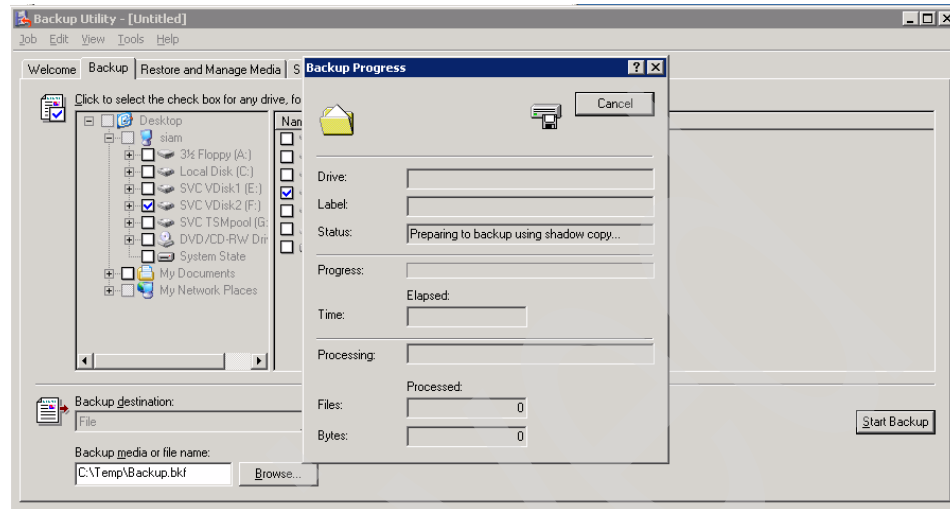


Figure 6-13 Preparing backup using shadow copy

Initiated by NTBackup, VSS checks for available VSS providers and triggers the FlashCopy interfacing with the IBM VSS provider. In Figure 6-14 on page 218 the FlashCopy mapping created by the VSS provider is listed.

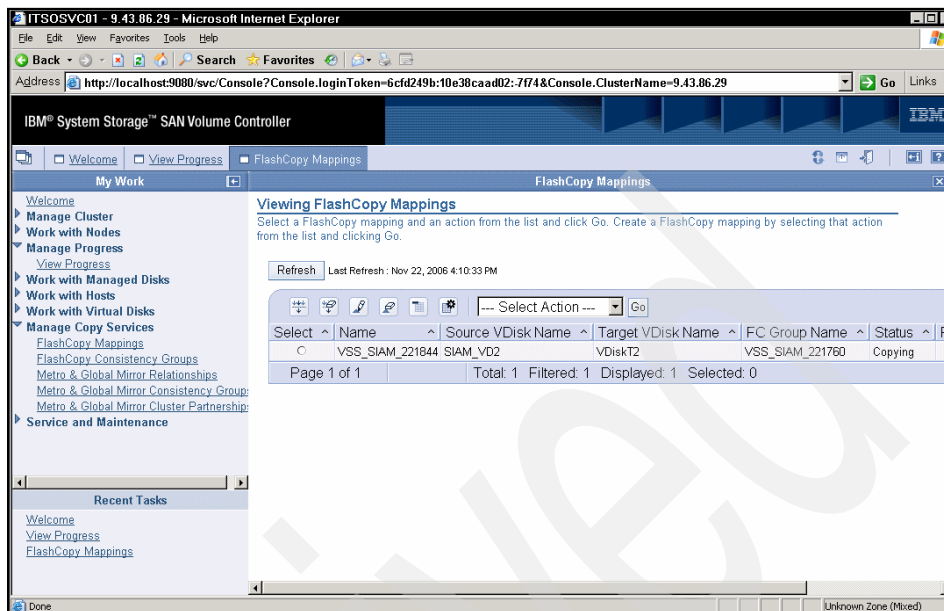


Figure 6-14 Viewing FlashCopy mapping created by the IBM VSS provider

After the FlashCopy mapping is triggered, NTBackup backs up the selected objects facilitated by the IBM VSS provider from the FlashCopy target. As you can see in Figure 6-15 on page 219 the FlashCopy target VDisk is mapped to the server, but with any drive letter assigned.

The NTBackup uses the data FlashCopied on to this disk to perform the backup. The disk has the same name and size as the original disk. The FlashCopy function on the SVC does not just copy the data on the disk; instead, it copies all disk information including the signature on the disk. The FlashCopy copies the VDisk block by block, and does not know the difference between data and ID information on the disk, so everything gets copied. To get more information about how SVC performs the FlashCopy check, refer to *IBM System Storage SAN Volume Controller*, SG24-6423.

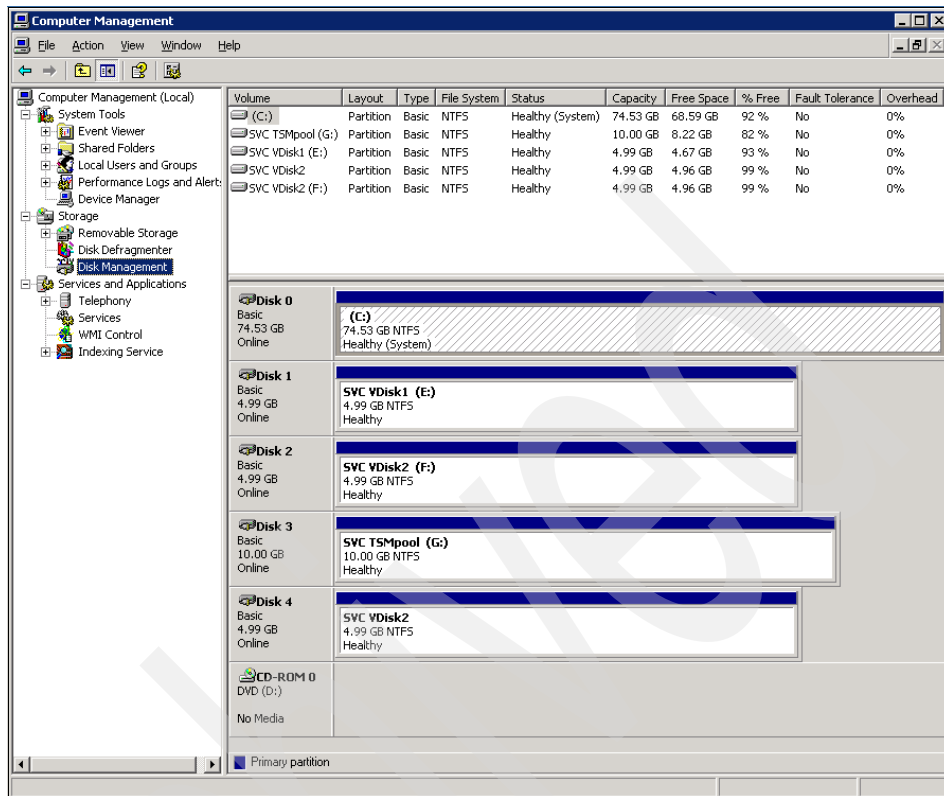


Figure 6-15 FlashCopy target VDisk mounted on server

While the VDisk is mapped, it is not assigned to the virtual host VSS_FREE. In our example you can see this by executing the **ibmvfcg listvols free** command, where you get the output as shown in Figure 6-16. The VDisk is moved to the VSS_RESERVED virtual host, as long as the NTBackup is performing the backup first. After the NTBackup notifies the IBM Hardware Provider that it is done, the VDisk is being moved back to the virtual host VSS_FREE.

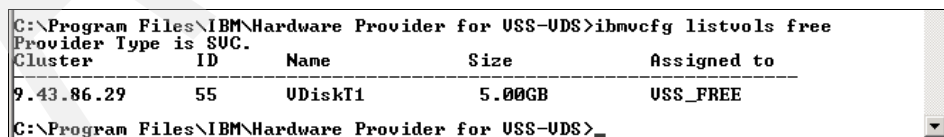


Figure 6-16 VSS_FREE only has one VDisk assigned

Figure 6-17 on page 220 shows that the backup is completed.

