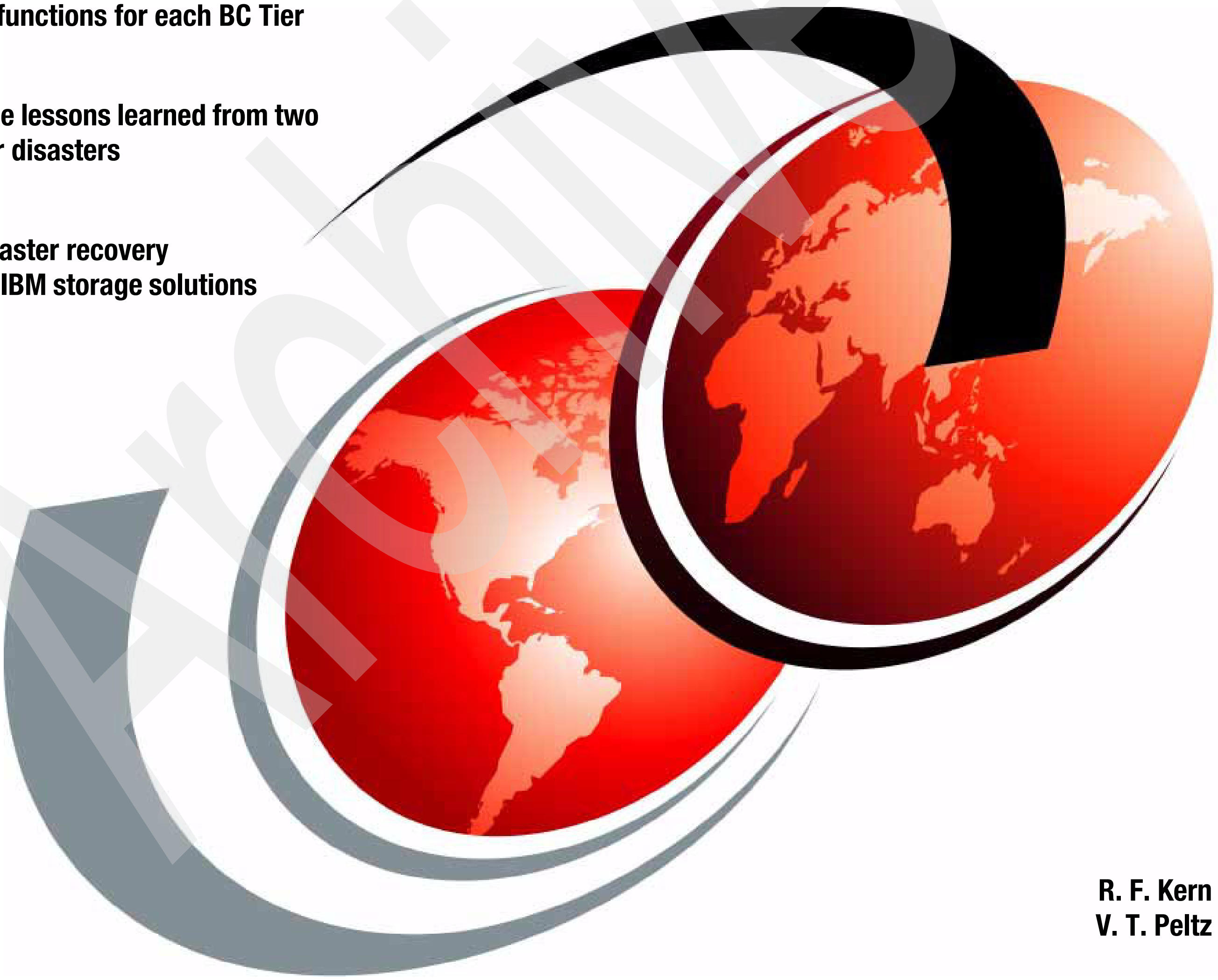
IBM Storage Infrastructure for Business Continuity

Describes IBM storage replication and automation functions for each BC Tier

Describes the lessons learned from two recent major disasters

Provides disaster recovery metrics and IBM storage solutions



Archived

R. F. Kern
V. T. Peltz

Redpaper

ibm.com/redbooks

Archived

International Technical Support Organization

IBM Storage Infrastructure for Business Continuity

January 2010

Archived

REDP-4605-00

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

Archived

First Edition (January 2010)

This document created or updated on January 29, 2010.

© Copyright International Business Machines Corporation 2010. 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

Archived

© Copyright IBM Corp. 2010. All rights reserved.

Archived

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.

Archived

© Copyright IBM Corp. 2010. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

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

AIX®
DB2®
DS4000®
DS6000™
DS8000®
ECKD™
Enterprise Storage Server®
ESCON®
FICON®
FlashCopy®
GDPS®
Geographically Dispersed Parallel Sysplex™
HACMP™
HyperSwap®
IBM®
IMS™
Parallel Sysplex®
PowerHA™
Redbooks®
Redpaper™
Redbooks (logo) ®
System p®
System Storage™
System z®
Tivoli®
TotalStorage®
XIV®
z/OS®
z/VM®
z9®

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.

Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

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

Archived

Preface

The importance of business continuity and disaster recovery remains at the forefront of thought for many executives and IT technical professionals. This IBM® Redpaper™ describes the lessons learned from recent disasters and how IBM storage technology can help businesses address many of the issues related to protecting their storage infrastructures and business-critical IT applications.

Two principal disaster recovery metrics, Recovery Time Objective and Recovery Point Objective, are defined and, along with the associated cost tradeoffs, are discussed from the vantage point of various IBM storage technology solutions.

Two IBM Business Continuance/Disaster Recovery (BC/DR) automation solutions, known as GDPS/PPRC with HyperSwap® and GDPS/PPRC HyperSwap Manager, are described and shown how they can help an installation move closer to attaining a goal of continuous operation[1]. For z/OS installations operating two or more sites, in the event of a storage subsystem, host, network or communications facility failure, a switch to processing at an alternate site can be made in almost real time by using GDPS/PPRC with HyperSwap.

Additionally, many Clustered Open Systems that are integrated with IBM Remote Copy technology can be configured to switch to a second site in almost real time. In these situations, when a site switch is executed, applications that have been cloned at both sites can continue running with minimal impact to the user.

The team who wrote this paper

This paper was produced by a team of specialists at the International Technical Support Organization, San Jose Center.

R. F. Kern is a member of the IBM Advanced Technical Support, Americas, (e-mail: bobkern@us.ibm.com). Mr. Kern is an IBM Master Inventor & Executive IT Architect. He has 36 years experience in large system design and development and holds numerous patents in storage-related topics. For the last 28 years, Bob has specialized in disk device support and is a recognized expert in continuous availability, disaster recovery and real time disk mirroring. He created the DFSMS/MVS subcomponents for Asynchronous Operations Manager and the System Data Mover.

In 2004, Bob was named a Master Inventor by the IBM Systems & Technology Group and is one of the inventors of Concurrent Copy, PPRC, XRC, GDPS® and zCDP solutions. He continues to focus in the Disk Storage Architecture area on HW/SW solutions focused on Continuous Availability, and Data Replication. He is a member of the GDPS core architecture team and the GDPS Customer Design Council with focus on storage-related topics.

V. T. Peltz works for the IBM Systems and Technology Group, San Jose, California, (e-mail: vpeltz@us.ibm.com). Mr. Peltz is a Consulting IT Architect in the IBM Systems and Technology Group. Vic worked for many years as a Systems Engineer with IBM's Finance and Insurance customers, and then in IBM storage product marketing organizations. He is currently working with some of the major IT industry consultants on a variety of storage-related topics, and with the IBM Tucson storage performance organization to analyze the performance and resiliency of IBM's asynchronous remote mirroring function.

[1] GDPS/PPRC with HyperSwap operates in z/OS® environments

Archived

© Copyright IBM Corp. 2010. All rights reserved.

Vic has been directly involved in working with companies to help them recover from computing center disasters. Over the years Vic has worked with numerous companies and governments to help them develop IT disaster recovery plans. He holds an M.A.Sc. in Electrical Engineering from the University of Toronto, joined IBM Canada Ltd. in 1975 and moved to the IBM San Jose Development Laboratory in 1992.

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

Mary Lovelace
International Technical Support Organization, San Jose Center

The authors would like to thank Craig Dahl and Jim Fisher, both with IBM Systems and Technology Group Tape Sales for their comments concerning tape solutions for business continuity and for supplying the graphics for Figure 3-2 on page 19.

Become a published author

Join us for a two- to six-week residency program! Help write a book 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.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

  ibm.com/redbooks

- Send your comments in an e-mail 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

Archived

Archived

1

Introduction

This chapter describes the content of this paper.

© Copyright IBM Corp. 2010. All rights reserved.

1.1 Paper content

Many people believe the events in New York City on the morning of September 11, 2001 were a watershed. For professionals involved in information technology (IT), their experiences resulting from the events of 9/11 changed their expectations and act as the benchmark for assessing what is required in a thorough and tested disaster recovery plan. Likewise, the Internet has changed customer expectations as to what constitutes acceptable availability for e-Commerce applications. Both of these new realities set standards for what is required to enable IT installations to recover from a disaster.

This Redpaper is divided into two parts:

- Part I summarizes the lessons learned by customers who had to recover their data centers following the WTC disaster. The points discussed in Part I result from having visited the site of the disaster and subsequent discussions with several customers who were affected by the disaster. See Figure 1-1.
- Part II describes seven tiers of disaster recovery defined by SHARE, and outlines some IBM products that can contribute to implementing disaster recovery plans appropriate to each disaster recovery tier.

The paper concludes with some remarks on vendor independence.

Figure 1-1 WTC Ground Zero on 19 June 2009

Archived

1.2 Terminology

Throughout this paper, frequent reference will be made to functions that are available on the IBM DS8000®, DS6000™, ESS800, DS5000, DS4000®, XIV® storage subsystems and the SAN Volume Controller (SVC). The authors believe that continuous and frequent reference to:

"IBM DS8000, DS6000, ESS800, DS5000, DS4000, XIV, SVC" would be cumbersome and serve to distract the reader. Accordingly, the term "eStorage" will be used instead, and in the context of this paper, is intended to mean reference to the IBM DS8000, DS6000, ESS800, DS5000, DS4000, XIV storage subsystems and the SAN Volume Controller (SVC). The term "eStorage" is not an official IBM term and is only used in the context of this paper. When the authors wish to limit the discussion to the DS8000, DS6000 and ESS800, the term DS8000 will be used.

Likewise, when discussing the IBM FlashCopy® local data replication function, we will make reference to z/OS volumes and Open Systems LUNs as the source or target of a FlashCopy operation. Rather than continually refer to a "Source volume/LUN" or a "Target volume/LUN", we will simply use "source" or "target" as appropriate to mean both z/OS volumes and Open System LUNs. In cases where a function refers only to a z/OS or Open Systems host environment, we will state that explicitly.

Archived

Archived

2

Part I: The lessons of 9/11

This chapter summarizes the lessons learned by customers who had to recover their data centers following the WTC disaster.

Archived

© Copyright IBM Corp. 2010. All rights reserved.

2.1 Business problems resulting from a disaster

The following sections describe business-related problems and issues resulting from a disaster.

2.1.1 Regional verses local disasters

Business can be affected in many ways following a disaster. The WTC attacks focused attention on vulnerabilities that must be addressed. However, these issues also existed prior to the 9/11 attacks. What was generally not understood prior to the events of 9/11 was that a regional disaster, such as what happened in Manhattan, can impose significant additional unanticipated difficulties and requirements that make plans for recovering from local disasters (such as a fire) inadequate.

The need to differentiate requirements for recovering from a regional disaster versus a local disaster has been further impressed on people's minds after the recent blackout in the Northeast U.S., SARS, the 2004 tsunami in Southeast Asia, and 2005 Hurricane Katrina. Other possible future events could be, for example: a major earthquake off the California coast and quarantine/travel restrictions resulting from an outbreak of H1N1 influenza. All of these catastrophic events interrupt normal business activities and change our thoughts on services that we often just assumed would be available after a disaster.

Communications, security, water, food, sanitary services, and so forth can be affected for hundreds of miles. Even if nothing is physically affected, for example in the case of a SARS epidemic, the IT Operations team cannot (or might not) be together in groups. In these situations remote operations from a number of distributed locations is required.

2.1.2 Business issues beyond the data center

Some examples of the business-related problems and issues that organizations must be prepared to deal with in the event of a disaster are as follows:

- The inability to restart or execute restoration of business operations. This can result in:
  - the business becoming unavailable to its customers because they cannot access the company's products and services
  - suppliers being unable to reach the business to resupply necessary materials
- As a consequence of a business being unable to restart operations, clients might look elsewhere for services, resulting in the loss of customer loyalty and lost revenue. Clearly, part of the requirement for successful restoration of business operations is the ability to deliver acceptable IT service. While it is reasonable to assume that most customers and suppliers will not expect a business to operate at its peak level following a disaster, degradation of service below a certain threshold will force customers to take their business elsewhere.
- Security vulnerability increases in times of business disruptions. For example:
  - Corporate Web sites are more vulnerable to attack, defacement, and having their contents changed maliciously. Cyber attackers might exploit disaster situations to infiltrate companies through their Internet environment.
  - Vulnerability to virus proliferation might increase, with the associated potential for the corruption or destruction of data.

Archived

- The erosion of business confidence based upon company reactions as reported in the press and internal communications to employees, for example, the perception of the company's ability to react to the crisis, minimize chaos, and provide timely updates on the status of company operations and company as a whole, and how Human Resource (HR) issues are being addressed.

The 9/11 attacks have caused many business located near U.S. landmarks or in densely populated areas to view their location as presenting increased risk. This increased risk must be incorporated into an organization's business continuity planning.

2.1.3 Problems with existing business continuity plans

Prior to 9/11, IT installations with business continuity plans felt, in most cases, that they had developed a sound, workable plan. In reality, when it came time to execute the plan, many companies found their plans were inadequate and/or incomplete. The following are examples of what companies found to be lacking or inadequate in their recovery planning.

- The business continuity plan was not comprehensive and only addressed recovering the IT installation. The plan did not address all of the other actions required to accomplish an end-to-end business recovery.
- Documentation describing how to recover critical business processes and IT applications was missing or inadequate.
- Personnel did not receive sufficient education and training about the business continuity plan and did not conduct disaster drills to practice executing the plan.
- The time assumed in the business continuity plan to be available to recover a critical IT application did not match the time imposed by the associated business requirement for how fast the IT application must be restored to operation.
- The business continuity plan was inadequate or lacked instructions for addressing organizational issues related to critical business functions. Examples of organizational items which companies found inadequate or missing include:
  - Defining the priority sequence for recovering business processes and associated IT applications,
  - Documenting a clear chain-of-command and an associated decision control matrix,
  - Documenting a succession plan in the event key personnel are incapacitated -- companies were unprepared for the loss or unavailability of key personnel and the skills they possess, and
  - Having an effective communication plan: internally for employees and externally for suppliers, customers and the media.
- Many business continuity plans did not include plans for the replacement of IT equipment outside of the data center. All key elements of the IT infrastructure were found to be essential to sustaining business, not just equipment in the data center.