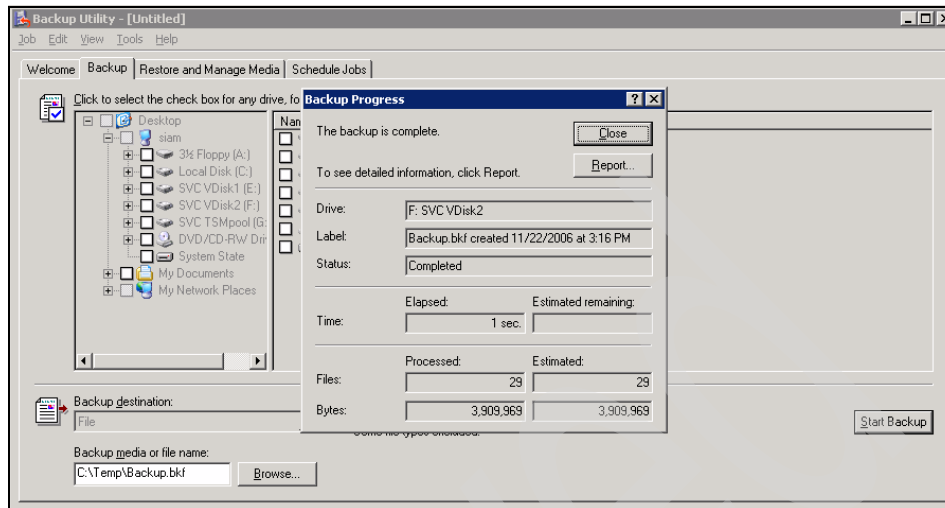


Figure 6-17 Backup completed

As the backup is completed, the IBM VSS provider deletes the FlashCopy mapping while the backup completion is notified by the initiating NTBackup, and the use of the FlashCopy target is no longer needed. Figure 6-18 shows that the FlashCopy mapping is removed.

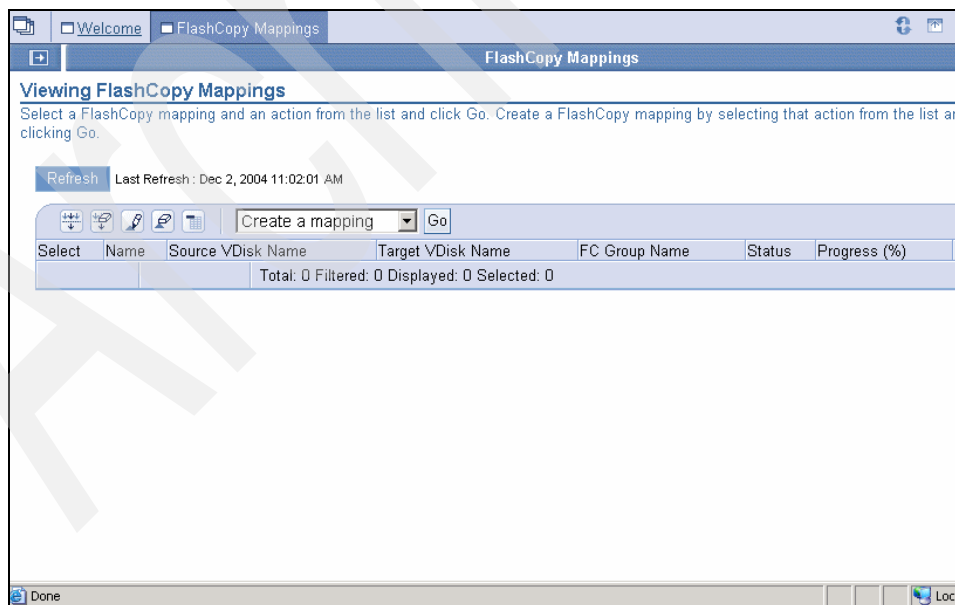


Figure 6-18 FlashCopy mapping was removed upon completion of the backup

6.5 Metro Mirror and Global Mirror

Metro Mirror and Global Mirror are designed to be used in a disaster recovery solution. Both solutions are designed to mirror data between two SVC clusters to secure data in two independent environments. The SVC supports an intra cluster Metro Mirror and Global Mirror configuration for testing, but the Metro Mirror and Global Mirror are not intended to be used in an intra cluster environment.

Metro Mirror for SVC uses synchronous replication, so the host does not receive an acknowledge until both the primary and the secondary SVC clusters have acknowledged the write.

Global Mirror for SVC uses asynchronous replication, so the host receives an acknowledgement as soon the write is secured in the primary SVC cluster. The Global Mirror function can increase the distance in your DR solution because the host does not have to wait for an acknowledge from the secondary SVC cluster.

The SVC keeps track of I/O write to the VDisks, at a block level. The SVC Metro Mirror or Global Mirror can be reversed in case of a disaster, so the secondary SVC cluster becomes the active SVC cluster. When the primary SVC cluster and all the needed systems are active again you can initiate a synchronization. The SVC starts synchronizing the data that was changed from the secondary SVC cluster to the primary SVC cluster. The SVC transfers only the data that was changed since the disaster. When all the changed data is replicated, you can reverse the direction back to normal, and you can do this at a time that is most suitable for you and your installation. You can also keep the environment like this and change the secondary SVC cluster into the primary SVC cluster without changing the configuration or replication.

For more information about SVC Global Mirror see Chapter 3, “SAN Volume Controller Mirroring Solutions” on page 51.

A two site DR solution does not only improve your application’s availability in case of a disaster, but it can also improve availability in a planned outage. To move the active application to the secondary site because of a planned outage is also a test that proves your DR solution is configured correctly. To ensure that your DR solution is reliable, you should perform regular tests. This also gives you an estimate on the RTO of your Business Continuity solution, even this test is performed in a controlled situation.

The methods you must use to perform the test depend on your environment. If you configured some form of automation, like using scripts or you have a cluster solution in your environment, you should activate the automatic failover. If you do not use any automation to control your DR, you can perform the test manually.

The use of Consistency Groups with both Metro Mirror and Global Mirror for the SVC, helps to avoid a rolling disaster. Metro Mirror and Global Mirror together with a Consistency Group ensures that data on the secondary site is usable, even if a VDisk cannot get updated.

In Figure 6-19 we illustrate that the first I/O gets to the secondary site, but the second I/O fails too. The use of Consistency Groups ensures that the third I/O does not get to the secondary site, even if the disk is available. If the third I/O does get to the secondary SVC cluster, this could result in corrupted data.

If the first I/O is a request for updating or adding a record in the log file, the second I/O is the data update or add to the database, and the third I/O is an update completed flag to the log file. If you need to use the data on the the secondary SVC cluster, your database might fail because the data that is believed to have been updated or added, although it is not updated in the database.

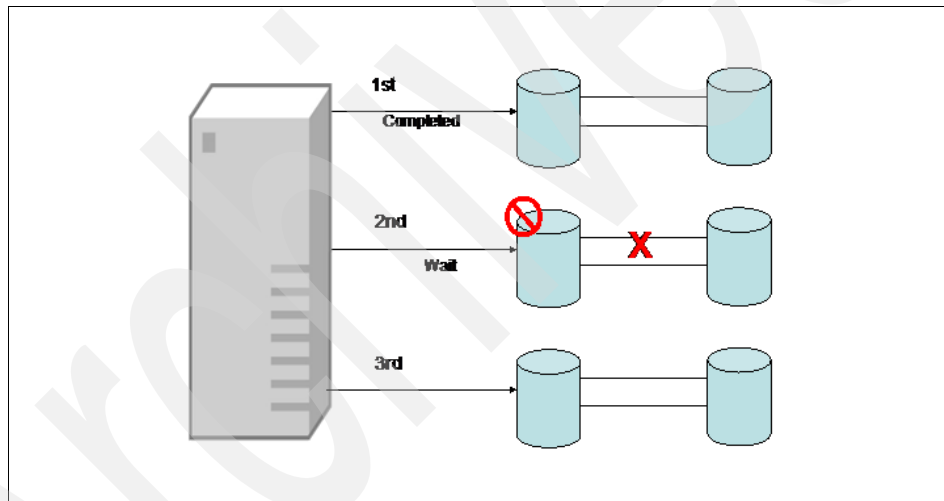


Figure 6-19 Rolling disaster

6.6 LVM versus Metro Mirror, Global Mirror, and DR

Logical Volume Manager (LVM) data is controlled by the servers operation system, while Metro Mirror (Metro Mirror) and Global Mirror (Global Mirror) are controlled by the SVC.

When using LVM in a DR environment, the server is in control of the data reads and writes, compared to Metro Mirror or Global Mirror where the mirrored writes

are performed and controlled by the SVC. The server often does not know about the secondary SVC or the status of the secondary write.

CPU cycle issue

When using LVM the server has to control the mirroring and perform two writes for every write the application performs. This server mirroring uses more CPU resources if the mirroring is performed by the disk subsystem, where the server needs to make one write. If a volume is unavailable, the server needs to perform the comparison of the two volumes and perform the synchronization between the two volumes in a LVM implementation. This also uses server resources, where the same function can be performed by the disk subsystems without the server knowing about it or utilize any resources on the server.

Bandwidth issue

When the server performs the writes, it uses standard SCSI commands to do this, but when you use disk subsystems to perform the mirroring they might use another protocol that is more efficient than SCSI to send the write to the secondary disk subsystem. If your solution is based on LVM the server tries not only to write to both volumes, but also to read from them. So if you have a two-site environment with one disk subsystem at each location and you use LVM, the server sends writes to both disk subsystems. It also tries to read from both. This might have an impact on the connection between the two sites you have. If you use the SVC Metro Mirror or Global Mirror solution, only writes are sent over the connection, while LVM sends both reads and writes that demand that your connection have a higher bandwidth using LVM.

Recovery issue

A very important issue when talking about recovering is data consistency. LVM can mirror data between two volumes, but if you have more volumes in your system, and have data spread across it, you need to ensure consistency. If you use Metro Mirror or Global Mirror you can define Consistency Groups on the SVC.

Consistency Groups in the SVC make sure that the mirrored data is valid and consistent. When using LVM, which does not have a consistency function, you might not be able to use the secondary copy of your volumes. If you have more volumes, and you use LVM to mirror the data volume by volume, and one volume fails at your secondary disk subsystem but the other volume at the secondary subsystem is available, LVM keeps mirroring data to these volumes. At this state you might not be able to recover using the data in your secondary subsystem, since the data is not consistent anymore.

LVM does not ensure the write order of your data on the secondary subsystem, so with LVM you can have a situation where your database cannot start after a

disaster on your primary subsystem. The database might need to perform the recovery, and might not always be able to do this. Often when you configure a database you have one or more volumes for data files, some for index files, and two or more for log files. In this case LVM can issue the write order as requested by the database. Update log files with records that a data table is going to be updated, write the data to the datafile, and update the logfile that the update is completed. When LVM also has to perform the mirroring to the secondary subsystem, and if the subsystem is located at a secondary site, the writes may not get to the secondary site in the correct order. This does not have any impact in normal operations, but if the connection between the two sites is lost, such as if the primary site is destroyed, the secondary site might have invalid data. The data the secondary subsystem received could be only the writes to the log file, so the log file gets the write saying data is going to be updated, and it also gets the write that the data update is completed, but the data file never receives the write update, because this is lost because of the disaster.

SVC ensures that the data is written in the correct order, so the SVC knows that the write to the data file was not performed. Therefore it does not acknowledge the last update to the log file, making the data consistent and recovery possible.

Because LVM does not support Consistency Groups and does not ensure data write order, with LVM you need to manually investigate what occurred before you can perform a recovery.

The SVC keeps track of what data block was changed and can compare this to the secondary disk subsystem and thereby determine the right solution for a recovery, so no or very little data is lost. Where LVM needs to compare and analyze the data to obtain what data is valid, the SVC uses bitmap tables to keep track of the data. This is more reliable and faster than the LVM approach.

Asynchronized mirroring issue

LVM only gives you the possibility to use synchronized mirroring, and the server needs to perform two writes to do the mirroring. If you use the disk subsystem remote mirroring solutions, the server only performs one write, but it still needs to wait for acknowledgment from both disk subsystems, even this is controlled by the primary disk subsystem.

With Global Mirror you can have asynchronized mirroring to the secondary site, which gives you the ability to have longer distances between the sites without adding latency to your applications or to use a connection with less bandwidth like FCP over IP. The distance LVM can support depends on the response time your application can handle and that you find acceptable. Some applications are very sensitive to response time, while others do not have an issue but just lets the user wait until the transaction is completed.

TPC performance metrics for SVC

This appendix lists the subsets of TPC metrics that are applicable to SVC 4.1 and are present in TPC 3.1.3. Please see Appendix F of *IBM TotalStorage Productivity Center User's Guide*, GC32-1775, for a current list of metrics and descriptions for your TPC release.

TPC metrics at SVC node, IOGroup, and cluster level

Table A-1 lists the available TPC performance metrics at the SVC node, IOGroup, and SVC cluster level.

Note: TPC refers to SVC clusters as Storage Subsystems.

Table A-1 Reported Metrics for SVC on the node, IOGroup, and cluster level

Node metric basename	Flavor	Flavor	Flavor
I/O Rate (overall)	Read	Write	Total
Data Rate	Read	Write	Total
Transfer Size	Read	Write	Overall
Response Time	Read	Write	Overall
Peak Response Time	Read	Write	
Queue Time	Read	Write	Overall
Cache Hits Percentage (overall)	Read	Write	Total
Percentage of Cache Hits	Readahead	Dirty Write	
Disk to Cache Transfer Rate			
Cache to Disk Transfer Rate			
Back end I/O Rate	Read	Write	Total
Back end Data Rate	Read	Write	Total
Back end Response Time	Read	Write	Overall
Back end Transfer Size	Read	Write	Overall
Port I/O Rate	Send	Receive	Total
Port Data Rate	Send	Receive	Total
Write Cache Delay ^a	Percentage	IO/Rate	
Write Cache Overflow	Percentage	I/O Rate	
Write Cache Flush-through	Percentage	I/O Rate	
Write Cache Write-through	Percentage	I/O Rate	
CPU Utilization Percentage			

Node metric basename	Flavor	Flavor	Flavor
Port to Host I/O Rate	Send	Receive	Total
Port to Host Data Rate	Send	Receive	Total
Port to Disk I/O Rate	Send	Receive	Total
Port to Disk Data Rate	Send	Receive	Total
Port to Local Node I/O Rate	Send	Receive	Total
Port to Local Node Data Rate	Send	Receive	Total
Port to Local Node Response Time	Send	Receive	Overall
Port to Local Node Queue Time	Send	Receive	Overall
Port to Remote Node I/O Rate	Send	Receive	Total
Port to Remote Node Data Rate	Send	Receive	Total
Port to Remote Node Response Time	Send	Receive	Overall
Port to Remote Node Queue Time	Send	Receive	Overall
Global Mirror Write I/O Rate			
Global Mirror Overlapping Write	Percentage	I/O Rate	
Global Mirror Secondary Write Lag			

a. This is presented in TPC panels as *Write-cache Delay*.

TPC metrics at the SVC VDisk level

Table A-2 lists the available TPC performance metrics at the SVC VDisk level.

Note: TPC refers to SVC Vdisks as Volumes.

Table A-2 Reported Metrics for SVC at the Volume level

Volume metric basename	Flavor	Flavor	Flavor
I/O Rate (overall)	Read	Write	Total
Data Rate	Read	Write	Total
Transfer Size	Read	Write	Overall
Response Time	Read	Write	Overall

Volume metric basename	Flavor	Flavor	Flavor
Peak Response Time	Read	Write	-----
Queue Time	Read	Write	Overall
Disk to Cache Transfer Rate	-----	-----	-----
Cache to Disk Transfer Rate	-----	-----	-----
Cache Hits Percentage (overall)	Read	Write	Total
Percentage of Cache Hits	Readahead	Dirty Write	-----
Write Cache Delay ^a	Percentage	I/O Rate	-----
Write Cache Overflow	Percentage	I/O Rate	-----
Write Cache Flush-through	Percentage	I/O Rate	-----
Write Cache Write-through	Percentage	I/O Rate	-----
Global Mirror Write I/O Rate	-----	-----	-----
Global Mirror Overlapping Write	Percentage	I/O Rate	-----
Global Mirror Secondary Write Lag	-----	-----	-----

a. This is presented in TPC panels as *Write-cache* Delay.

TPC metrics at the SVC MDisk level

Table A-3 lists the available TPC performance metrics at the SVC MDisk level.

Table A-3 Reported Metrics for SVC at the MDisk level

Managed Disk metric basename	Flavor	Flavor	Flavor
Back end I/O Rate	Read	Write	Total
Back end Data Rate	Read	Write	Total
Back end Transfer Size	Read	Write	Overall
Back end Response Time	Read	Write	Overall

TPC metrics at the SVC port level

Table A-4 on page 229 lists the available TPC performance metrics at the SVC port level.

Table A-4 Reported Metrics for SVC at the port level

Port metric basename	Flavor	Flavor	Flavor
Port I/O Rate	Send	Receive	Total
Port Data Rate	Send	Receive	Total
Port to Host I/O Rate	Send	Receive	Total
Port to Host Data Rate	Send	Receive	Total
Port to Disk I/O Rate	Send	Receive	Total
Port to Disk Data Rate	Send	Receive	Total
Port to Local Node I/O Rate	Send	Receive	Total
Port to Local Node Data Rate	Send	Receive	Total
Port to Remote Node I/O Rate	Send	Receive	Total
Port to Remote Node Data Rate	Send	Receive	Total

Relationship between TPC metrics and SVC counters

The following tables present the general relationship between the performance counters reported by SVC and the TPC performance metrics calculated from them. The short identifiers used in the expressions refer to the SVC performance counters as described in Chapter 4.3.3 Per-node statistics of *IBM System Storage SAN Volume Controller*, SG24-6423. Their value is to be considered a difference between two counter readings over the time.