Archived

2.2 Lessons learned and some key business considerations

There are hardware and software disaster recovery actions that require consideration and planning. The following sections discuss recovery site considerations based on lessons learned from the WTC attacks.

2.2.1 Ten most important lessons learned from September 11th

The WTC attacks precipitated a "regional" disaster that affected many city blocks, hundreds of companies and thousands of people's lives. Power was cut off, major telephone company switching centers were destroyed, travel restrictions were imposed with resulting delays and access to buildings and data centers was restricted or impossible for several days following the disaster. All of these factors contributed to how individuals and corporations functioned after the morning of September 11, 2001. Out of this experience came some hard-won lessons. The following list details the ten most important lessons learned:

- Successful recovery necessitates a greater dependency upon automation rather than people.

  In a regional disaster, it is risky to assume that critical skilled personnel will be available to recover the business. As lives are affected in a multitude of ways by a disaster, they might not be able to get to their workplace or might be forced to deal with personal situations resulting from the disaster, which must take priority.

  Consequently, many IT installations are now looking at ways to automate switching all resources from one site to another from a single trigger point. The ability to switch all critical resources from one site to another is discussed in Part II under the topic: 3.4, “IBM geographically dispersed parallel SYSPLEX” on page 40.

- The recovery site must have adequate hardware.

  IT installations must plan for the recovery site to have sufficient server processing capability and storage capacity to enable them to run all business-critical applications at the recovery site.

  IBM System z® servers have an option called the Capacity Backup Option (CBU). CBU makes it easier to match host processing capacity at the recovery site to the processing capacity at the primary site. CBU provides extra CPU engines that can be activated dynamically when needed. Using CBU, a customer can have the additional processing power (MIPS) available on standby and only start paying for them when they are actually required.

  When features like CBU are not available, as with many Open System servers, IT installations do not necessarily purchase two duplicate servers: one for production at the primary site and a second for the disaster recovery site. Disaster recovery plans must take into account that the processing capability at the recovery site might be less than the processing capacity at the primary site.

- The recovery site must have hardware facilities that are compatible with the hardware at the primary site.

  An example of a potential incompatibility is encryption hardware. If the recovery site does not have the necessary encryption hardware, it might not be possible to process data that was encrypted at the primary site. This is an example of a requirement that is easy to overlook.

Archived

- A disaster might cause multiple companies to initiate recovery actions, thereby stressing the capacity of local commercial business recovery services and other services in the local area.

  Business continuity companies typically work on a "first come, first served" basis. Hence, in the event of a regional disaster the business continuity facilities can fill up quickly. Furthermore, the contracts with some business continuity facilities might stipulate that the customer using the facility must vacate within a specified period of time, for example, within sixty days. For these reasons, a number of companies have been prompted to:

  – Explore what would be required to create their own disaster recovery site,

  – Consider changing business continuity firms, or,

  – Initiate negotiations with their business continuity firm to modify their existing contracts to make it more likely that the necessary hardware and space are available if needed.

- Having executed a successful recovery from disaster, it also is necessary to have a secondary disaster recovery plan.

  While having a secondary disaster plan might seem like overkill, companies that recovered to a secondary site having lost their primary data center realized that now all of their data was located in one location. Not knowing the extent of the terrorist actions, several customers soon realized that if their recovery site was incapacitated, they had no further recourse of action (for example, no secondary disaster plan).

  Furthermore, those companies that had recovered to a third party disaster recovery site were also faced with having to vacate the recovery facility within the time limit specified in their contractual agreements. A regional disaster of this magnitude caused raised floor space and office space in the greater New York city area to become scarce. So, if companies needed subsequently to move their data center to another facility, it became a difficult problem that needed to be addressed in a short period of time. A key question that all customers should ask relative to any disaster/recovery plan is "If I ever invoke my disaster recovery plan, what is my secondary disaster recovery plan?"

- Some companies are giving serious consideration to a three-site strategy.

  Realizing that losing one of two sites results in not having a backup site, several companies have implemented a three site disaster/recovery strategy. In this scenario, there would be two sites within the same general geographic area to facilitate quick failover and high availability plus a third out-of-region site to protect against a wide-spread regional disaster. Telecommunication costs are a major factor when considering a three-site strategy. Clearly, not all corporations can justify the cost of such a solution. However, the cost versus risk factors should at least be analyzed. In some cases some interesting solutions have started to emerge, including corporations sharing raised floor space in a common facility.

- Investigate using true asynchronous remote mirroring.

  Another consequence of the WTC disaster is some IT installations with their own backup site are considering moving the site further away from their primary data center.

  Increasing the distance between the primary IT site and a backup site introduces some technical considerations and tradeoffs concerning which remote mirroring technique to use (that is, synchronous versus true asynchronous remote mirroring). While synchronous remote mirroring is popular for reasons discussed in Chapter 3, “Part II: Storage technology for disaster recovery” on page 13, many installations that run business-critical applications on System z hosts are giving serious consideration to changing to true asynchronous remote mirroring as part of their considerations for moving their backup site.

Archived

- Tape, as well as disk, is a crucial part of the recovery capability.

  Many companies are now rethinking the importance of their data that is stored on tape. For example, numerous companies are considering use of functions like IBM's TS7700 Grid tape mirroring. When data that was stored on tape at the primary site was lost, many companies realized that the only other copy of their critical data was stored off-site but was too old to be useful. In many cases the data on off-site tapes was days or weeks old. Even data that was only a few hours old was, in many cases, of no use to the recovery effort.

- The success of recovering data for distributed systems and desktop PCs ranged from grade A to F.

  Recovery of data for distributed systems and desktop PCs was successful if the systems and PCs were actually backed-up on a scheduled basis. Two products that are popular for this purpose are the Tivoli® Storage Manager (TSM) and Veritas' Net Backup DataCenter. In many cases a great deal of data was lost due to lack of any backup, or only a backup that was performed once a week, or once a month, and so forth.[1]

  An installation's backup policy should be evaluated on a server-by-server and desktop-by-desktop basis. The important thing is to have an explicit policy that was arrived at by investigating what backup frequency is required to facilitate a successful disaster recovery plan. When an automated backup scheme is used, the required backup policies can be implemented effectively across the business.

- The IBM Consistency Group functions for preventing a rolling disaster were exercised and worked.

  Installations in the WTC that made use of Consistency Groups and the Freeze/Run facilities available with IBM Systems Storage were able to use the data at their remote site and recover successfully. These facilities helped prevent what is known as a rolling disaster. A discussion is presented in Chapter 3, "Part II: Storage technology for disaster recovery" on page 13 illustrating why these techniques are necessary for successful data recovery following a disaster.

2.2.2 Additional recovery considerations

The lessons articulated above became clear soon after 9/11. Subsequently, in discussions with a number of other customers worldwide, some additional insights can be added to the list in 2.2.1, "Ten most important lessons learned from September 11th" on page 8.

- End-to-end business recovery might also involve connecting with various vendors and business partners outside of the IT organization to complete business transactions successfully.

  For example, if government regulations require that a financial institution must be in a position to close and settle outstanding transactions within a given timeframe after a disaster, this might require the financial institution to connect to vendors or business partners at their primary or disaster recovery site either locally or out-of-region. Managing this consideration will likely require additional testing scenarios and extending the disaster recovery plan to include additional considerations relative to a company's business partners and vendors.

Archived

[1] One of the authors (VTP) makes it a habit to execute a TSM incremental backup of his laptop PC every night before leaving the office. The process typically takes less than five minutes. It has enabled him to avoid a disaster on more than one occasion, not because of a hard disk failure, but rather to recover data accidentally lost due to software bugs and human error.

- Primary corporate IT data centers that are located in high risk areas might be swapped with existing secondary sites in lower risk locations to minimize the risk of a disaster in the high risk area.

  Several corporations performed risk evaluations and consequently have upgraded their secondary site facilities and then swapped their primary and secondary production sites to minimize the risk of a possible disaster in the area in and around the original data center.

- Protection of skilled labor.

  Many corporations are now exploiting the improvements in high speed Internet connections and the various high quality conference calling capabilities to permit a percentage of their skilled workforce to work remotely, be mobile, or work from home one or two days a week. The intent is to minimize the direct impact on the entire skilled labor workforce in the event of a disaster to their data center or operations center. This idea originated after September 11th as protection from physical attacks, and became a new method of operation after the SARS virus affected various work areas in various major cities.

- Alternative communication scenarios are being investigated.

  Alternative land line and cell phone carriers are being investigated. The intent is to minimize or eliminate dependence upon a specific carrier or carriers in the event of a disaster.

- The possibility of having to recover from an attack on data by way of a virus or from the malicious intent of an employee is an increasing concern for a number of corporations.

  An area receiving increasing focus is the physical and logical security of data within the data center and its security when in transit outside the corporate data center by vendors and business partners.

  A new focus area for research and development activities is the detection, protection, isolation, and recovery from data corruption. This work is being undertaken by various storage and backup software vendors, database companies and virus detection/protection software companies. Techniques, which include the creation of point-in-time local copies and the creation of local copies of changed data only (that is, so-called space efficient copies) enable more frequent backups of data. An example of a resulting benefit in a database environment is the ability to minimize the time required to recover using the database log.

  Continuous Data Protection (CDP) functionality is being integrated into existing backup software as well as virus detection and protection software.

2.2.3 Some key questions for your business

In light of recent disaster recovery experiences, here are some questions to consider as you explore the issues and concerns related to making plans so that your business can recover and continue in the event of a disaster.

- What are your most critical business processes and applications? Do you have a documented and tested plan to recover and restart them in the event of a disruption or disaster?
  - What dependencies do your business-critical processes and applications have on your on-demand business infrastructure?
  - What is the cost of an outage of your critical business operations? How long before you are out of business?
  - What risks pose the greatest threat to your business?

Archived

- How comprehensive is your business continuity program?
  - How comfortable are you that the necessary processes and procedures are in place to support the ability of your company to respond effectively to a disaster from both the IT and business continuity perspectives?
  - What is your communication strategy to respond quickly to employees, customers, and the media? What means would you employ for communicating to employees, customers, and the media quickly?
- If a disaster rendered your facility unavailable, where would you recover your business operations, including, for example, call centers, help desks, and workgroups?
  - Where would employees work if their offices were not available? How would you communicate this information to employees?
  - If the call center were unavailable, how would your customers reach you?
- Do you have adequate resources for managing Internet security and intrusion, including ongoing monitoring and management? If your Internet environment gets hacked, how will you respond?
- Is your IT recovery strategy consistent with your business objectives? Is the time allotted by the IT organization to recover business-critical applications consistent with the maximum amount of time the line-of-business can sustain for the business-critical application to be unavailable?
- Do your business and IT operations hinge on the availability of one or a few individuals skills?

2.2.4 In summary

A corporation's backup/recovery strategy, the tools and procedures for affecting recovery, the recovery point, and the recovery time from various data corruption scenarios should all be considered as part of a comprehensive business continuity/disaster recovery plan.

Today, there are many choices among storage hardware and software capabilities that one must understand to make the appropriate business tradeoffs regarding business impact of a disaster versus the cost of recovering from a disaster.

Chapter 3, “Part II: Storage technology for disaster recovery” on page 13 examines how storage technologies available today can address various disaster recovery requirements.

Archived

3

Part II: Storage technology for disaster recovery

This chapter describes seven tiers of disaster recovery and IBM products that can contribute to disaster recovery plans appropriate for each tier.

Archived

© Copyright IBM Corp. 2010. All rights reserved.

3.1 Impact and requirements of on-demand business IT applications

The following sections describe levels of service at a recovery site.

3.1.1 Impact of on-demand businesses on system availability requirements

The rapid rise of on-demand business and the Internet has given rise to the requirement that various business units be available "24 x Forever". This, in turn, has caused a number of businesses who use IBM System z systems as well as various open clustered servers to move towards an IBM on-demand business availability solution encompassing servers, storage, networks, and an application called the Geographically Dispersed Parallel Sysplex™ (GDPS) and Geographically Dispersed Open Clusters (GDOC). GDPS/GDOC provides the ability for an installation to manage end-to-end application and data availability across multiple geographically separate sites. GDPS/GDOC is discussed in greater detail later in this chapter.

In many cases new applications are developed in response to growing on-demand business requirements. As part of a move to on-demand business, enterprises should evaluate the availability requirements of various applications and weigh the options available through today's technology.

3.1.2 Response Time Objectives (RTO) and Recovery Point Objectives (RPO)

Two metrics are commonly used to establish IT system requirements for a business recovery plan are Recovery Time Objective (RTO) and Recovery Point Objective (RPO). RTO measures how long the business can tolerate an IT application being inoperable. RTO values can range from a few minutes to several hours. Surveys across multiple industry sectors have established that RTOs typically range from 15 minutes to 12 hours, depending on the industry and application.

RPO measures the amount of time the secondary storage system lags behind the primary storage system. When the primary storage system receives a write command for a volume/LUN that is part of a remote copy pair, the primary system initiates the process of transmitting the data to the secondary storage system at the remote site. This process, including actual data transmission, takes time. Hence at any given moment in time, if asynchronous mirroring is being used, the currency of the data at the secondary storage system might lag behind the primary. This time lag is typically a few seconds to several seconds, depending on several factors including the workload at the primary and secondary storage systems and the remote data link bandwidth.

Since RPO measures the amount of time the secondary storage system lags behind the primary, the corresponding amount of data that must be recreated at the secondary in the event of an unplanned system outage is directly proportional to the RPO. Hence, RPO is an important measure of how fast an installation can recover from a system outage or disaster and therfore has a direct influence on each application's RTO and on the RTO of the business as a whole.

Archived

When planning for recovery from outages, one is faced with reconciling the need to recover quickly and completely against the cost to affect the recovery. Typical business tradeoffs include:

- The operational telecommunications cost of running remote mirroring between a primary and secondary data center versus the impact to the I/O performance of the installation's primary business applications.
- Determining an installation's realistic RTO versus the potential cost in additional hardware, software, and infrastructure to achieve the desired RTO. The time available to recover critical applications and have all critical operations up and running again must include the time to recover servers, network, workload and data. More and more applications now run across multiple server platforms of the same or different operating systems. Therefore, all platforms must be fully recovered before the application is available.
- Determining the installation's RPO. How much data must be recovered in terms of the time required to affect the data recovery, or, can some of the data simply be lost?