Note: The formula representations shown here are only meant to express the general relationship between the TPC metrics and SVC counters and are *not* to be regarded as arithmetic equations or definitions. Aspects such as proper unit conversion, delta calculation, and aggregation are not fully expressed here. The following tables should be considered informational, and *not* normative.

Table A-5 MDisk level metrics

Metric basename	Read	Write	Total / Overall
Backend I/O Rate	ro/t	wo/t	(ro+wo)/t
Backend Data Rate	rb/t	wb/t	(rb+wb)/t

Metric basename	Read	Write	Total / Overall
Backend Response Time	re/ro	we/wo	(re+we) / (ro+wo)
Backend Queue Time	(rq-re)/ro	(wq-we)/wo	(rq+wq-re-we) / (ro+wo)
Backend Transfer Size	rb/ro	wb/wo	(rb+wb)/(ro+wo)

Table A-6 VDisk level metrics

Metric basename	Read	Write	Total / Overall
I/O Rate	ro/t	wo/t	(ro+wo)/t
Data Rate	rb/t	wb/t	(rb+wb)/t
Response Time	rl/ro	wl/ro	(rl+wl)/(ro+wo)
Peak Response Time	rlw	wlw	-----
Transfer Size	rb/ro	wb/wo	(rb+wb)/(ro+wo)
Cache Hit Percentage	ctrh/ctr	ctwfw/ctw	(ctrh+ctwfw) / (ctr+ctw)
Global Mirror Write I/O Rate	-----	gws/t	-----
Global Mirror Secondary Write Lag	-----	gwl/gws	-----

Table A-7 VDisk level metrics (cont.)

Metric basename	Percentage	i/O Rate
Readahead % of Cache Hits	ctrhp/ctrh	-----
Dirty Write % of Cache Hits	ctwh/ctwfw	-----
Disk to Cache	-----	(ctp+ctrm)/t
Cache to Disk	-----	ctd/t
Write-cache Delay	(ctw-ctwfw) / (ctr+ctw)	(ctw-ctwfw)/t
Write-cache Overflow	ctwfwsh/ctw	ctwfwsh/t

Metric basename	Percentage	i/O Rate
Write-cache Flush-trough	ctwft/ctw	ctwft/t
Write-cache Write-trough	ctwwt/ctw	ctwwt/t
Global Mirror Overlapping Write	gwo/gws	gwo/t

Table A-8 Port level metrics

Metric basename	Send	Receive	Overall
Port I/O Rate	$(het+cet+lnet+rnet) / t$	$(her+cer+lner+rner) / t$	$(het+cet+lnet+rnet + her+cer+lner+rner) / t$
Port Data Rate	$(hbt+cbt+lnbt+rnbt) / t$	$(hbr+cbr+lnbr+rnbr) / t$	$(hbt+cbt+lnbt+rnbt + hbr+cbr+lnbr+rnbr) / t$
Port to Host I/O Rate	het/t	her/t	$(het+her)/t$
Port to Host Data Rate	hbt/t	hbr/t	$(hbt+hbr)/t$
Port to Disk I/O Rate	cet/t	cer/t	$(cet+cer)/t$
Port to Disk Data Rate	cbt/t	cbr/t	$(cbt+cbr)/t$
Port to Local Node I/O Rate	lnet/t	lner/t	$(lnet+lner)/t$
Port to Local Node Data Rate	lnbt/t	lnbr/t	$(lnbt+lnbr)/t$
Port to Remote Node I/O Rate	rnet/t	rner/t	$(rnet+rner)/t$
Port to Remote Node Data Rate	rnbt/t	rnbr/t	$(rnbt+rnbr)/t$

Table A-9 Node level metrics

Metric basename	Send	Receive	Overall
CPU Utilization	-----	-----	busy/t
Port to Local Node Response Time	we/wo	re/ro	(we+re)/(wo+ro)
Port to Local Node Queue Time	(wq-we)/wo	(rq-re)/ro	(wq+rq-we-re) / (wo+ro)
Port to Remote Node Response Time	we/wo	re/ro	(we+re)/(wo+ro)
Port to Remote Node Queue Time	(wq-we)/wo	(rq-re)/ro	(wq+rq-we-re) / (wo+ro)

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

For information about ordering these publications, see “How to get IBM Redbooks” on page 233. Note that some of the documents referenced here may be available in softcopy only.

- ▶ *IBM System Storage SAN Volume Controller*, SG24-6423
- ▶ *IBM System Storage Business Continuity Solutions Overview*, SG24-6684
- ▶ *IBM System Storage Business Continuity: Part 1 Planning Guide*, SG24-6547
- ▶ *IBM System Storage Business Continuity: Part 2 Solutions Guide*, SG24-6548
- ▶ *SAN Multiprotocol Routing: An Introduction and Implementation*, SG24-7321
- ▶ *IBM TotalStorage Productivity Center V3.1: The Next Generation*, SG24-7194
- ▶ *SAN Volume Controller Distance Considerations*, TIPS0599

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

500 44, 134

A

absorption 35
abstraction 28
acceptance 21
access 7, 21, 29, 45, 48–49, 54, 56, 63, 73,
111–114, 143, 159, 164, 172, 174, 182, 196, 198,
204, 207, 213, 215–216, 220
accident 213
activate 192, 234
activation 228–229
active 26, 71, 75, 77, 131, 136, 139–140, 158, 160,
166, 179, 233
adapters 23, 28, 163, 200
address translation 45–46
addressing 46, 71, 138
adds 22, 24, 26, 132, 140, 185, 189
Admin 201
administration 28, 41, 87, 165
administrator 122, 163–164, 201
affordable 4, 14, 46
Agent 218
agent 162, 164, 166, 223–224
AIX 12, 27, 160–161, 165, 171, 193–195, 217
alert 122, 147–148
alerts 121, 147–149, 151, 162
analysis 2, 5, 13, 20–21, 26, 47, 120–122,
135–136, 162, 209
API 165, 219
application
 availability 15
 performance 120, 132–133
application availability 14–16, 163, 206–207
application testing 16
applications 2–3, 6, 19, 22–23, 55–56, 68, 122,
148, 156, 159, 163, 212, 214, 217
arbitration 202
architecture 36–37
area 29, 32–33, 45–46, 156
areas 76, 136, 140–141, 144, 150
arrays 23, 86–87, 208

Asynchronous 22–23, 25, 52, 57–58
asynchronous 16, 22, 24–25, 51–52, 58
Asynchronous remote copy 58
asynchronous replication 22, 61
asynchronously 70
attached 15, 28, 32, 38, 54, 78, 84, 162, 194, 196,
215
attention 26, 34, 134
authority 122
authorized 46, 174
automatic failover 197
Automation 23, 117, 160–161, 207
automation 12, 15–16, 27, 155–157, 219, 233
auxiliary 54, 60, 101, 140
auxiliary VDisk 54
availability 2–5, 13–16, 18–20, 22–23, 25–26, 31,
63, 77, 82–83, 121, 157, 160–161, 163, 194, 199,
205–207, 212, 217, 233

B

background copy 24, 56, 60, 67, 131–132, 138
background copy progress 114–116
background copy rate 138, 140
backup 3, 11, 15, 58, 131, 158–160, 196–197,
206–207, 212–215
backup time 206, 216
backup window 158
balance 13, 23, 26, 68, 122, 132, 136
balanced 13, 136
balancing 6
band 14, 28, 71, 204
Bandwidth 24, 26, 34, 235
bandwidth 23, 25–26, 29–31, 34–35, 57, 67, 79, 83,
86, 117, 132–134, 144–146, 208, 235–236
bandwidth requirements 146
baseline 120, 136, 144
basic 4, 26, 52, 122, 127, 136, 207
BB 38
BB_Credit 37–38
BB_Credit_CNT 38
best practices 85, 87, 117
between 5, 7–9, 12, 18–19, 21–23, 51–53, 55, 62,
66, 121, 123–124, 132, 136, 157–158, 160, 166,

- 183, 185, 212–213, 233, 235–236, 243
- block 15, 42, 55, 96, 138, 142, 144, 168, 205, 230, 233, 236
- blocked 138, 204
- block-for-block translation 96
- broadcast 46
- Brocade 33, 85
- buffer 37–38, 218
 - credits 38
- buffer credit 37
- buffers 37–38, 46
- bus 23, 28, 38, 53, 163, 200
- business
 - objectives 13–14, 20
- business continuity 9–12, 199
- business recovery 2, 14
- business requirements 13, 18, 20, 193