Factors involved in determining the RTO and RPO include:

- The cost of sustaining some data loss while maintaining cross-volume/cross-subsystem data consistency. Maintaining data consistency enables the ability to perform a database restart (typically seconds to minutes in duration).
- The cost of achieving no data loss that will either impact production on all operational errors in addition to disaster recovery failure situations, or yield a database recovery disaster (typically hours to days in duration) as cross-volume/cross-subsystem data consistency is not maintained during the failing period.
- A common restart point for an applications data on a single server or across multiple server platforms might also be a consideration. This would enable each server platform to be recovered independently, but all of the data to be recovered to a common point in time, eliminating the need to adjust recovery points between various parts of the application's end-to-end data.

3.1.3 Requirement for cross-volume data integrity and Consistency Groups

It is crucial that computers write data to disks with full integrity, even in the event of hardware failures and power failures. To accomplish this, systems designers employ a variety of different techniques, including:

- Designing the storage subsystem cache so that it is mirrored to prevent data loss in the event of a cache hardware failure,
- Backing up the power to the cache by batteries to prevent cache data loss in the event of a power failure, and
- Disk mirroring or parity-based RAID schemes for protecting against HDD failures.

Another, more subtle, requirement for preserving the integrity of data being written is making sure that dependent writes are executed in the application's intended sequence. Consider the following typical sequence of writes for a database update transaction, as shown in Figure 3-1 on page 16.

1. Execute a write to update the database log indicating that a database update is about to take place.
2. Execute a second write to update the database.
3. Execute a third write to update the database log indicating that the database update has completed successfully.

Archived

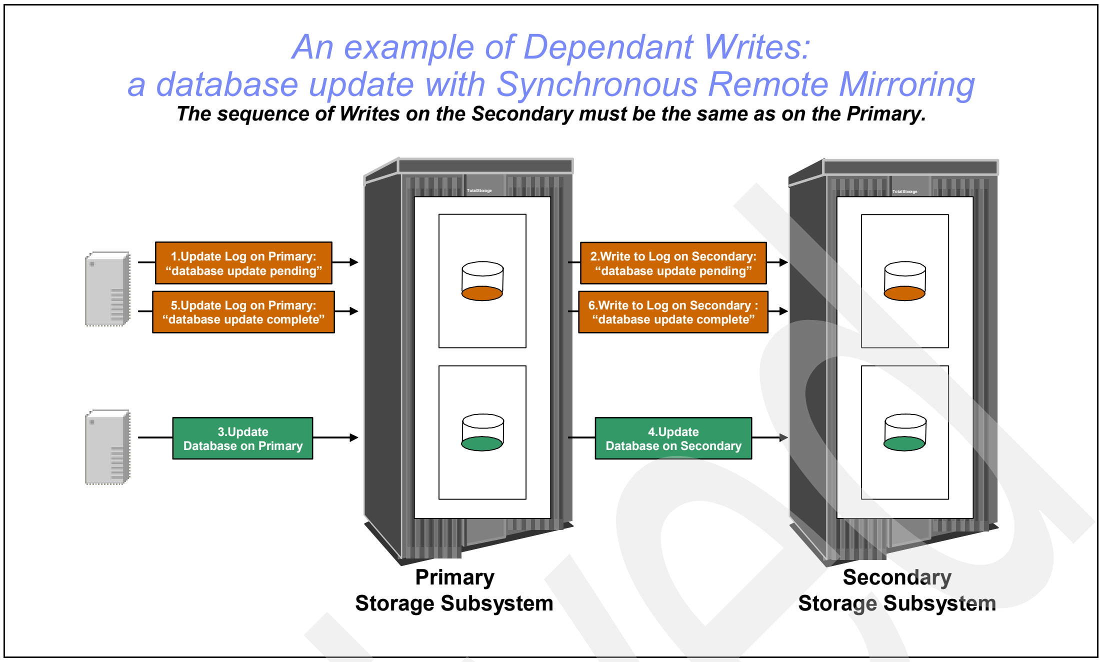

An example of Dependant Writes:
a database update with Synchronous Remote Mirroring
The sequence of Writes on the Secondary must be the same as on the Primary.
1.Update Log on Primary: "database update pending"
5.Update Log on Primary: "database update complete"
3.Update Database on Primary
2.Write to Log on Secondary: "database update pending"
6.Write to Log on Secondary : "database update complete"
4.Update Database on Secondary
Primary Storage Subsystem
Secondary Storage Subsystem

Figure 3-1 Dependant writes with Synchronous Mirroring

It is imperative that these dependent writes be written to disk in the same sequence that they were issued by the application as described in the sequence.

Continuing with the example, there is no guarantee that the database log and the database itself reside on the same storage subsystem. Failure to execute the write sequence correctly might result in writes (1) and (3) being executed, followed immediately by a system failure. When it comes time to recover the database, the database log would indicate that the transaction had completed successfully when, in fact, that is not the case. The transaction would be lost and the integrity of the database would be in question.

The technique for ensuring that the write order sequence is preserved at the secondary center is known as creating a consistency group.

When IBM Dual Copy (for example, RAID-1 disk mirroring) was introduced in the 1980's, the necessity of solving the dependent write issue was the main roadblock to implementation of disk mirroring across multiple storage subsystems. Creating consistency groups through the use of timestamping and data freeze techniques have overcome these issues.

In any disaster recovery solution involving data replication it is critical to understand the need for cross-volume data integrity and data consistency. Essential elements for cross-volume data integrity and data consistency include:

- The ability to create Point-in-Time (PiT) copies of the data as necessary
- The ability to provide for a secondary site "data freeze" in the case where a secondary site will be used for disaster recovery
- The ability to create data set/file consistency groups
- The ability to have a common restart point for data residing across multiple operating system platforms, reducing end-to-end recovery time

Cross-volume data integrity/data consistency enable a database restart to be performed in the event the second copy of the data is used.

Archived

Solutions that employ cross-volume mirroring and remote disk mirroring must address the issue of data consistency to support cross-volume and cross-storage subsystem data integrity.

Most customers, when designing a multi-site solution, need to minimize the time it takes to restart applications once the data at the secondary site has been recovered. Database recovery options that typically used to take hours to days can be avoided with today's technology.

3.1.4 Requirement for a common restart point

A new emerging requirement is the need for a common restart point for applications whose execution and data span multiple server platforms with the same or different operating systems. An example could be a branch office banking application. The user transaction at the teller might involve a Windows® system. The Windows system, in turn, routes the transaction to a System p® system within the branch office, which routes it to a z/OS system located at the corporate head office.

Several IBM remote copy software management packages, for example GDPS, can provide a common data freeze across z/OS and many different open system LUNs, thus providing a common restart point to reduce the time required to restart the application across the multiple system platforms. The IBM Global Mirror remote mirroring function provides the ability to manage both the z/OS Count-Key-Data format and the Distributed Systems Fixed Block data format within the same consistency group.

3.2 Tiers of multi-site service availability

In the 1980s, the Share Technical Steering Committee, working with IBM, developed a white paper that described levels of service for disaster recovery. In the paper, six tiers of disaster recovery were defined. These tiers are summarized beginning in 3.2.1, “Tier 0: No disaster recovery” on page 17.

An additional new seventh tier, representing the industry's highest level of availability driven by technology improvements, is outlined in 3.2.7, “Tier 7: IBM geographically dispersed parallel SYSPLEX (GDPS)” on page 20.

Today's on-demand business world has placed renewed emphasis on recovery time objectives and recovery point objectives. These factors, among others, are metrics that are used to evaluate an application's robustness.

3.2.1 Tier 0: No disaster recovery

Most customers in the industry today understand the need for disaster recovery as well as the need for backup of critical data. However, Tier 0 is still common, especially with non-critical data, that is, data that can be easily recreated or obtained from other sources if lost. Generally, for Tier 0 data, the cost of recreating the data or obtaining it from another source is less expensive than backing-up the data.

Archived

3.2.2 Tier 1 and Tier 2: Physical transport

The vast majority of customers today use a traditional method of creating tapes nightly and then transporting them to a remote site overnight. This method is known unofficially within the industry as: Pickup Truck Access Method (PTAM)

- Tier 1

  Tier 1 users send the tapes to a warehouse or 'cold' site for storage.

- Tier 2

  Tier 2 users send the tapes to a 'hot' site where the tapes can be quickly restored in the event of a disaster.

Various schemes have been developed to improve the process of offloading data nightly from production sites and production volumes to tape. Some of these solutions provide full integration with various databases (Oracle, DB2®, UDB, and so forth), as well as applications, for example SAP. Numerous names have been created to describe these off-line backup solutions including:

- Server-less backup
- LAN-less backup
- Split Mirroring
- SnapShot

Hardware vendors have created products to fit into this marketplace. The IBM FlashCopy function[2] provides this capability and, when coupled with disk mirroring solutions, can create a point-in-time copy of data within the same logical disk storage subsystem without impact to production applications. IBM FlashCopy is discussed in 3.3.2, "IBM enterprise storage server local data replication" on page 23.

3.2.3 Tier 3: Electronic vault transport

Today there are two ways to accomplish electronic tape vaulting. Both replace physically transporting tapes, with the tradeoff of the added expense for telecommunication lines.

- The first method is to write directly from the primary site to a tape storage subsystem located at the remote secondary site.
- A second method is introduced with the TS7700 Grid Tape replication solution. With IBM tape replication, a host writes data to a tape locally in the primary site. The TS7700 Grid hardware then transmits the data to a second TS7700 located at the secondary site. The TS7700 actually supports more than one target location in a grid. Various options are provided as to when the actual tape copy takes place, for example:
  - During tape rewind/unload
  - Asynchronously after rewind/unload

Figure 3-2 on page 19 illustrates an example of a TS7700 Grid configuration that would be used as part of a High Availability and disaster recovery solution[3].

[2] IBM FlashCopy is supported on the IBM DS8000, IBM DS6000 and IBM ESS800 storage subsystems.

[3] Some vendors recommend replacing electronic tape vaulting with electronic disk vaulting, which really is disk mirroring. The various capabilities of disk are marketed against the virtues of tape. With TS7700 tape one achieves an automatic second copy of tape data for disaster recovery purposes with minimal impact to the primary application production environment.

Archived

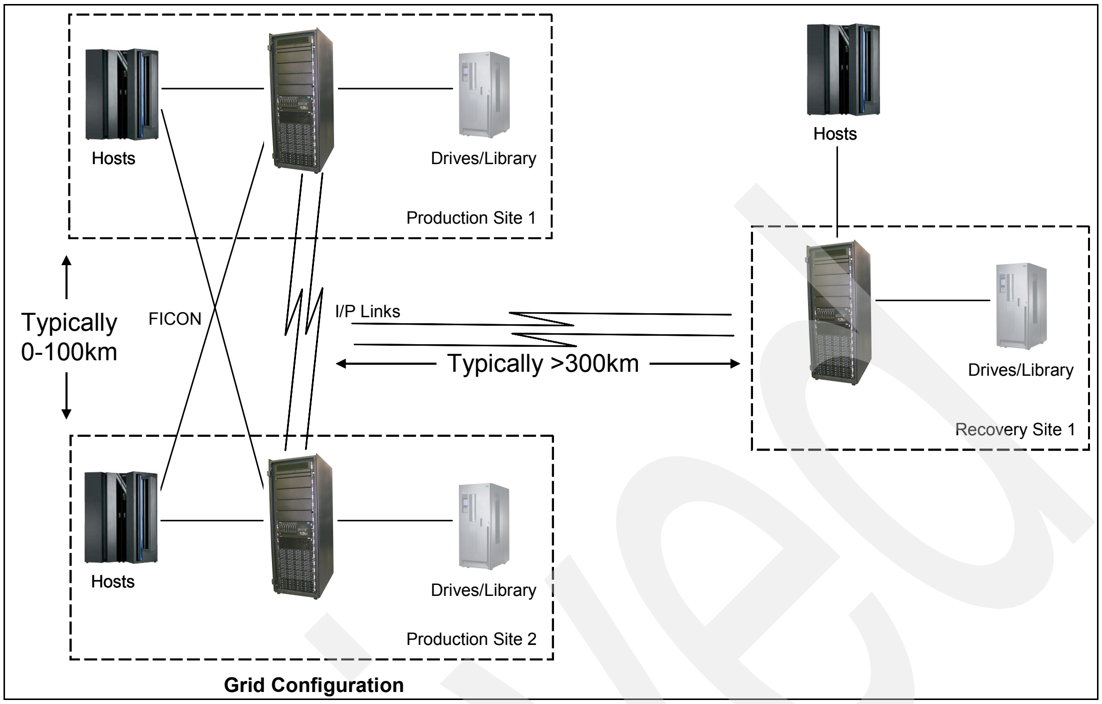


Figure 3-2 Typical TS7700 Grid deployment for HA and DR

3.2.4 Tier 4: Two active sites with application software mirroring

Various database, file system, or application-based replication techniques also have been developed to replicate current data to a second site, however these techniques are limited to data contained in the particular database or file system for which they were designed. An example of this in the Distributed/Open Systems world is software mirroring at the file system level. If all of the application data resides within the file system, these techniques can be a fast and efficient method of replicating data locally or remotely. Software-based file system mirroring can also be fully integrated with various host base server clustering schemes like AIX® PowerHA™(PowerHA was previously known as HAGEO.). Host failover causes the software mirroring to fail over as well.

These techniques are effective for a specific server platform or file system. But, from an operational point of view, in an enterprise with multiple server types and host operating systems, it can be difficult to understand which technique is used with each server and when things change, maintenance of the solution can be a challenge.

3.2.5 Tier 5: Two-site, two-phase commit

This technique is specific to the database used and its configuration in conjunction with the application environment. Various databases provide specific data replication of database logs, coupled with programs to apply the log changes at the secondary site. Databases using this approach include IMS™ with IMS RSR and IBM Tivoli's Replidata[4] solution for DB2.

Typically one only gets data consistency within the specific database and transactions across multiple databases are not supported.

[4] The formal names for these products are: Tivoli Infosphere Replication Server for zOS (DB2 Replidata) and Tivoli Classic Infosphere Replication Server for zOS (for DB2, IMS and other database systems).

Archived

3.2.6 Tier 6: Disk and tape storage subsystem mirroring

In this section we discuss the use of disk mirroring and tape mirroring as part of disaster recovery solution planning.

Disk mirroring

Disk mirroring is popular today because it can be implemented in the storage subsystem and, as a result, is independent of the host applications, databases and file systems which use the storage subsystem. There are six popular storage subsystem mirroring architectures, each with various options, available in the industry today:

- IBM Metro Mirror (PPRC) Remote Copy
- IBM Global Mirror (GM) remote mirroring
- IBM Extended Remote Copy (z/OS Global Mirror (XRC))
- EMC Symmetrix Remote Data Facility (SRDF/SRDF/A)
- Hitachi HARC
- Hitachi Universal Replicator (HUR)

The appeal of hardware-based disk mirroring is it can be used in conjunction with other enterprise disaster recovery packages and solutions to address all data within an enterprise.