C

- cable 36
- cache 23, 57, 70–71, 120, 133, 135, 137–139, 191, 214, 217, 241–242, 244
- cache algorithm 74
- cache flush 140–141
- capacity 15, 21, 23, 26, 29, 34, 67, 80, 94, 121, 131–132, 134–135, 137, 160–164
- CAW 199
 - patented third heartbeat 202
- certification process 204
- certified 204
- challenge 35, 156, 163, 214, 217
- challenges 44, 47
- changes 20, 44, 52, 77, 120, 147, 166, 189
- channel 31, 34, 62, 66–67
- channels 28, 32–33, 62, 66
- choice 24, 48, 194
- CIM 220, 223–224
- CIM Host 222
- CIM User 222
- CimAgent 218
- CIMOM 122, 166–167, 182, 191, 222
- Cisco 33, 45, 85, 194
- Cisco MDS 9000 194
- Class 1 38
- Class 2 38
- Class 3 38
- classes 69
- CLI 24–25, 90–91, 106–107, 110, 165–166,

- 170–171, 191–193, 200, 203, 209, 224
 - commands 209
- Client 219
- client 159, 200
- clock 2, 40
- cluster 16, 24, 26–27, 29, 44, 52–53, 55, 57, 62–63, 120, 126, 132, 137–138, 143–146, 157–159, 163, 179, 182, 195, 199, 201, 233, 240
 - adding nodes 151
 - creation 63, 84
- cluster partnership 62, 85–86
- clustering 74, 199, 202
- clustering software 194
- clusters 25–27, 51, 59, 62, 64–65, 146, 149–150, 152, 193, 197, 199–200, 206, 240
- combination 9, 13, 29, 200, 206
- combines 201, 203
- command 42, 75, 85, 107–108, 135, 164–165, 170–171, 191, 209, 218, 225–227, 231
- command line interface 209
- command prompt 192
- commit 12, 157
- common interface model 206
- communication 3, 24, 27–28, 57–58, 62–63, 65, 120, 146, 161, 167, 181, 201, 204, 212
- compatibility 78, 206
- complexity 27, 162
- compression 35, 41–42, 44
 - algorithms 46
- concepts 2, 52–53
- concurrent 4, 133, 142, 194
- configuration 26, 56, 62, 65, 69, 82, 84–87, 121–122, 132–133, 144–145, 150, 156, 159, 161–163, 182, 196–199, 201, 203, 213, 216–217, 219, 225–226, 233
- configuration changes 166
- configure 148, 161, 163, 177, 179, 201, 236
- conflicts 46
- congestion 34, 39
- connected 24, 29, 32, 59–63, 182, 200
- connected state 60
- connection 36, 38–39, 45, 68, 81, 133, 146, 177, 182, 235–236
- connections 32, 63–64, 66
- connectivity 29, 33
- consistency 12, 54–56, 120, 141, 159, 167, 170, 234–236
- consistency group 54, 56, 60, 141, 185, 189
- consistency groups 56, 61, 114, 234

- consistent 7, 15–16, 27, 51–52, 55, 141–142, 214, 216–219
- consistent copy 141
- ConsistentDisconnected 61
- ConsistentStopped 60, 76
- ConsistentSynchronized 60
- consolidate 47
- container 167
- continuous availability 13, 16, 19, 199
- Continuous Availability for AIX 195
- Continuous Availability for AIX Management of Metro Mirror pairs 197
- Continuous Availability for Windows 199
- Continuous Availability for Windows in greater detail 201
- control 31, 34, 37, 44–45, 62, 66, 72, 80, 121, 161, 163, 165, 183, 196, 204, 207–209, 234–235
- copy 4, 7, 14–15, 22–24, 51–52, 54–55, 120, 131–132, 157–160, 165, 213–216
- copy bandwidth 145
- copy process 56, 60, 139
- copy processes 67, 81
- copy rate 72, 138, 140
- copy service 23, 136
- Copy Services 15, 17, 23–24, 26–27, 51–53, 173, 200
- copy services 120, 136, 147–148, 163, 165
- copying the grain 72
- copy-on-write process 138
- core 29, 209
- correctly configured 62
- corrupted 160, 217, 234
- corruption 2, 55, 202, 204
- cost 13–15, 18, 21, 28, 32, 45–46, 156
- counters 38, 120, 126, 243
- credits 37
- critical 6, 9, 11, 29, 46, 142, 148, 193, 199
- Current 33, 35, 134, 142
- current 7, 9, 11, 26, 51, 79, 82, 120, 127, 131, 239
- CWDM 29–30, 32

D

- dark 32, 36
- dark fiber 31–33, 81
- data 2–4, 18–19, 22, 51–53, 120–122, 156, 158–159, 211–212
 - consistency 12, 24, 27, 54–55, 137, 159, 165, 217, 234

- data collection 119, 121–122
- data consistency 55–56
- data flow 23, 52, 113
- data integrity 6–8, 201, 214, 217
- data migration 15
- data rate 35, 41, 127, 131–132
- data replication 28, 52, 58, 61
- database 7–8, 11, 84, 86, 121–122, 136, 171, 173, 175, 212, 234–235
- datagram 40
- date 21, 57–58, 60
- debug 204
- decibel 36
- decibel milliwatt 36
- decrypt 44
- default 40, 54, 67, 69–70, 72, 87, 121, 127, 167, 172, 220, 228
- defined 9–10, 19, 53, 55, 64, 121, 127, 137, 162, 173–174, 225
- degraded 146
- delay 35, 62, 67, 69, 142, 213
- deleted 213
- demand 2, 15, 23, 29, 120, 122, 132, 156–157, 212
- dependency 23, 147, 159
- dependent writes 55–56
- design 13, 16, 18, 20, 22, 120, 122, 130, 156
- destage 138–141
- destroyed 156, 208, 212–213, 236
- detected 55, 69, 147, 162–163
- device 32, 34–37, 40–41, 54, 63, 96, 162
- diagnose 146
- diagnostic 204
- diagnostics 204
- different vendors 24, 47
- director 36
- directors 37
- disable 70
- disabled 31, 61, 138
- disaster 5–7, 22, 27, 101, 156, 233–235
- disaster recovery 2, 5, 7–8, 24–25, 195
- disciplines 19
- disconnected 61
- discovery 40
- disk 11, 14–15, 23–25, 53–54, 68, 120, 136, 139, 157, 160, 168, 191, 194–195, 212–216
- disk subsystems 15, 207, 213–214, 216–217, 235–236
- disruption 2, 4, 13, 55, 69, 76, 206
- disruptive 4–5, 15, 45

- distance 23–24, 26, 29, 34–35, 37–38, 42, 44, 62, 68, 70, 80, 83, 107, 120, 131–133, 145, 194, 233, 236
- distances 22–23, 31, 33, 35, 46, 51, 59, 194, 201, 208
- documentation 2, 10
- domain 45, 182
- downtime 6, 15, 44, 197, 199–200, 206
- driver 15, 78
- drops 39
- DS Storage Manager 165
- DS4000 12, 53, 61, 136, 160, 199–200, 206
- DS4500 86, 123
- DS6000 53, 61, 166, 179, 194
- DS8000 53, 61, 166, 179, 194, 199–200
- DWDM 29–30, 32–33, 36–37, 44, 80

E

- E_Port 45
- edge 40
- EE Flow Control 38
- EE_Credit 37–38
- EE_Credit_CNT 38
- efficiency 32, 46, 121
- element 28, 84
- elements 3, 14, 18, 28, 33, 120
- enable 2, 15, 53, 56, 112, 163, 179, 200, 218
- encapsulated 29
- encrypt 44
- encrypted 41
- Encryption 44
- encryption 44
- End to End Credit 37
- Enterprise 53, 162
- error 2, 20, 27, 37, 44, 56, 69, 76, 78, 144, 156, 160–161, 163–164, 213–214, 216, 218
- error detection 34
- errors 69, 81, 86, 146, 157–159, 163, 206, 209, 211–212
- ESS 53, 165–166, 179
- Ethernet 30–31, 40–41
- ethernet 40
- event 5–8, 56, 69, 77, 121, 145, 147, 163, 202
- events 2, 5, 69, 75, 132, 138, 148, 162
- exchange 37, 48, 62
- excludes 69
- expand 2, 164, 209
 - a volume 164

- expansion 164
- exporting 135
- Extended 44
- Extended Fabrics 44
- extenders 38, 80
- extension 21, 23–24, 27, 29, 194, 209
- extent 6, 54, 93, 141, 143
 - size 93
- extent size 93
- extents 53–54, 91, 93–94, 96, 125