Tape mirroring

Tape data can be mirrored through the IBM TS7700 Grid technology or by various software packages. For example, for System z hosts, DFSMShsm has a feature known as Aggregate Backup and Recovery Support (ABARS). ABARS can be used to create remote tape volumes for data sets that are not part of z/OS Global Mirror (XRC) or Metro Mirror (PPRC) remote copy pairs. Typically this is non-critical data, but data that still needs to be recovered in the event of a disaster that prevents moving back to the primary site for an extended period of time.

All tape replication schemes have the potential for some data loss. In the event of a disaster, the tapes actually being written to the remote site at the time of the disaster as well as any open tape data sets will not have all of their data copied successfully to the remote site. To address this issue, most customers place critical production data on disk and use one or more of the various disk mirroring solutions available. A similar software product for Open Systems is the Tivoli Storage Manager feature called the Disaster Recovery Manager.

3.2.7 Tier 7: IBM geographically dispersed parallel SYSPLEX (GDPS)

The highest level in the multi-site availability hierarchy can be implemented using GDPS. GDPS can help an installation provide the means to support the highest level of application availability. GDPS combines software and hardware to provide the ability to manage execution of a complete switch of all resources from one site to another automatically, providing continuous operations as well as disaster/recovery support for both planned and unplanned outages. GDPS is discussed in 3.4, “IBM geographically dispersed parallel SYSPLEX” on page 40.

Archived

3.3 IBM disk storage advanced copy services

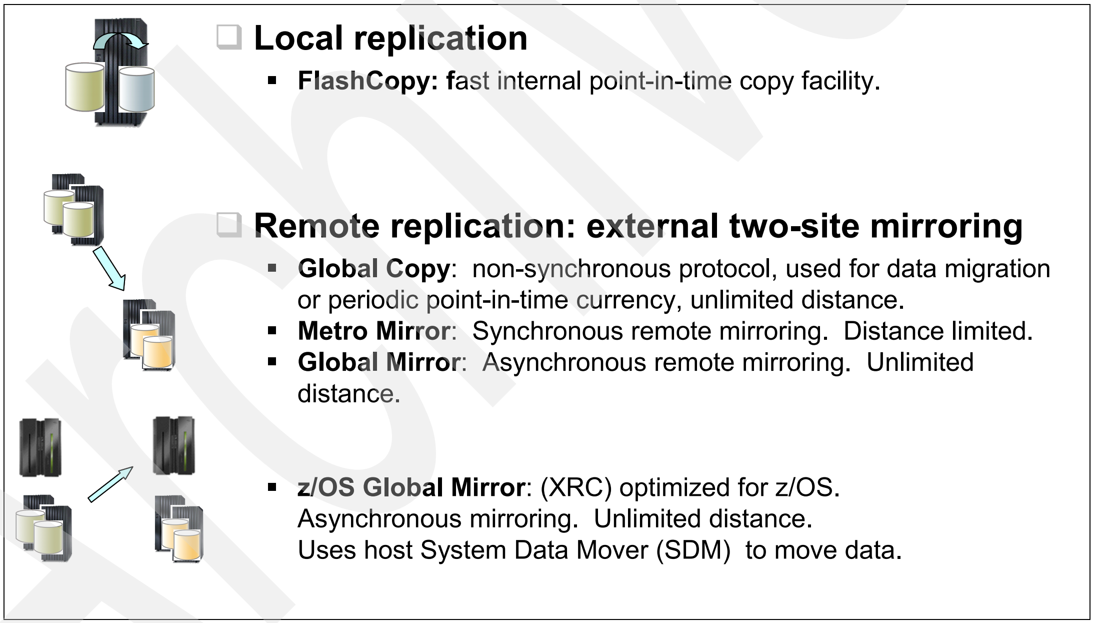
In the following sections we discuss IBM disk storage advanced copy services.

3.3.1 Core technologies for IBM storage copy services

Data replication is a core function for business continuance/disaster recovery. Data replication functions might be characterized as performing local replication or remote replication. The following two figures and table summarize the core technology copy services functions available on IBM eStorage.

An example of a local replication function is the IBM FlashCopy function shown in Figure 3-3. In the case of the IBM SAN Volume Controller (SVC), local replication can be accomplished within a single storage subsystem or between two storage systems that might be from the same vendor.

Remote replication might perform synchronous mirroring that is distance-limited or, asynchronous mirroring, which is designed to function without significant distance limitations. Figure 3-3 shows examples of two-site remote replication functions. The SVC can perform remote replication between like storage subsystems or between subsystems from different vendors.



❑ **Local replication**

- **FlashCopy:** fast internal point-in-time copy facility.

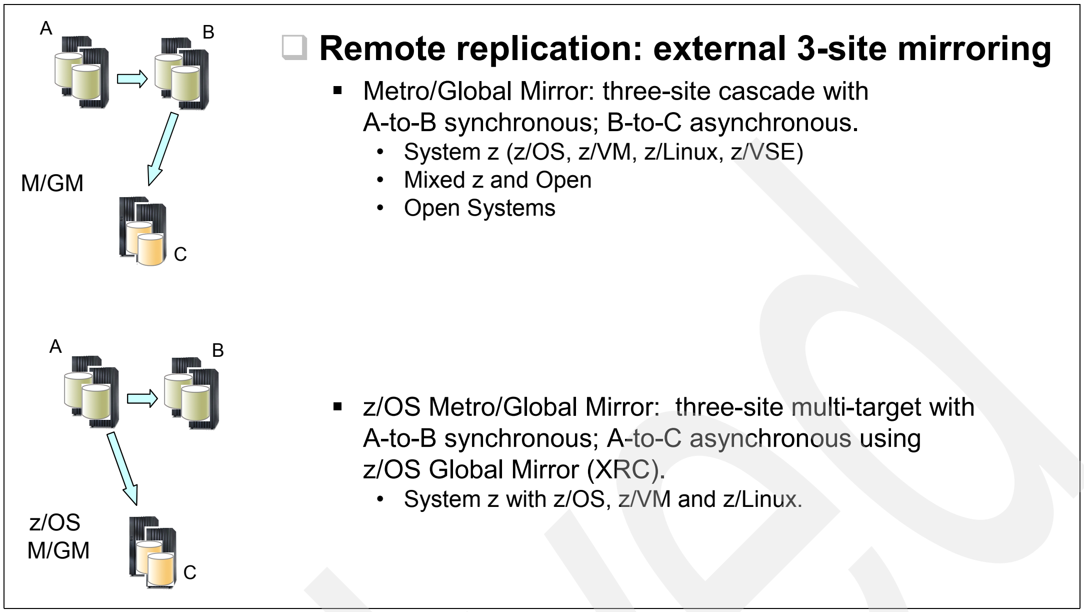
❑ **Remote replication: external two-site mirroring**

- **Global Copy**: non-synchronous protocol, used for data migration or periodic point-in-time currency, unlimited distance.
- **Metro Mirror**: Synchronous remote mirroring. Distance limited.
- **Global Mirror**: Asynchronous remote mirroring. Unlimited distance.
- **z/OS Global Mirror**: (XRC) optimized for z/OS. Asynchronous mirroring. Unlimited distance. Uses host System Data Mover (SDM) to move data.

Figure 3-3 Local and Two-site Remote Replication

Archived

Figure 3-4 shows two examples of three-site remote replication.

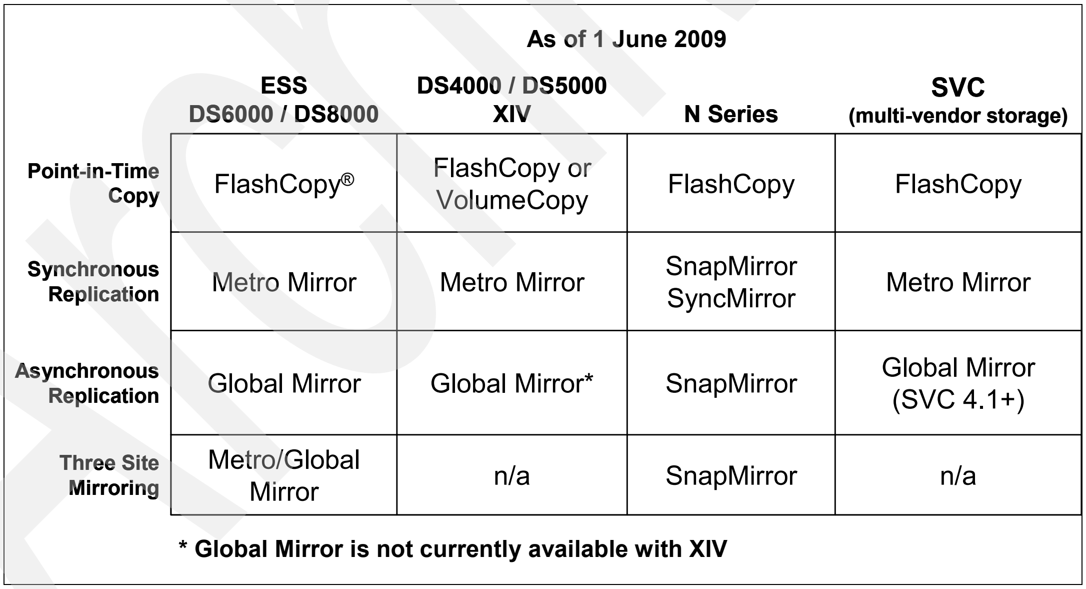


❑ **Remote replication: external 3-site mirroring**

- Metro/Global Mirror: three-site cascade with A-to-B synchronous; B-to-C asynchronous.
  - System z (z/OS, z/VM, z/Linux, z/VSE)
  - Mixed z and Open
  - Open Systems
- z/OS Metro/Global Mirror: three-site multi-target with A-to-B synchronous; A-to-C asynchronous using z/OS Global Mirror (XRC).
  - System z with z/OS, z/VM and z/Linux.

Figure 3-4 Three-site remote replication

The table in Figure 3-5 provides a summary of the function names for the various data replication options and features available on IBM eStorage systems. Additional detail relative to specific data replication features and functions will be discussed in the sections that follow.



As of 1 June 2009

	ESS DS6000 / DS8000	**DS4000 / DS5000 XIV**	**N Series**	**SVC (multi-vendor storage)**
Point-in-Time Copy	FlashCopy®	FlashCopy or VolumeCopy	FlashCopy	FlashCopy
Synchronous Replication	Metro Mirror	Metro Mirror	SnapMirror SyncMirror	Metro Mirror
Asynchronous Replication	Global Mirror	Global Mirror*	SnapMirror	Global Mirror (SVC 4.1+)
Three Site Mirroring	Metro/Global Mirror	n/a	SnapMirror	n/a

*** Global Mirror is not currently available with XIV**

Figure 3-5 Advanced copy services functions for IBM storage

Archived

3.3.2 IBM enterprise storage server local data replication

This section provides detailed information on the FlashCopy function.

FlashCopy for IBM eStorage[5]

The IBM FlashCopy function provides the ability to make a fast internal copy of a z/OS volume or Open Systems LUN. The FlashCopy function is available on IBM disk storage systems. This section discusses some of the capabilities of the different versions of FlashCopy and how customers can exploit them to address business continuity requirements. A summary of the various functional capabilities available with IBM FlashCopy is shown in Figure 3-9 on page 27.

Basic or traditional FlashCopy

The FlashCopy function copies data from a source to a target. The size of the target must be equal or greater size than the source. Once the source/target relationship is established both the source and target are available to servers for read and write access.

IBM FlashCopy has a unique capability called the COPY/NOCOPY option. This function is designed to permit the user additional flexibility to decide at the time FlashCopy is invoked whether or not to make a physical copy of the data by way of background processing. If no physical copy is made, the source physical copy is maintained with pointers to the active copy of the data as well as to the virtual point-in-time copy. This feature is designed to make copies of the complete volumes that can then be varied online to systems and used for multiple purposes. In addition a command, `Start Background Copy Now`, can be used to invoke a background copy when required. As a result, most executions of FlashCopy use the NOCOPY option. Figure 3-6 shows how IBM FlashCopy works.

A common use for the No Background (NOCOPY) copy option is for backing-up data. A full background copy would be recommended in situations where the point-in-time copy is actively used by a number of users, for example, database queries.

Once the second copy of the data has been created by using FlashCopy, it might be necessary to copy the data to tape. Software products that provide the ability to dump data from disk to tape include DFSMSdss and DFSMShsm on System z (z/OS) hosts and the Tivoli Storage Manager (TSM), which is available on a variety of host platforms. Figure 3-6 illustrates the FlashCopy function on the IBM DS8000. The same figure could be used for the DS4000/DS5000, SVC and XIV disk storage subsystems.

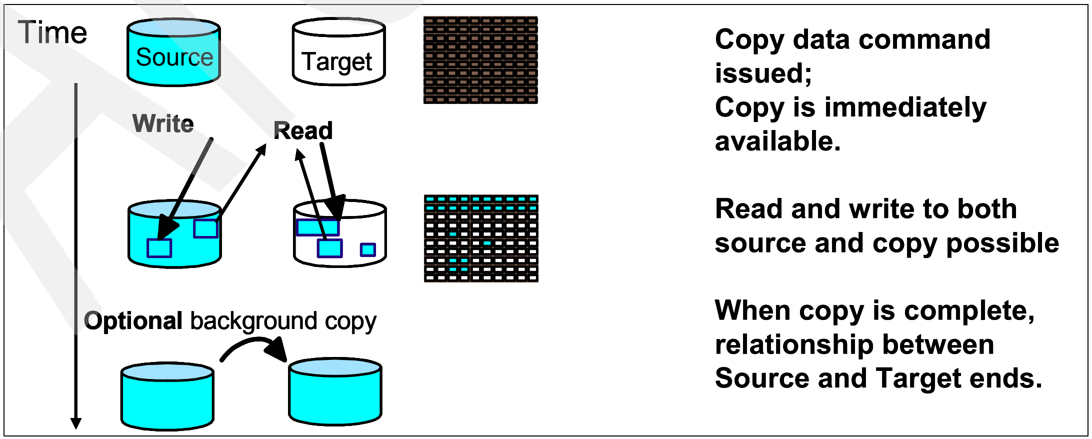

Time
Source
Target
Write
Read
Optional background copy
Copy data command issued;
Copy is immediately available.
Read and write to both source and copy possible
When copy is complete, relationship between Source and Target ends.

Figure 3-6 IBM eStorage FlashCopy

[5] The IBM SAN Volume Controller (SVC) is not strictly a disk subsystem since it does not contain disks itself. References in the text to Source and Target for the SVC are for Open Systen LUNs only, not z/OS volumes.

Archived

Space Efficient FlashCopy

The DS800 also offers a second version of FlashCopy, known as Space Efficient FlashCopy (SEFC). This feature is offered for z/OS volumes but not Open Systems LUNs. SEFC provides the ability to replace the need for a target volume of the same or larger physical size with a common repository that holds multiple targets. The difference is that the common repository only holds the target volume updates and does not require the same amount of space as the source. This capability can be considered to be a form of thin provisioning. The data held in the common repository is maintained for the duration of the source/target relationships. The common SEFC repository must be of sufficient size to contain all write I/O updates against all FlashCopy Source volumes for the duration of the source/target FlashCopy relationships.

Consistent FlashCopy

IBM Consistent FlashCopy is designed to provide the ability to create a copy of data such that the copy is I/O-consistent. The installation identifies a set of volumes or LUNs across any number of physical volumes or storage subsystems, such that when the FlashCopy source/target relationship is established, all I/O to all source volumes is temporarily busied, permitting the I/O-consistent FlashCopy to be created.

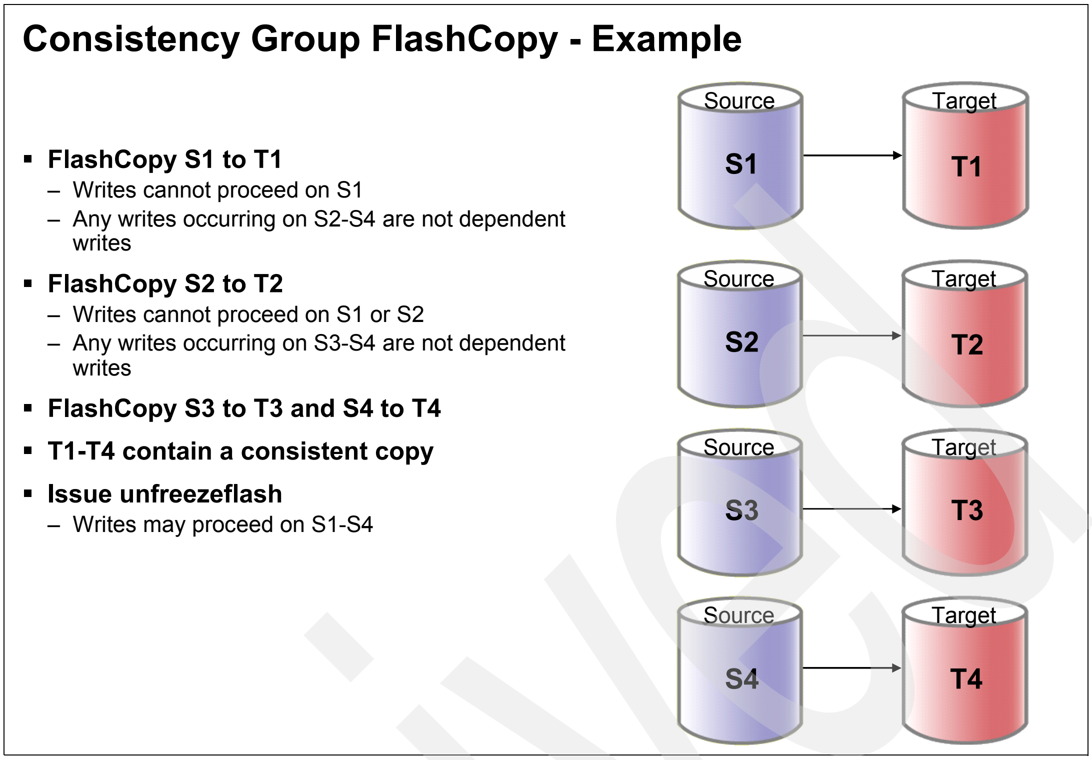
Consistent FlashCopy has been a help to many customers who have lost track of the relationship between applications and the data actually used. This typically is discovered through a root cause analysis from disaster recovery tests that restore numerous backup takes and fail as the result of data from the various tapes being inconsistent across the entire application or suite of applications.

Consistent FlashCopy can be used to ensure consistency for many tasks, for example:

- Backup
- Cloning a database
- Moving work to other hosts or servers

Consistent FlashCopy provides this capability within a single eStorage subsystem or across multiple eStorage subsystems. An illustration of how Consistent FlashCopy works is shown in Figure 3-7 on page 25.

Archived



Consistency Group FlashCopy - Example

- **FlashCopy S1 to T1**
  - Writes cannot proceed on S1
  - Any writes occurring on S2-S4 are not dependent writes
- **FlashCopy S2 to T2**
  - Writes cannot proceed on S1 or S2
  - Any writes occurring on S3-S4 are not dependent writes
- **FlashCopy S3 to T3 and S4 to T4**
- **T1-T4 contain a consistent copy**
- **Issue unfreezeflash**
  - Writes may proceed on S1-S4

Figure 3-7 Creating FlashCopy Consistency Groups

Point-in-Time Copy

Point-in-Time Copy is also provided through a combination of FlashCopy and Metro Mirror. Consistent FlashCopy functionality helps reduce the requirement for an additional set of Metro Mirror (PPRC) or z/OS Global Mirror (XRC) secondary volumes. However, the data freeze technique used in combination with Metro Mirror might have less impact to the primary application I/O than Consistent FlashCopy.

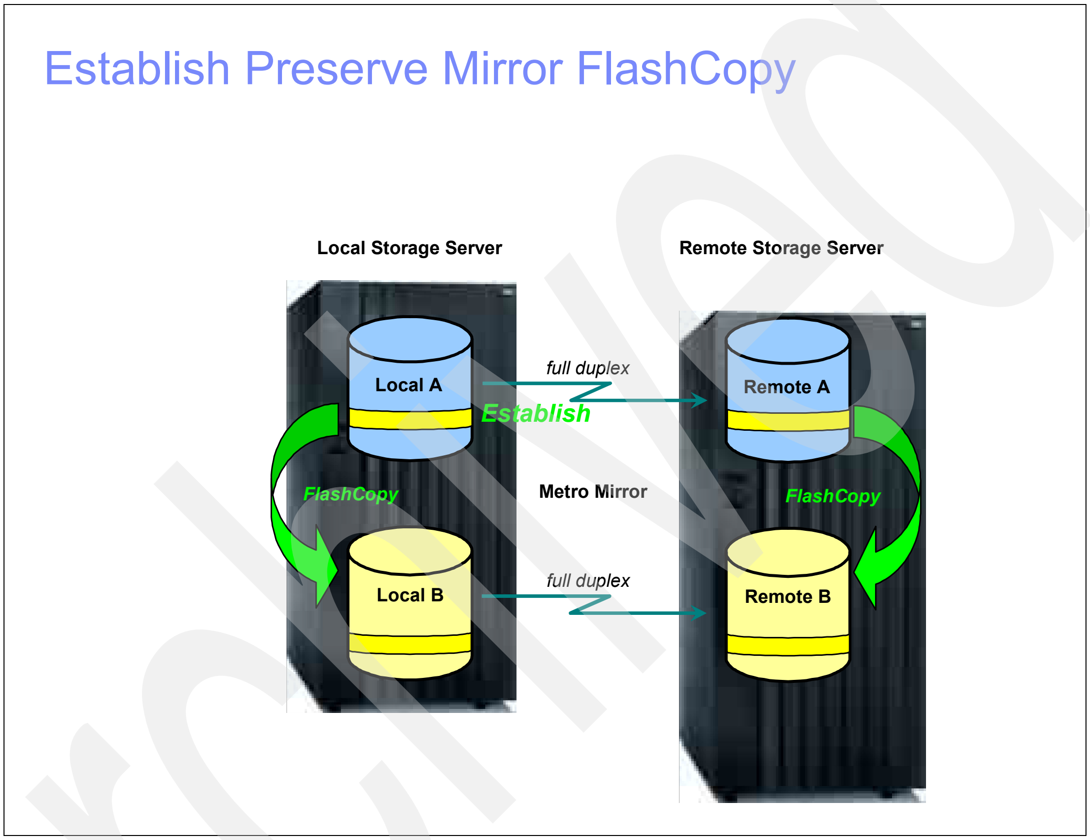
Recent FlashCopy Enhancements for DS8000

The DS8000 FlashCopy function now supports additional capabilities:

- FlashCopy relationships can now be established between any source and target volume independent of the DS8000 Logical Storage Subsystem (LSS).
- A FlashCopy Target can be a Metro Mirror (PPRC) or Global Copy (PPRC-XD) primary volume.
- A single source might be established with up to twelve targets. In a z/OS environment, a single volume might not be both a source and a target. However, when using Data Set Level FlashCopy, various extents of a volume that are part of a Data Set Level FlashCopy pair might simultaneously be a source or target for the Metro Mirror or Global Mirror remote mirroring functions.
- If a full copy is done between a source and target, it can be established with an attribute called *persistent*. Subsequent FlashCopies to the target can then be performed as incremental changes to the initial full FlashCopy.
- DFSMSdss now exploits Data Set Level FlashCopy for the DFSMSdss COPY function.

Archived

- When two primary volumes are in a FlashCopy relationship, and are simultaneously being mirrored to a pair of remote secondary volumes, it is now possible to issue a FlashCopy command at the primary volume that will cause a corresponding FlashCopy to be executed at the secondary volume. This avoids having to send incremental changes continually from the FlashCopy target at the primary volume to the corresponding FlashCopy Target at the secondary volume, reducing the demands on link bandwidth and reducing the time lag (for example, RPO) between the target at the primary volume and the target at the secondary volume. This function is called Remote Pair FlashCopy and is shown in Figure 3-8.

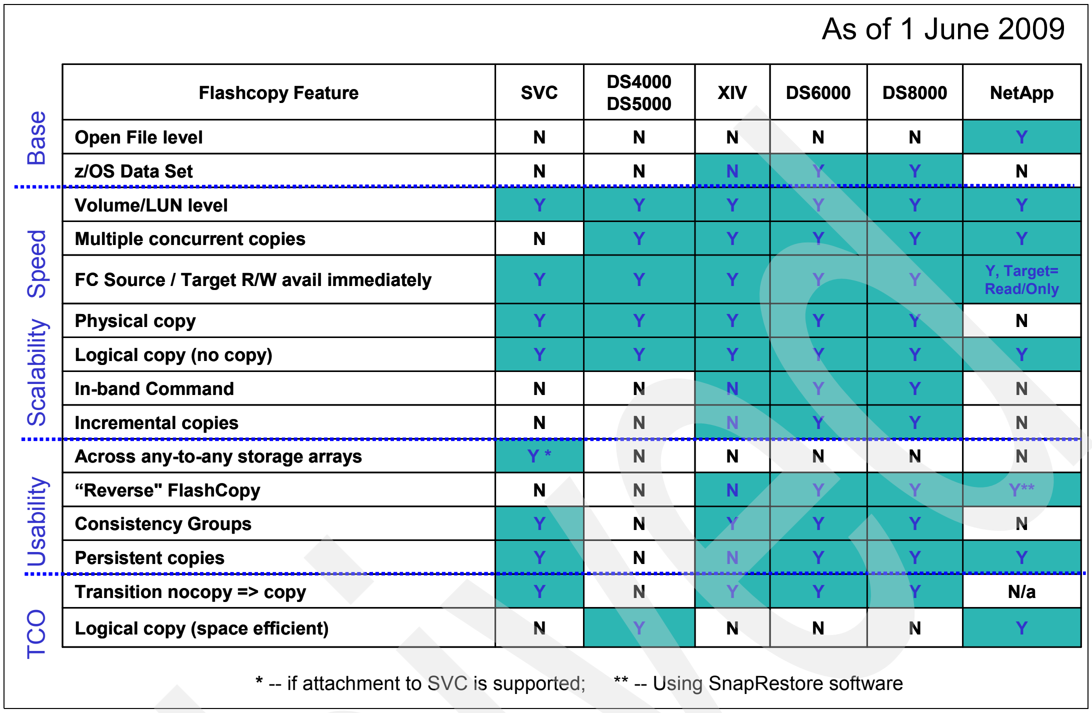


Establish Preserve Mirror FlashCopy

Figure 3-8 Remote Pair FlashCopy

Archived

A summary of the various functional capabilities available with IBM FlashCopy is shown in the table in Figure 3-9.



As of 1 June 2009

	Flashcopy Feature	SVC	DS4000 DS5000	XIV	DS6000	DS8000	NetApp
Base	Open File level	N	N	N	N	N	Y
	z/OS Data Set	N	N	N	Y	Y	N
Speed	Volume/LUN level	Y	Y	Y	Y	Y	Y
	Multiple concurrent copies	N	Y	Y	Y	Y	Y
	FC Source / Target R/W avail immediately	Y	Y	Y	Y	Y	Y, Target= Read/Only
Scalability	Physical copy	Y	Y	Y	Y	Y	N
	Logical copy (no copy)	Y	Y	Y	Y	Y	Y
	In-band Command	N	N	N	Y	Y	N
	Incremental copies	N	N	N	Y	Y	N
Usability	Across any-to-any storage arrays	Y *	N	N	N	N	N
	“Reverse" FlashCopy	N	N	N	Y	Y	Y**
	Consistency Groups	Y	N	Y	Y	Y	N
	Persistent copies	Y	N	N	Y	Y	Y
TCO	Transition nocopy => copy	Y	N	Y	Y	Y	N/a
	Logical copy (space efficient)	N	Y	N	N	N	Y

* -- if attachment to SVC is supported; ** -- Using SnapRestore software

Figure 3-9 FlashCopy functions for IBM storage

3.3.3 IBM system storage remote data replication

This section outlines the various IBM remote copy solutions available on IBM eStorage subsystems. These solutions can be selected to implement the Tier 1 through Tier 7 multi-site availability options described in 3.2, “Tiers of multi-site service availability” on page 17.

Synchronous remote mirroring

Synchronous remote replication, also called synchronous mirroring, is one of two generic types of remote mirroring. The other type of remote mirroring, known as asynchronous remote replication or asynchronous mirroring is discussed in “IBM DS8000 asynchronous remote mirroring” on page 32.

Synchronous remote mirroring characteristics

The most important characteristic of synchronous data replication is when a write I/O is issued by a server to the primary disk storage subsystem:

1. The primary issues a write I/O to the secondary.
2. The primary waits until it has received notification from the secondary that the data was successfully received by the secondary.
3. Upon receiving notification from the secondary, the primary notifies the server that the write I/O was successful.

If cross-volume/cross-LUN data consistency is required, then typically a 'software' management function, called Consistency Group (CG) processing, also is required in conjunction with synchronous data replication. Consistency Groups are discussed above.

Archived

Synchronous remote disk mirroring is the most common storage-based data replication solution and has been implemented by IBM, EMC, Hitachi, Sun Microsystems, and additionally by numerous other companies who have implemented block-based synchronous data mirroring software in the host or in SAN switches and appliances. Vendor implementation of synchronous mirroring varies, but generally accomplish the same essential goal (for example, securing data at both sites before notifying the host that the I/O operation completed).

Mitigating remote link latency for synchronous mirroring

An inherent delay in synchronous mirroring is due to the time required to send an electrical signal or optical pulse from the primary subsystem to the secondary and the corresponding reply from the secondary back to the primary. This delay, known as *latency*, is different from the time required to transmit data due to the limitations imposed by link bandwidth. The speed of light in optical fiber is about 5 µsec/km. This works out to a latency of 0.5 msec/100 km, or, 1 msec/100 km round-trip.