F

- F_Ports 37
- Fabric 27, 33–34, 62, 82, 86, 161–162, 166
- fabric 24, 33–34, 36–37, 44, 62, 65–66, 82–83, 85, 162
 - isolation 46
 - local 46, 82
 - reconfiguration 45
 - remote 46
 - services 34, 45–46
- Fabric Manager 162
- fabrics 44–46, 63, 65–66, 80, 82
- failover 58, 156, 158, 160, 166, 193–195, 197, 199–200, 212, 234
- failure situations 4
- Fast Ethernet 30
- fast restore 160
- fault tolerant 4, 18
- FC 29, 37–38
- FCIP 29, 33, 39–42, 44, 80, 133
 - link 41
 - performance 40
 - tunnel 40–41
- FCP 42–43, 236
- FDDI 30
- features 4–5, 14, 81, 120–121, 137, 163, 197, 204
- fiber 30–32, 81
- Fibre Channel 28, 33, 35, 37, 40, 44, 82, 194
 - frame 40
 - interface 40
 - protocol 33
 - routers 40
 - traffic 33, 44–45
- Fibre Channel frame 40
- file
 - server 160, 192, 217
- file system 15, 164, 209

- file systems 15, 161–162, 164
- filter 109, 127–128, 134
- filtering 124–125
- firmware 78–79, 213
- flag 108, 234
- flash 216
- FlashCopy 11, 15–16, 51, 104, 120–121, 130–132, 158, 163, 166, 168–170, 214–217
 - applications 120, 133, 207, 216
 - commands 170
 - create 171, 218
 - how it works 218
 - indirection layer 138
 - mapping 131, 138, 140, 217
 - prepare 140
 - source 132, 138
 - Start 191
 - target 132, 138, 140, 168
 - trigger 191
- FlashCopy mapping 140, 229–230, 232
- flexibility 13
- flow 23, 32, 37, 52, 67, 113
- focal point 69, 81
- focal point node 69
- format 70, 98–99, 167
- frame 34, 36, 38, 40, 164, 216
 - Fibre Channel 40
- frames 29, 37–38, 40, 69
- FTP 133
- full bandwidth 38
- full-duplex 196
- function 24, 51–52, 58, 121, 124, 160, 163, 166, 213–215
- functions 4, 16, 24–25, 71, 151, 157–158, 160, 214, 217

G

- gateway 34
- gateways 34
- GB 48, 80, 213
- Gb 38
- geographically 193, 200
- geographically dispersed 193, 199
- geographically distributed 45
- Gigabit Ethernet 31, 41
- Global 11–12, 31, 51–52, 61, 130, 132–133, 168–169, 194, 233, 241–242, 244
- Global Mirror 12, 16, 51–52, 61, 77, 81, 120–122,

- 132–134, 161, 166, 168–169, 208, 233–234, 241–242, 244–245
- Global Mirror relationship 144–145
- Global Mirroring 208
- GM 25, 43, 51–52, 54, 133–134, 145, 207–208, 233–235
- gmlinktolerance 144
- grain 54, 56, 68, 131, 138, 140
- grains 54, 68, 72, 138
- grant 56
- graph 120, 143
- graphing 121
- group 9, 24–25, 53–54, 56, 126, 133–134, 162, 167, 171, 234
- groups 48, 53–54, 56, 120, 134, 149, 185, 187, 189, 218, 234–235
- growth 2, 15, 28, 120, 132
- GUI 94, 97, 101–102, 105–107, 135, 165–166, 170–171, 181, 189, 191–192, 218, 224
- GUI interface 104

H

- hackers 172
- HACMP 12, 27, 160, 193–194
 - Fallover and Fallback 197
 - resource groups that include PPRC replicated resources 195
 - sites 196
- HACMP/XD
 - solution description 193
- Hardware 6, 41, 44, 169, 200, 204, 218, 225–226
- hardware 3–4, 7–8, 10, 18, 20, 27–28, 41, 44, 69, 83, 169, 194–195, 200–201, 205, 212–213, 218–219
- HBA 38, 63–64, 66
- HBA port 63–64, 66
- HBAs 23, 63, 67, 78, 81, 163, 200
- header 40
- health 200, 203
- heartbeat 69, 179, 198, 201
- help 4, 13–14, 16, 26, 120–121, 131, 135, 156, 161–164, 192, 216
- heterogeneous 45, 121, 162
- High Availability 14, 16, 19, 120, 160, 193
- high availability 5, 13, 16, 26, 31, 199
- highly available 4, 196
- hold time value 36
- hop 37, 40, 83

- hop count 36, 82–83
- hops 36–37, 80, 82–83
- host 23, 25, 28, 53–55, 123, 136, 138, 162–164, 217–218, 220
 - information 111, 163
 - systems 78, 217
- HTTP 172, 177
- hubs 28

I

- I/O group 72, 76–77, 126, 149
- I/O groups 149–150
- I/O performance 137, 144
- IBM TotalStorage Productivity Center 123, 149, 161, 163, 239
- IBM VSS Provider 218
- identification 128, 145
- identify 13, 26, 101, 120, 137, 151, 161–162
- idling 114
- IdlingDisconnected 61
- IETF 33–34
- iFCP 34, 46
- image 2, 8, 51–52, 60
- implement 19, 24, 71, 74, 84, 157–159, 212
- implementing 17–18, 41, 54, 71, 84, 164
- improvements 20
- in-band 28, 204
- inconsistent 55–56, 114
- InconsistentCopying 60
- InconsistentDisconnected 61
- InconsistentStopped 60
- independent fabrics 80
- information 7, 10, 21, 28, 37, 55, 61–62, 121–123, 156, 161–162, 214, 217–218
- infrastructure 2–4, 11, 13, 18–19, 26, 28, 31, 120–121, 131, 136, 147, 161, 163, 204–206, 212
- initiate 121, 141, 217, 228, 233
- initiating 232
- initiator 42–43, 218
- initiators 45
- input/output 42
- install 161, 171, 209, 218, 220, 222, 225
- installation 48, 84, 122, 161, 171–172, 211, 213, 219
- Installing IBM VSS Provider 220
- Integration 15
- integration 12, 15, 45, 204
- integrity 6–8, 44, 46, 56, 157, 201, 214, 217

- intelligence 131
- interactions 135
- intercluster 24–25, 27, 62, 65, 67
- intercluster link 67, 69, 71
- intercluster link bandwidth 67
- interconnection 27, 33–34, 80, 136
- interface 15, 24–25, 40, 72, 104, 121, 135, 163, 203, 206, 217–219
- Internet 29, 200
- inter-switch link 44
- interval 22, 39, 47, 122–123, 126
- intracluster 25, 62, 107
- introduction 26, 119
- investment 32, 45
- IP 29, 33–34, 133, 146, 167, 172, 181–182, 194, 220, 236
- IP address 182–183, 186–187, 189, 220, 222
- IP packets 44
- iSCSI 46
- ISL 44, 82
- islands 33
- ISLs 83
- isolation 45–46

J

- jumbo frame 40

K

- key 34, 55, 75, 136, 167, 206, 219

L

- LAN 161, 200, 212
- latency 22–26, 35, 37, 44, 48, 52, 57, 67, 78–80, 83, 107, 117, 133–134, 138, 142–144, 146, 151, 191, 201, 236
- layers 70, 72
- LBA 68
- level 7–8, 19–21, 78, 120, 126, 131, 159–160, 162, 233, 240–242
 - storage 137, 164
- levels 2, 7–8, 127, 148, 150, 156, 160, 212
- liberates 46
- library 209, 213, 216
- license 81, 201
- licensed 81
- Licensing 81
- light 26, 31–32, 35

- light propagation 35
- limitation 48, 80
- limitations 31, 46, 57, 79
- limits 45, 83
- linear 36
- link 16, 23–24, 26, 57, 61–62, 132–133, 144, 181, 201–202
 - latency 26, 81, 83, 133
 - speed 26
- link extenders 38
- links 23, 26, 33, 62, 65–66, 145, 194, 200, 203
- Linux 171, 209
- list 79, 88–89, 122, 126, 144, 186, 190, 192, 221, 225–226, 239
- load balance 23, 68
- loading 135
- local cluster 69
- local fabric 82
- location 5, 32, 35, 86, 160, 168, 175, 195, 204, 212–213, 224, 235
- log 7, 48, 85–86, 179, 234, 236
- logged 7, 65, 178
- login 37–38, 65, 177, 179, 182
- logins 62, 69, 72
- logs 8, 69
- LUN 15, 27, 53, 70, 84, 86–88, 160
- LUN mapping 87
- LUNs 35, 53, 86–88, 166
- LVM 207–208, 234–236

M

- mainframe 48
- maintenance 4–5, 13, 18–20, 28, 69, 77, 120, 138, 195, 201, 209
- MAN 28–29
- manage 26, 35, 91, 162, 165, 179, 219
- managed disk 53, 92, 96
- managed disk group 53
- management 2, 7, 14, 18, 28, 46, 91, 107, 121, 161–163, 165–166, 179, 196–197
 - applications 163
- managing 27–28, 121, 162, 179, 186
- map 76–77, 208, 220, 224–225
- mapping 71, 84, 87, 131–132, 140, 229–230, 232
- mappings 120, 141, 189
- maps 13, 77
- mask 4, 16, 20
- master 54, 60, 101, 109, 140, 218, 220

- maximum distance 38
- MB 35, 49, 67, 149–150
- MDisk 53, 84, 89–90, 120, 123, 126, 242
 - showing 123
- MDisk group 84, 91–92, 100, 138, 149
- MDS 9000 194
- media 29, 40, 213
- member 174
- members 76
- memory 37, 73, 139, 141, 175
- merging 45
- message 40–41, 69, 73, 164, 171
- messages 27, 41, 77, 86
- messaging 199
- metric 48, 128–129, 131, 240–242
- Metro 12, 16, 26, 51–52, 61, 120–121, 130, 161, 163, 165, 233–234
- Metro Mirror 12, 51–52, 82, 132, 137, 144, 165–166, 168–170, 233
- Metro Mirror Read-from-secondary 204
- Metro Mirror relationship 138, 168–169, 183
- Metropolitan Area Network 28
- microcode 45, 138, 207, 213, 216
- Microsoft Cluster 160, 199–200
- Microsoft Volume Shadow Copy Service 218
- migration 15, 138
- mirrored 5, 7, 25, 31, 49, 115, 159, 166, 188, 212–213, 234
- mirrored copy 197
- mirroring 11, 27, 52, 54, 62, 78, 114, 157, 159, 194, 197, 235–236
- mkpartnership 86
- mkrcrelationship 106–108
- mode 16, 25, 32, 45, 88, 96–97, 140, 150, 189, 191, 215
- module 63–64, 209
- monitor 114, 120–123, 133, 161–163, 165
- monitor performance 136
- monitored 69, 117
- monitoring 3, 27, 78, 119–121, 136, 150, 166
- monitors 69, 122, 200, 204
- mount 217
- MSCS 160, 199–201, 218
- MTU 40–41
- multiplexing 31–33
- multi-vendor 45

N

- N_Ports 37
- names 31, 87, 91, 170
- naming 93, 170
- nanometers 32
- Navigation Tree 122–124
- network link 41
- new MDisk 91, 93
- node 38, 63–64, 66, 120, 126, 144, 147, 194–196, 240, 243
 - failure 194
 - port 63
- node level 126, 146–147
- nodes 27, 38, 62, 64–65, 134, 147, 151, 194–198
- non 4–5, 13, 27, 38, 58, 69, 73, 133, 151, 162, 194
- non-redundant 69
- NTBackup 228
- NTP 47

O

- offline 77, 111–112, 114, 159, 196, 198
- online 11, 15, 20–22, 75–76, 100, 115–116, 121, 158–159, 207, 218–219
- operating systems 15, 18, 78, 161, 165, 214, 217
- optimize 13, 162
- ordered list 96
- ordering 55, 75, 195, 201
- OS 12, 78, 166, 218
- overlap 75
- overloading 133, 138
- overview 2, 59, 121, 134, 178–179, 193, 226
- OXID 42

P

- packet 27, 39, 133
 - segments 40
- parallel SCSI 28
- Parallel Sysplex 160
- parameters 67–68, 175, 226
- partitions 161
- partnership 24, 29, 62, 85–86, 145
- password 171–174, 177, 181–182, 220–221, 223–224
- path 15, 28, 36, 39–41, 70–71, 142, 146, 201, 224
- paths 5, 23, 82, 179, 200
- payload 40
- peak 47–48, 127–128
- peak workloads 131

- Peer-to-Peer Remote Copy 160
- performance 15, 18, 23–24, 26, 29, 35–39, 57, 63, 68–70, 119–122, 162–164, 239–243
 - degradation 123, 137
- performance characteristics 153
- performance data 47, 121–122, 162
- performance improvement 41–43
- Performance Monitor 136
- performance monitoring 120–121, 136
- permanent 4, 20
- physical 28, 53, 63–64, 204, 212, 214
- ping 39, 133
- pipe 31, 37–38
- PiT 157–159, 214–216, 218
- PiT copy 159–160, 207, 214, 216–217
- planning 6, 8, 14, 26–27, 34, 83, 117, 120–121, 130
- point-in-time 11, 15, 24–25, 138, 140, 219
- point-in-time copy 15
- point-to-point 29
- policies 96, 196–197, 209
- policy 56, 96, 162, 196–197
- port 37–38, 63–64, 66, 120, 126, 150, 162, 167, 172, 177, 242–243
- port numbers 172
- ports 37, 63, 65, 82, 126, 146, 150, 171–172
- power 2, 5, 36, 55
 - failure 5, 55
 - supplies 5
- power supply 5
- PPRC 12, 160
- preferred 40
- primary 23–24, 26–27, 51–54, 131–132, 137, 141, 160, 186–188, 212, 233, 235–236
- priority 61, 197
- private 29, 201–202
- problems 23, 26, 45–46, 55, 161, 206
- profile 47
- progress 55–56, 69, 165, 191–192
- propagation 35, 44
- properties 6–7, 14, 35, 127, 192
- protect 6, 91, 144, 194, 199, 211–212
- protecting 193
- protection 6, 24, 28, 32, 45, 83, 159, 216
- protocol 34–35, 37, 43, 46, 80, 235
- protocol conversion 45
- protocols 29, 46, 72, 74
- provisioning 46, 162
- PVID 195

Q

Quality of Service 133
quickly 4, 15, 20, 56, 70, 107, 140, 206
quiesce 202, 218
quorum disk 202

R

R_RDY 38
RAID 53, 71, 133, 138
ranges 54
reboots 15, 207
receive 38, 40, 43, 74, 83, 142, 148, 150, 233
recovery 2–3, 5–6, 18, 20, 22, 24–25, 34, 55, 76,
101, 156–157, 165, 194–196, 212–214
recovery point 7, 55
Redbooks Web site 247
 Contact us xvi
redundancy 23, 132, 138
redundant 4–5, 16, 18, 20, 23, 69, 145
redundant fabrics 85
redundant power supplies 5
relationship 3, 21, 27, 35, 53–56, 123–124,
131–132, 157, 168–169, 183, 187, 243
remote cluster 66, 69, 72
remote copy 22, 24, 35, 55, 57, 179, 208
remote disk 27
remote fabric 82, 85
remote mirroring 236
remotely 25
remount 216
removed 41, 73, 232
rename 89–91
replicate 47, 59, 195
replication 7, 12, 22–23, 25, 28, 46, 48, 51–52,
57–58, 61, 113, 166, 168, 233
reporting 119, 121
reports 47, 121–123, 161–162
reset 69
resiliency 13
resilient 29
resolve 206
resource allocation time-out value 37
resources 18–19, 21, 45, 67–68, 141, 196, 199,
202–204, 212, 235
restart 7, 141
restarting 176
restarts 202
restore 11–12, 159–160, 207, 212, 214–215

restore procedure 217
restricted 2
retransmission 39
ring 29–31
risk 9, 202, 213, 216
role 1, 27–28, 60–61, 168, 170
roles 168–169
root 137
round 35, 38–39, 68–69, 72, 142–143, 146
round trip 43
route 31
routers 35, 41, 46
Routing 45–46
routing 46
RPQ 194, 200
RSCNs 45
RTO 6–8, 22–23, 39, 157–158, 212, 233
rules 48, 56, 85

S

SAN 1, 17, 23–24, 28–29, 51, 53–54, 63–64, 71,
79, 84, 122, 157, 161–163, 186, 194, 207, 211–212
 availability 23
 fabric 44
 islands 45
SAN extension 24, 27, 29
SAN routers 46
SAN Volume Controller 194, 199, 207, 209
scalability 32
scalable 194
scale 36, 46
scattering 35
scripting 117, 209
scripts 24–25, 90, 159, 164, 192, 208–209, 233
SCSI 23, 28, 35, 42–43, 53–54, 70–73, 134, 204,
235
 commands 71, 235
 protocol 42
SDD 200, 203–204
SDH 29–30
secondary 7, 16, 23–26, 52–56, 132–134, 137,
160, 179, 186–188, 195, 213, 233–235
secondary copy 235
secondary site 29, 52, 62, 67, 197, 200, 202,
233–234, 236
secure 11, 156, 172, 177, 212–214
security 2, 46, 207
segment 3, 39–40, 138