Hardware vendors try to mitigate the overheads and latency inherent in synchronous mirroring data transmission by using techniques such as reduction in the synchronous mirroring link overheads per se, pipe lining I/Os, and exploiting different link protocols. Generally, the more complicated the transmission algorithm (for example, pipe-lining techniques) the more time is spent in the synchronous link overhead to maintain I/O data integrity and cross-volume data consistency.

IBM storage subsystems use a feature known as *Pre-Deposit Write*, which improves the normal Fibre Channel Protocol (FCP) from the standard two protocol exchanges to an optimized single protocol exchange. All IBM disk storage subsystems exploit the Pre-Deposit Write feature. Several I/O requests are sent as one I/O package, reducing the number of handshakes required to execute an I/O chain. This technique, coupled with the high speed microprocessors within the storage subsystems and the microprocessors in the storage subsystem adapter cards, have helped extend Metro Mirror synchronous distance to 300 km with good performance.

A key design point to remember when looking at various synchronous copy implementations is that all vendors attempt to hide the synchronous remote copy overheads from the application, but the architecture is still synchronous. Installations with performance-sensitive applications should consider an asynchronous disk mirroring solution like Global Mirror or z/OS Global Mirror (XRC), which are designed to minimize the impact to the primary application I/O at any distance.

IBM DS8000 synchronous mirroring

IBM's synchronous remote disk mirroring function is called Metro Mirror[6]. IBM DS8000 Metro Mirror works with data formatted as System z count-key-data (CKD) volumes and with Distributed/Open Systems Fixed Block (FB) volumes. IBM also offers synchronous mirroring on other storage products such as the IBM DS5000, IBM DS4000, IBM N series and the IBM SAN Volume Controller (SVC).

The IBM Metro Mirror architecture as implemented for System z is available across multiple disk vendors including EMC, HDS and Sun Microsystems. Specific features and functions can vary across vendors. Check with each vendor for specifics as it relates to their implementation.

[6] Previously the IBM DS8000 version of Metro Mirror was known as Peer-to-Peer Remote Copy (PPRC).

Archived

IBM DS8000 Metro Mirror for System z

In addition to the Pre-Deposit Write function, an excellent solution to help mitigate Metro Mirror overheads is the use of the HyperPAV/ Parallel Access Volume (PAV), Multiple Allegiance and I/O priority queuing functions that are available on the IBM when used in conjunction with a System z host running z/OS. Several customers have implemented Metro Mirror with these options and have seen an increase in the overall throughput of an application, because I/O operations are no longer issued serially to a device. IOSQ time has been greatly reduced or eliminated through the use of HyperPAV/PAV, Multiple Allegiance and I/O Priority Queuing.

It also is possible to create point-in-time copies of data at the Metro Mirror secondary site. TPC-R, the native DS8000 command line interface (CLI) or GDPS/PPRC can be used to invoke FlashCopy while specifying the Metro Mirror secondary logical volumes as the FlashCopy Source volumes. This technique makes it possible to create a point-in-time copy of all data at the secondary site. Subsequently, the FlashCopy Target volumes can be used to perform various functions like backup, input for end-of-day reports, database query servers, application testing, disaster recovery testing, and so forth.

IBM DS8000 Metro Mirror for Open Systems

The IBM Metro Mirror architecture also works with Open Systems (fixed block) devices in the same manner that it operates with System z Count-Key-Data (CKD) devices. Management of the Open Systems Metro Mirror environment is done via a browser-based management tool, Tivoli Storage Productivity Center for Replication (TPC-R), or via a DS8000 function called "Open LUN" support under GDPS/PPRC. Metro Mirror is also fully integrated with some IBM high availability clustering solutions like PowerHA for AIX (HACMP™), Windows-NT MSCS, Sun Veritas Clusters or HP-UX clusters.

Similar to the technique discussed above, TPC-R, the native DS8000 command line interface (CLI) or GDPS/PPRC Open LUN support can be used to invoke FlashCopy at the Metro Mirror secondary and specify the Metro Mirror secondary logical volumes as the Source volumes for the FlashCopy. This technique enables a point-in-time copy of all data at the secondary site similar to what is described for a z/OS environment.

As with System z Metro Mirror, the Distributed/Open System Metro Mirror will operate at distances up to 300 km with reasonable performance. The Pre-Deposit Write feature is implemented to reduce the FCP protocol exchanges to only one exchange for a write I/O. Beyond approximately 300 km one would typically look at an asynchronous data replication technique to mitigate application performance issues.

AIX HACMP V5.1 provides a feature that fully integrates DS8000, DS6000, ESS 800, DS4000, DS5000 and SVC Distributed/Open Metro Mirror with PowerHA, such that a cluster failover/fallback will also invoke a Metro Mirror failover/fallback. IBM has similar offerings with MSCS clusters. The Metro Mirror architecture also has been integrated with Veritas Cluster Manager, Sun's cluster manager and IBM Tivoli AppMan solutions under the GDOC automation.

IBM DS8000 Metro Mirror Fibre Channel Link Support [7]

In 2003, IBM added the capability to utilize Fibre Channel connections for Metro Mirror (PPRC) links. Prior versions of Metro Mirror (PPRC) used ESCON® links. ESCON is a 200 Mb/sec. serial channel that can achieve instantaneous data transfer rates up to about 18 MB/sec. ESCON channels are single threaded. They only allow one I/O to use the channel at a time. As a result additional Metro Mirror (PPRC) write concurrency could only be achieved

[7] This section is extracted from: "IBM TotalStorage® Enterprise Storage Server® Model 800 Version 2.3 Performance White Paper", by Lee La Frese, ESS Performance Evaluation, IBM Corporation, Tucson, AZ, USA. For a copy of this White Paper, please contact your IBM representative.

Archived

by using more ESCON links. For ESCON links on the DS8000/DS6000/ESS800, link propagation delay adds about 20 microseconds per km to the Metro Mirror (PPRC) write service time on a logical track basis. Because the ESCON link is single threaded, no other I/O can use the channel during the propagation delay. As a result, there is a substantial drop in throughput on an ESCON channel as the distance increases.

Fibre Channel links are a significant technological improvement over ESCON. The eStorage host adapters support full duplex fibre channel operation at two or four Gb/sec. The instantaneous data transfer rate is up to about 200/400 MB/sec. Fibre Channel links are multithreaded, so many I/O operations can operate concurrently on a single Fibre Channel link. The protocol used with Fibre Channel links is more efficient than ESCON so only about ten microseconds per km of propagation delay is added to the Metro Mirror (PPRC) write service time. Multithreading helps prevent the aggregate throughput capability of a Fibre Channel link from degrading with distance. Theoretically, this applies to any distance. However, IBM requires an RPQ for Metro Mirror (PPRC) distances beyond 300 km, so the anticipated I/O response time implications might be reviewed with the customer before implementing. Practically speaking, synchronous Metro Mirror (PPRC) might provide acceptable performance for many users even at distances up to 300 km when using Fibre Channel links with dark fiber or Dense Wave Division Multiplexing (DWDM) [8].

Fibre Channel links for the IBM Disk Storage Subsystems enable a substantial reduction in the infrastructure required for implementing a Metro Mirror solution on System z. This is because a single Fibre Channel link, depending on the workload, can handle between four and 12 times the throughput of an ESCON link at 0 km based on laboratory measurements. At distance, the improvement ratios can be even higher.

Asynchronous Remote Data Copying

The following section provides information on IBM asynchronous remote data copying solutions.

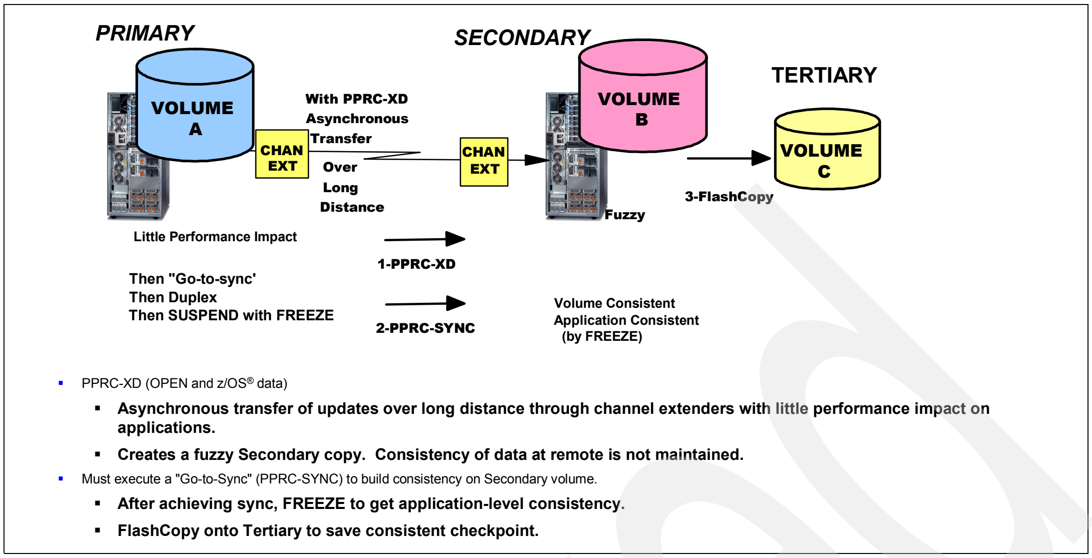
IBM DS8000 Global Copy (PPRC-XD)

Long distance data migration can be achieved using Global Copy (PPRC-XD). Figure 3-10 on page 31 illustrates Global Copy (PPRC-XD) operation:

1. A write command and data is received by the storage subsystem at the primary site. The primary storage subsystem immediately signals "I/O successful completion" back to the host application.
2. The data is transmitted asynchronously from the primary storage subsystem to the secondary storage subsystem at the backup site.

[8] Repeaters might be needed for the Fibre Channel links to function at extended distances.

Archived



- PPRC-XD (OPEN and z/OS® data)
  - **Asynchronous transfer of updates over long distance through channel extenders with little performance impact on applications.**
  - **Creates a fuzzy Secondary copy. Consistency of data at remote is not maintained.**
- Must execute a "Go-to-Sync" (PPRC-SYNC) to build consistency on Secondary volume.
  - **After achieving sync, FREEZE to get application-level consistency.**
  - **FlashCopy onto Tertiary to save consistent checkpoint.**

Figure 3-10   IBM Global Copy overview

The secondary will tend to lag behind the primary. The amount of data queued to be sent to the secondary will depend on the intensity of the primary write activity and the distance between the primary and secondary. Commands are provided to enable the installation to force the volume, or set of volumes, participating in the Global Copy (PPRC-XD) session go synchronous at a given point in time to allow the secondary to catch-up and then have a consistent set of data. Additionally, the volumes will naturally go synchronous if the primary site applications are quiesed or the application workload tapers off.

Global Copy (PPRC-XD) works with both distributed systems and System z data and will scale to the bandwidth available between the primary and secondary sites. It is mainly used for the following purposes:

- Data migration of one storage subsystem to another within a single site or across multiple sites.
- To send database logs to a remote site in a Log Apply D/R solution.
- Global Copy in combination with Incremental FlashCopy to provide a point-in-time copy of data at a remote site that is refreshed on a regular interval.

Archived

Combinations of FlashCopy and Global Copy

Figure 3-11 illustrates using the combination of Incremental FlashCopy with Global Copy to provide a point-in-time copy of data at a remote site. The incremental FlashCopy can then be repeated at a regular interval (for example, hourly, once a shift, or nightly) to refresh the data at the remote location. For many applications, this procedure might provide a low cost alternative data replication solution and be a cost-effective improvement over nightly backups to tape.

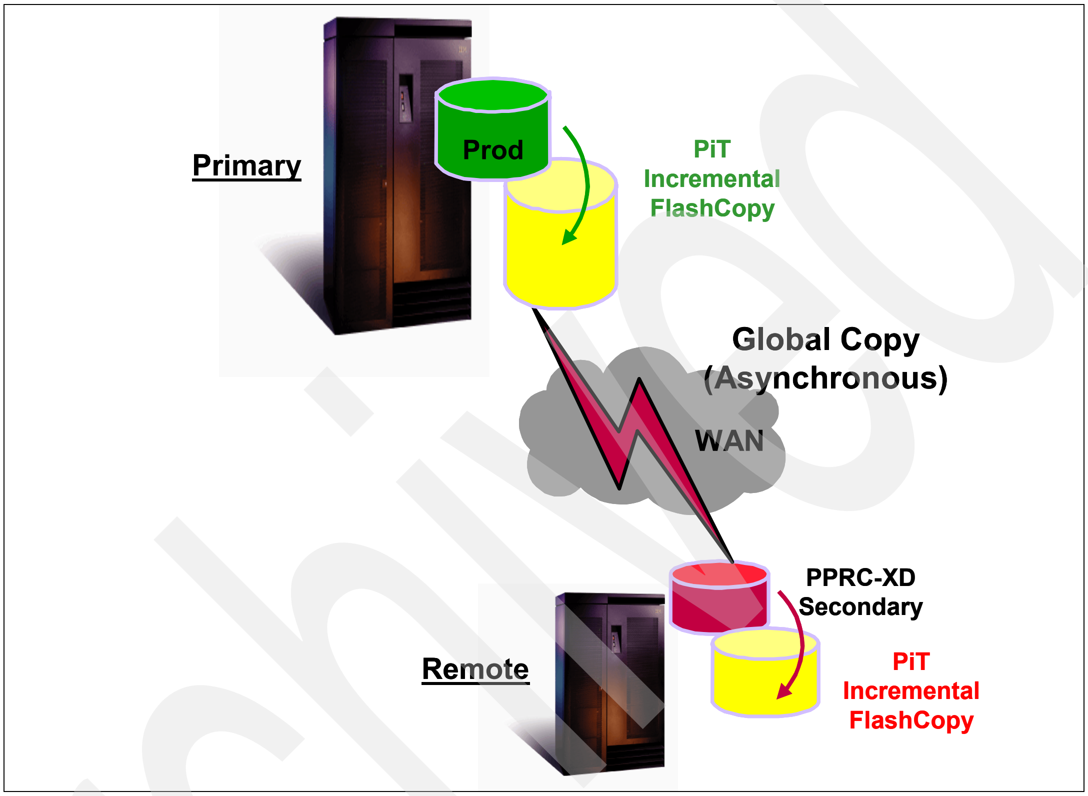


Figure 3-11 Remote replication using Global Copy plus Incremental FlashCopy

IBM DS8000 asynchronous remote mirroring

In addition to synchronous mirroring, the other generic form of remote mirroring is asynchronous mirroring. IBM offers two different forms of asynchronous mirroring for DS8000 storage subsystems:

- IBM DS8000/DS6000/ESS Global Mirror

  This is a function that runs on IBM DS8000 storage systems.

- z/OS Global Mirror (XRC)

  This runs on a System z host in conjunction with IBM DS8100/DS8300 storage as well as selected non-IBM storage systems that can attach to System z hosts.

IBM DS8000 Global Mirror

IBM DS8000 Global Mirror is a two-site unlimited distance data replication solution for both System z and distributed systems data. An example of a Global Mirror configuration is shown in Figure 3-12 on page 33.

Archived

Archived

Global Mirror is designed to provide cross-volume and cross-storage subsystem data integrity and data consistency. Data consistency is managed outboard by the disk storage subsystems. The design objectives for DS8000 Global Mirror are as follows.

- Keep the currency of that data at the secondary consistent to within three to five seconds of the primary, assuming sufficient bandwidth and storage system resources.
- Do not impact the storage subsystem's host response time when insufficient bandwidth and system resources are available.
- Be scalable. Provide data consistency and predictable host response time across multiple primary and secondary disk subsystems.
- Verify end-to-end data integrity. This is accomplished through:
  - Transmission and verification of Cyclic Redundancy Checking (CRC) codes, which are included with each changed data block or record.
  - Detection of dropped FCP frames.
  - For ECKD™ devices, transmission and verification of the track format is maintained in metadata.
- Removal of duplicate writes within a consistency group before transmitting the data to the secondary at the remote site.
- In situations where the primary-to-secondary link bandwidth is constrained, continue operating by allowing the RPO to increase.
- Support creation of a consistent set of data at the secondary when the primary and secondary storage subsystems are partitioned to contain both System z CKD data and Open Systems Fixed Block data.

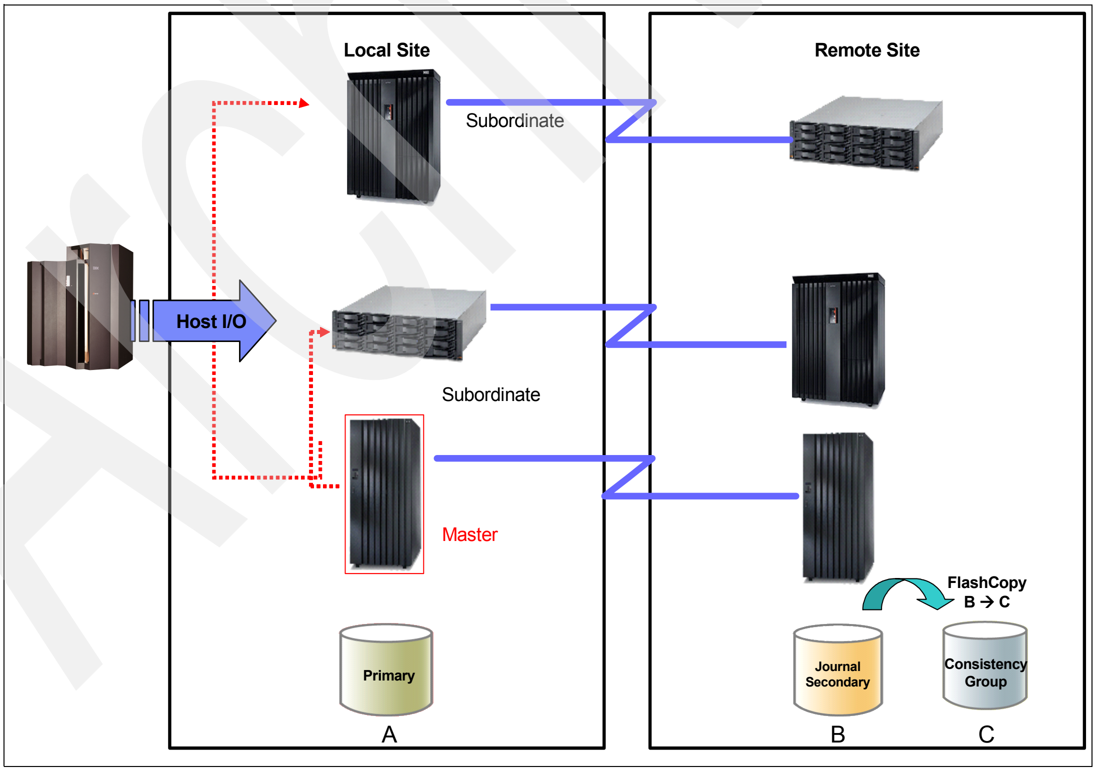


Figure 3-12 DS8000 Global Mirror architecture

It is possible to form consistency groups across all data managed by Global Mirror. These consistency groups are automatically managed by the DS8000 microcode through a policy defined as part of the Global Mirror setup and then implemented when the data replication master and subordinate session is initialized. The policy can also be modified dynamically. To set up Global Mirror, the installation specifies:

1. The volume mirroring pairs across the primary disk storage subsystems and the Target Remote site disk subsystems, and
2. How often to form consistency groups yielding the desired RPO. This, in turn, defines the maximum application pause time (that is, the time used by the master logical storage subsystem to coordinate subordinate logical storage subsystems within the disk storage subsystem or across multiple disk storage subsystems (scalable to 16 primary disk storage subsystems) through a Fibre channel connection).

Software such as the IBM GDPS/GM, GDPS/MGM and TPC-R help to setup, manage, monitor, and report on the end-to-end data replication function.

IBM DS8000 Global Mirror operation

All DS8000 Global Mirror operations are controlled and managed outboard by the master Logical Storage Subsystem (LSS) in a master/subordinate relationship that is defined during the initial setup of a Global Mirror session.

Each LSS defined in the Global Mirror session controls the flow of changed records/blocks to their corresponding LSSs at the remote site. The master LSS coordinates the formation of consistency groups based on the policy provided for the session. The master LSS will cause a pause to occur across all LSS-to-LSS pairs long enough to coordinate each subordinate LSS pair. This coordination involves swapping bitmaps that denote which tracks/data blocks have changed but have not yet been transmitted to the secondary.

The coordination initiates the transition from one consistency group to the next. Once all data for the consistency group is secured on the remote site B volume, the master LSS coordinates the hardening of the data to the C volume. This is accomplished by sending an order to the secondary subsystems to execute a FlashCopy to copy the contents of the B volumes to the C volumes at the secondary.

This process repeats itself at regular intervals. The time interval is controlled by the installation-specified consistency group interval time parameter. The data transfer process aggregates record/block write transfers to the remote LSS, scaling and optimizing the data transmissions to the bandwidth provided. For a specific write I/O, its consistency group is marked through the volume bitmaps at the primary site, but the actual consistency group is formed at the remote site through the B to C FlashCopy operation.

IBM DS8000 System z extended distance remote copy (XRC)

The IBM DS8000[9] Global Mirror for System z asynchronous disk mirroring architecture is currently implemented by IBM, Hitachi, and EMC. This function is also commonly referred to as z/OS Global Mirror or XRC. A simplified example of the XRC architecture is shown in Figure 3-13 on page 35.

Customers who have z/OS and/or zLinux applications that require remote copy over long distances[10] with minimal impact to the primary application I/O activity should consider z/OS Global Mirror[11]. IBM's XRC implementation produces remote copies that are time-consistent and enable the recovery of databases at the remote site.

[9] z/OS Global Mirror (XRC) is not supported on the IBM DS6000.

[10] z/OS Global Mirror (XRC) has been tested to 24,000 miles.

[11] Currently, the overhead of asynchronous mirroring using DS8000/ESS800 z/OS Global Mirror (XRC) is approximately 0.04ms or 40 microseconds.

In an XRC configuration, the primary disk storage subsystem must be capable of running the XRC function. The target secondary disk subsystem might be any disk and need not be capable of running XRC.

If the installation also has the requirement for planned site switches, that is, mirror data from site 2 back to site 1 after having switched the affected applications from site 1 to site 2, in this case the secondary disk storage subsystem must also be capable of running XRC.

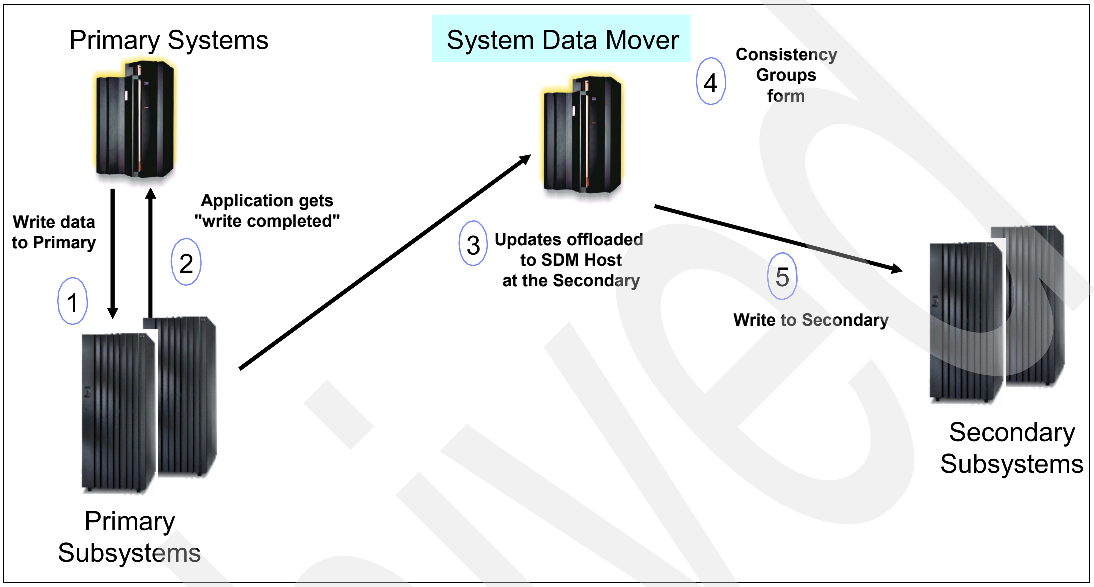


Figure 3-13 z/OS Global Mirror (XRC) architecture

The IBM DS8000 has lead the industry with many essential z/OS Global Mirror (XRC) improvements. Consequently, some customers have adopted the strategy of installing new DS8000 storage subsystems for their primary site and moving existing disk subsystems to the secondary site. In this way customers preserve their investment in their existing storage, while taking advantage of the latest improvements at their primary site.

Here are some of the principal benefits of IBM's z/OS Global Mirror (XRC).

- An IBM XRC primary volume can also exploit the performance capabilities of PAV, HyperPAVs, Multiple Allegiance, and I/O Priority Queuing. There are no restrictions that prevent XRC from operating concurrently with PAV or HyperPAVs for a given disk volume.
- The z/OS host has software that manages all cross-volume/cross-disk subsystem data consistency. The z/OS System Data Mover (SDM) software typically runs at the remote location, pulls the updates from the primary site disk storage subsystems and in turn writes the updates to the secondary storage.
- The z/OS host and the DS8000 subsystem make use of algorithms that enable them to cooperate to optimize data transfer, while minimizing the impact to application I/O activity and use of host processing cycles.
- SDM MIPS can now be offloaded to zIIP engines, when the SDM system is run on a z9/z10 zIIP-capable processor.
- XRC is designed to be highly scalable, thus yielding the low RPOs with little to no impact to host application I/O activity.
- GDPS/XRC and XRC provide full inter-operability with OEM disk storage subsystems.

Archived

IBM z/OS Global Mirror operation

A host application issues a write request to a disk on the primary subsystem that is part of a z/OS Global Mirror configuration. The DS8000 z/OS Global Mirror function intercepts the write request and captures all of the information required by the System Data Mover (SDM) to recreate the write operation at the secondary site. Information collected for the SDM includes:

- Type of write command [12]
- Sequence number
- Host timestamp [13]

All of this information is collected in about .04 ms after which the write command continues and is completed.

Independent of the application I/O running on the host at the primary, the SDM communicates with the DS8000 at the primary site. The SDM reads all record updates across all volumes in a LSS. The updates are then collected through the SDM across all primary storage subsystems, journalized in consistency groups and subsequently written to the secondary target disks. z/OS Global Mirror implementation details might vary by vendor.

IBM z/OS Global Mirror (XRC) performance

The architecture of the DS8000 z/OS Global Mirror (XRC) function has evolved through working with a number of customers who have had the requirement of little to no application impact by XRC. This has been achieved through several DS8000 capabilities:

- Device-level I/O pacing
- Fixed and floating utility devices
- Suspend without creating a Long Busy
- Multiple Reader Support
- Zero suspend FlashCopy
- Unplanned outage algorithms.

XRC has been successfully tested in configurations running in production across 20,000+ XRC volume pairs.

z/OS Global Mirror (XRC) is highly scalable. As of June 1, 2009, one customer is running 650+ TBs in a single XRC configuration at a telecommunications distance of greater than 1,900 miles. This customer has some 50+ coupled System Data Movers (SDMs) in a single GDPS/MzGM HyperSwap configuration.

Typically, a SDM can manage between 1,200–1,800 volumes and up to five SDMs can be active on a single z9® or z10 LPAR. Host MIPS usage of SDM with the DS8000 is in the range of 1–3 MIPS per 100 write SIOs. Distance is achieved using channel extension technology converting the FICON® I/O protocol to a telco protocol and then back to the FICON I/O protocol at the remote site. Encryption and/or compression can be provided through the channel extender when required.

Archived

[12] Examples of write commands are: Write Any, Format Write, Update Write, and Erase.

[13] To create the timestamp, the host system executes a Store Clock instruction into each write I/O chain. Additionally, all systems participating in the z/OS Global Mirror configuration are connected to a common time source, the SYSPLEX timer.

3.3.4 IBM Three-site Remote Replication functions

The following sections provide details on the IBM Three-site Remote Replication functions.

IBM DS8000 Metro Global Copy (MGC)

IBM Metro Global Copy (MGC) is a combination of Metro Mirror between two local or regional volumes and Global Copy (PPRC-XD) to a remote location. Examples are shown in Figure 3-14. Metro Global Copy (MGC) can be used with System z CKD data and distributed systems fixed-block data. Additionally, the data freeze command set is designed to provide point-in-time data consistency across any combination of System z and distributed system volumes/LUNs, helping to enable complete cross-volume and cross-storage subsystem data integrity and data consistency for business-critical application data.