- sequence 37, 40, 43, 52, 55–56, 169
- sequential 35, 68, 96
- Server 12, 53, 134, 171, 180, 218–219, 226
- server 7–8, 15, 23, 28, 57–58, 77, 122, 128, 134, 156–158, 213–215
- Servers 3, 179, 206
- servers 2, 15, 28, 47, 53, 78–79, 113, 131, 134, 138, 146, 158–159, 161, 163, 166, 179, 212, 227, 234
- service 9, 13, 18–20, 52, 120, 136, 149, 165, 182, 195, 218, 226
- Service Location Protocol 227
- settings 82, 86, 101, 204
- setup 165, 201, 220
- share 74
- shared 31, 44, 133, 145, 161, 196
- sharing 199
- shrinking 15
- shut down 159
- shutdown 164
- signal power 36
- Simple Network Management Protocol 162
- single point of failure 5, 166
- site 6–7, 10–11, 24–26, 29, 33, 52, 58, 62, 65, 137–138, 153, 156–157, 159–160, 168, 212–213, 233
- sizing 47, 121, 130–132, 134
- SMI-S 121
- snapshot 206, 218–219
- snapshots 206
- SNMP 147, 162
- SNS 45
- Software 12, 41, 44, 200
- software 3–4, 18, 20, 23, 55, 69–70, 121, 159, 161, 192, 214
- solution 4–6, 17–19, 55, 62, 71, 120–121, 127, 155–157, 212–214
- solutions 1, 3, 6, 27, 29, 32, 51, 119–120, 157–158, 213–214, 233
- SONET 29–31
- sort 89–90
- source 15, 24–25, 40, 52, 54, 57–58, 72, 78, 131, 133, 138–140, 189, 191, 196–197, 203, 215–217, 220
- sources 131, 135
- space 13, 15, 38, 48, 54, 76, 80, 122, 136, 138, 160, 162, 164, 215
- spare 21
- speed 18, 29, 31, 34–35, 44, 46, 213
- speed of light 35
- speeds 34–35
- split 32, 72, 138, 140, 142
- spoofing 42
- SSL 224
- stack 71, 134
- standards 33–34
- start 26, 38–39, 53, 84, 111, 132, 140, 175, 177–178, 229, 235
- state 7, 27, 59–61, 131, 141, 180, 185, 189, 216, 235
 - consistent 7, 60
 - ConsistentSynchronized 60
 - inconsistent 114
- state changes 44, 189
- states 59, 61, 167, 196, 200
- statistics 121–122, 126, 131, 134, 136, 142, 144–146, 243
- status 84, 86, 115–116, 147, 165, 178–180, 183–184, 235
- stoprcconsistgrp 114
- storage 3–5, 23–24, 26, 53, 55, 61, 120–122, 156, 160–161, 212, 218
- Storage Area Networks 28
- storage capacity 139, 163
- storage controller 62, 139
- storage controllers 28, 70
- storage level 208
- Storage Manager 11, 15, 159, 165, 218–219
- storage network 131, 136
- striped 91, 96
- Subsystem Device Driver 200
- superuser 220
- support 13–16, 36, 40, 45–46, 78–81, 83, 120, 133, 137, 147, 159–160, 166, 194, 199, 201, 206, 215, 217–219, 236
- surviving node 194
- suspended state 198
- SVC 2, 14–17, 23–24, 51–53, 119–121, 155, 161, 163–164, 215–218, 239–242
- SVC cluster 27, 57, 62–63, 65, 69, 144–146
- SVC configuration 85, 209
- SVC intracluster 62
- SVC node 63, 65, 71, 120, 147, 150
- SVC nodes 67, 81–82
- SVC support for VSS 218
- svcinfo 85–86, 114, 170
- svctask 86, 88, 91, 110, 114, 191, 224–225
- switch 33, 36, 38, 52, 63, 65, 162–163, 194

- fabric 36
- switch port 38
- switch ports 85
- switches 28, 35, 37, 40–41, 45, 52, 63–64, 79, 83, 162, 194
- Switching copy direction 111
- switchrconsistgrp 114
- symmetric 85
- symmetrical 82
- synchronization 47, 54, 56, 76, 132, 140–141, 197, 233, 235
- Synchronized 76, 109–110
- synchronized 27, 76–77, 110, 180, 189, 236
- synchronizing 138, 180, 233
- Synchronous 22, 25, 30, 52, 57, 107, 194
- synchronous replication 22, 48
- system 2, 4–5, 18–20, 55, 68–69, 121, 123, 156, 158–159, 217, 234–235
- system design 22
- system performance 48, 151

T

- T0 15
- target 8–9, 15, 24–25, 42–43, 52, 54, 57–58, 70, 72, 75, 78, 111, 120, 131–132, 137–141, 159–160, 168, 187, 198, 215–217
- targets 45, 91, 140, 203, 216
- tasks 131, 163, 165, 178, 184, 192
- TCP/IP 33
- TCP/IP network 34
- telephone network 36
- test 16, 42, 69, 128, 133, 159, 233
- This 1, 3, 6, 17, 20–21, 51–53, 119–121, 155, 159–160, 212–213, 217, 239, 241–242
- Threshold 149, 151
- threshold 69, 121, 132, 148–150
- thresholds 122, 148–150, 162
- throughput 38–40, 68, 83, 120
- tier 9–11, 17, 25, 46
- time 4–8, 10–11, 20–25, 51–52, 55, 61, 67–69, 71, 120–121, 127, 131–134, 137–139, 156–158, 163–164, 179, 181, 183, 191, 195–196, 212–218, 243
- time outs 37
- Tivoli 11, 15, 162, 218–219
- Tivoli Storage Manager 11, 159, 218–219
 - server 218–219
- Token Ring 30

- tools 28, 47, 117, 136, 163, 179, 204
- Topology 123
- topology 28–29, 31, 123–124, 162
- TotalStorage Continuous Availability for AIX
 - solution description 193
- TotalStorage Continuous Availability for Windows
 - overview 199
 - solution components 200
 - solution description 199
- traditional 21, 28, 46
- traffic 28, 31, 33–34, 45, 67–68, 83
 - Fibre Channel 45
- transfer 24, 38, 42, 201, 233
- transitions 60
- transport 29, 31, 145
- trends 120, 161
- trigger 121, 148, 160, 164, 191
- trunking 83
- truststore file 224
- tunnel 41
- tunneling 46

U

- unique identifier 53
- UNIX 165
- unmount 217
- upgrade 4, 44
- upgrades 4, 15
- user interface 218–219
- users 4–6, 11, 16, 19–20, 22–23, 31, 156, 158, 174, 221

V

- vaulting 11, 157
- VDisk 15, 25, 52, 54–55, 60, 120, 123, 125, 168, 186–187, 216, 218, 220, 230–231, 234, 241
 - creating 54, 141
 - information 123, 168, 230
- verification 226, 229
- virtual disk 24–25, 54
- Virtualization 70–71
- virtualization 14, 28, 96
- virtualized storage 14, 209
- volume group 195–196
- Volume Shadow Copy Service 218
- VSS 218
- VSS activation 228
- VTs 12

W

WAN 28–29, 42, 133, 146, 212
warning 148, 180, 189
wavelength 32
wavelengths 32
WDM 32
Windows 2003 171, 200, 218, 226
wire speeds 34
Wizard 94–95
workload 21, 26, 47, 79, 120–121, 127
workloads 75, 131, 133–134
write acceleration 35, 43
Write ordering 75
write ordering 55, 75
writes 25, 35, 38, 47–48, 55–56, 60, 67, 69, 132,
139, 142–144, 204, 214, 217, 234–235

Z

zone 45–46, 82, 85–86, 163
zoned 44
zones 21, 65–66, 85
zoning 44, 84, 163



Using the SVC for Business Continuity

(0.5" spine)
0.475" <-> 0.875"
250 <-> 459 pages



Using the SVC for Business Continuity



Redbooks

Learn how to size and design a scalable Business Continuity solution

This IBM Redbooks publication gives a broad understanding of how you can use the IBM® System Storage SAN Volume Controller (SVC) as the foundation for IT Business Continuity in your enterprise.

Learn when to use Global Mirror, Metro Mirror, and Flash Copy

This book will help you select the appropriate techniques and technology including how to size and design your IT Business Continuity Solution, while utilizing the SVC Advanced Copy Services and obtaining a highly flexible and scalable storage infrastructure.

Business Continuity solutions with the SVC

**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