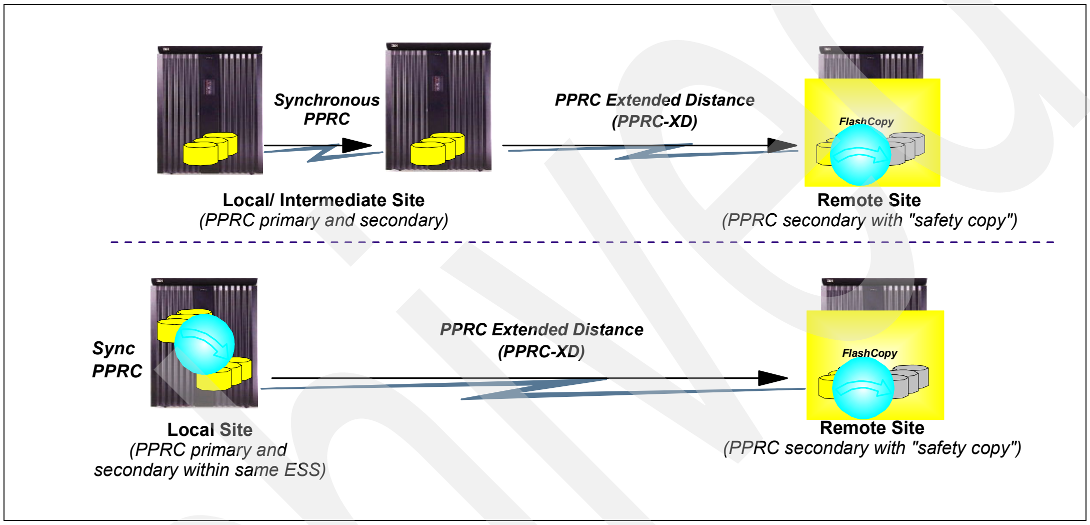


Figure 3-14 Two-site and three-site DS8000 Metro Global Copy

Metro Global Copy (MGC) operation

As shown in Figure 3-14, Metro Global Copy (MGC) might be configured as a two-site or a three-site operation. In both configurations, there are three, or optionally four, copies of an application's business-critical data.

- IBM Metro Mirror is used to propagate data from volume A to volume B. Volume B might be within the same disk storage subsystem eStorage, or a separate disk storage subsystem within the same site or an intermediate site.
- Volume B, which is the Metro Mirror secondary volume associated with volume A, also acts as the primary volume for Global Copy (PPRC-XD) operation between the intermediate site and volume C, which resides at the final remote site. Typically, the remote site is at some distance from the primary site.
- The communication bandwidth used by Global Copy (PPRC-XD) is highly scalable. It will use the bandwidth provided.

If volumes A and B are in two separate sites, in the event of a disaster where only site 1 or the intermediate site 1A is lost, but not both, no data is lost in this configuration. One might also add a volume D at the remote site, which can be used to create a point-in-time copy of the data. Volume D would be used for recovery in the event that both site 1 and the intermediate site 1A are lost in a regional disaster.

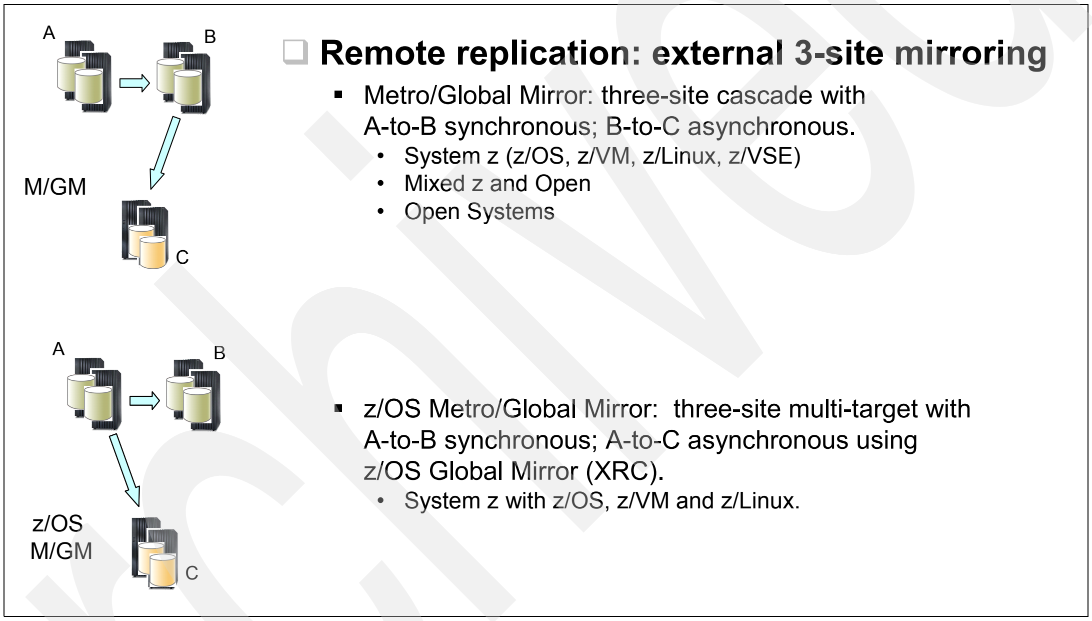
Archived

To form a hardened consistency group on the set of D volumes at site 2, a data freeze is invoked for the Metro Mirror operation between volume A and volume B, suspending the Metro Mirror A-to-B relationship with change control. Then one waits until all Global Copy (PPRC-XD) data in-flight synchronizes between volumes B and C. Once volumes B and C are in full duplex mode (in Sync), FlashCopy is used to copy volume C to D. When all volume C to D logical FlashCopy copies complete, a resynchronization can occur between volumes A and B.

Metro Global Copy (MGC) is managed by TPC-R. TPC-R is a Tier 6 IBM product designed to provide end-to-end management of the data mirroring solution.

IBM DS8100/DS8300 three-site "no-data-loss" solutions

Many customers have a requirement for no data loss at all in the event of a disaster. This requirement can be met using a three-site implementation strategy like one of the configurations shown in Figure 3-15.



❑ **Remote replication: external 3-site mirroring**

- Metro/Global Mirror: three-site cascade with A-to-B synchronous; B-to-C asynchronous.
  - System z (z/OS, z/VM, z/Linux, z/VSE)
  - Mixed z and Open
  - Open Systems
- z/OS Metro/Global Mirror: three-site multi-target with A-to-B synchronous; A-to-C asynchronous using z/OS Global Mirror (XRC).
  - System z with z/OS, z/VM and z/Linux.

Figure 3-15 External three-site mirroring with M/GM or z/OS M/GM

Metro Mirror supports zero data loss because it operates synchronously. However, most installations choose to accept some data loss so as to avoid any impact to applications running on the host at the primary site in the event of an operational error that is not actually a disaster, while at the same time requiring the minimum practical RESTART time in the event of a site switch. For this reason, many installations choose asynchronous mirroring. z/OS Global Mirror (XRC) typically will yield three to five seconds of data loss.

The key in any solution, however, is that cross-volume data integrity and data consistency must be maintained no matter what. As discussed at length above, IBM Metro Mirror, IBM Global Mirror, and IBM z/OS Global Mirror all maintain data consistency in the event of an unplanned outage.

Three-site System z/Open System solutions

For distributed systems, the DS8100/DS8300 Metro Global Mirror (M/GM) solution provides for no data loss for both distributed and System z systems with the assumption that only one site is lost in the event of a disaster. This solution will also provide a specific RPO (loss of data) if site 1 and site 1A are both lost in a regional disaster.

Archived

Three-site System z solutions

For z/OS customers that require the highest level of continuous operation across two local sites as well as having an out-of-region disaster recovery capability, a popular option being implemented by many customers is GDPS/XRC and GDPS/PPRC for mirroring combined with HyperSwap for automated failover/fallback, as shown in Figure 3-16 on page 40.

GDPS/XRC is an IBM product that is designed to provide a complete multi-site disaster recovery capability from a single point of control while managing a single logical session across multiple disk vendors' z/OS Global Mirror implementations. GDPS/XRC will mirror all of the production site volumes to an out-of-region site 3, while GDPS/PPRC with HyperSwap will perform real time synchronous data replication across two local sites or across two disk storage subsystems within the same data center floor[14].

Many customers are now moving to this combination of GDPS/XRC and GDPS/PPRC to provide continuous operations locally while maintaining disaster recovery protection outside the region. GDPS/XRC is a scalable, repeatable, proven, auditable business continuance solution running in many customer locations world wide. Additionally, since multiple disk storage vendors have implemented the IBM PPRC and/or XRC architectures, customers are also provided with the comfort of a multi-vendor disk storage subsystem strategy fully supported by GDPS.

IBM SVC and DS4000/DS5000 Global Mirror

The IBM SAN Volume Controller (SVC) and the DS4000/DS5000 disk subsystems also provide an asynchronous Global Mirror function which, similar to the DS8000 Global Mirror function described above, is an "A" to "B" solution. However, unlike the D8000 solution discussed above, this solution orders write I/O operations in the primary SVC prior to transmitting the data to the Target SVC. The same is true with the DS4000/DS5000.

When DS4000/DS5000 storage subsystems are used with SAN Volume Controllers, the SVCs control remote replication. The SVC Global Mirror function transmits data from one SVC to another, not from the disks under control of the SVCs at the primary and remote sites.

The SVC and DS4000/DS5000 disk subsystems do not participate in a Metro Global Mirror three-site configuration.

Rules of thumb for remote mirroring

As discussed above, one of the first issues that must be considered when planning to back up business-critical data using remote mirroring is the potential increase in the time it takes to complete a write operation. Synchronous remote mirroring is popular due to its relatively simple operational considerations and its ability to support zero data loss when transmitting data to the backup site. Synchronous remote mirroring is the only remote mirroring scheme which can avoid data loss.

However, when synchronous remote mirroring is employed, the time it takes for a write operation to complete is proportional to the distance between the primary and secondary storage subsystems[15]. Asynchronous remote mirroring avoids the problem of increased write I/O time, but introduces operational considerations that the IT installation must take into account when planning for long distance remote mirroring.

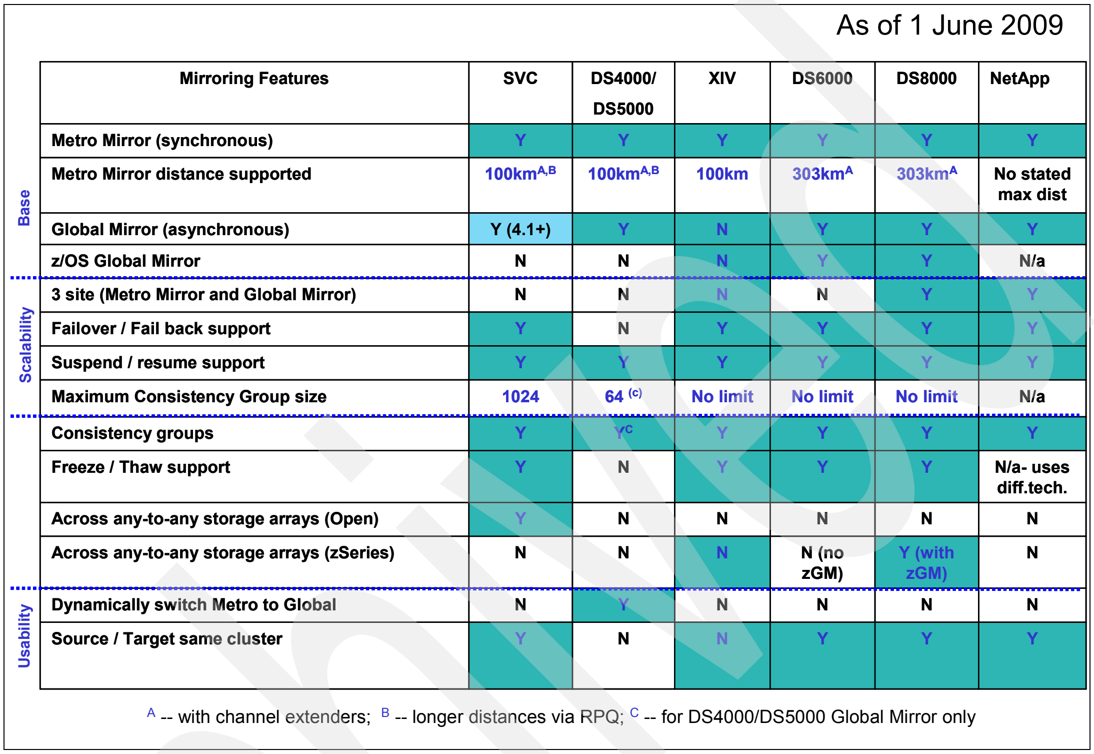
Archived

[14] A popular lower cost option to mask storage subsystem failures is to implement the GDPS/PPRC HyperSwap Manager discussed above in place of full-function GDPS HyperSwap.

[15] Electrical signals can only travel so fast; increasing the distance between the primary and secondary storage subsystems increases the time it takes signals to transit the distance between the subsystems. When synchronous remote mirroring is used, write operations are not considered complete until the primary has received notification from the secondary that all of the data for the write operation has been received successfully at the secondary.

If dark fibre is available between the sites, a rule would be to consider synchronous Metro Mirror first if the fibre distance between the boxes is less than 300 km and the application can tolerate the synchronous write delays. For distances over 300 km or if dark fibre is not available, then an asynchronous data replication solution is normally deployed.

The table in Figure 3-16 provides a summary of remote mirroring functions for IBM storage.



As of 1 June 2009

	Mirroring Features	SVC	DS4000/ DS5000	XIV	DS6000	DS8000	NetApp
Base	Metro Mirror (synchronous)	Y	Y	Y	Y	Y	Y
	Metro Mirror distance supported	100km^{A,B}	100km^{A,B}	100km	303km^{A}	303km^{A}	No stated max dist
	Global Mirror (asynchronous)	Y (4.1+)	Y	N	Y	Y	Y
	z/OS Global Mirror	N	N	N	Y	Y	N/a
Scalability	3 site (Metro Mirror and Global Mirror)	N	N	N	N	Y	Y
	Failover / Fail back support	Y	N	Y	Y	Y	Y
	Suspend / resume support	Y	Y	Y	Y	Y	Y
	Maximum Consistency Group size	1024	64 $^{(c)}$	No limit	No limit	No limit	N/a
	Consistency groups	Y	Y^{C}	Y	Y	Y	Y
	Freeze / Thaw support	Y	N	Y	Y	Y	N/a- uses diff.tech.
	Across any-to-any storage arrays (Open)	Y	N	N	N	N	N
	Across any-to-any storage arrays (zSeries)	N	N	N	N (no zGM)	Y (with zGM)	N
Usability	Dynamically switch Metro to Global	N	Y	N	N	N	N
	Source / Target same cluster	Y	N	N	Y	Y	Y

^{A} -- with channel extenders; ^{B} -- longer distances via RPQ; ^{C} -- for DS4000/DS5000 Global Mirror only

Figure 3-16 Remote Mirroring capabilities comparison

Archived

3.4 IBM geographically dispersed parallel SYSPLEX

In this section we discuss managing z/OS data replication and Open Systems data replication with GDPS.

3.4.1 Managing z/OS data replication with GDPS

As stated in 3.2.7, “Tier 7: IBM geographically dispersed parallel SYSPLEX (GDPS)” on page 20, the highest level in the multi-site availability hierarchy can be supported using GDPS. For a disaster recovery plan to achieve a proven, repeatable RTO, all of the resources must be managed by end-to-end automation. This includes all of the affected servers, workloads, networks, and the various data management components for both planned and unplanned site failover and fallback situations.

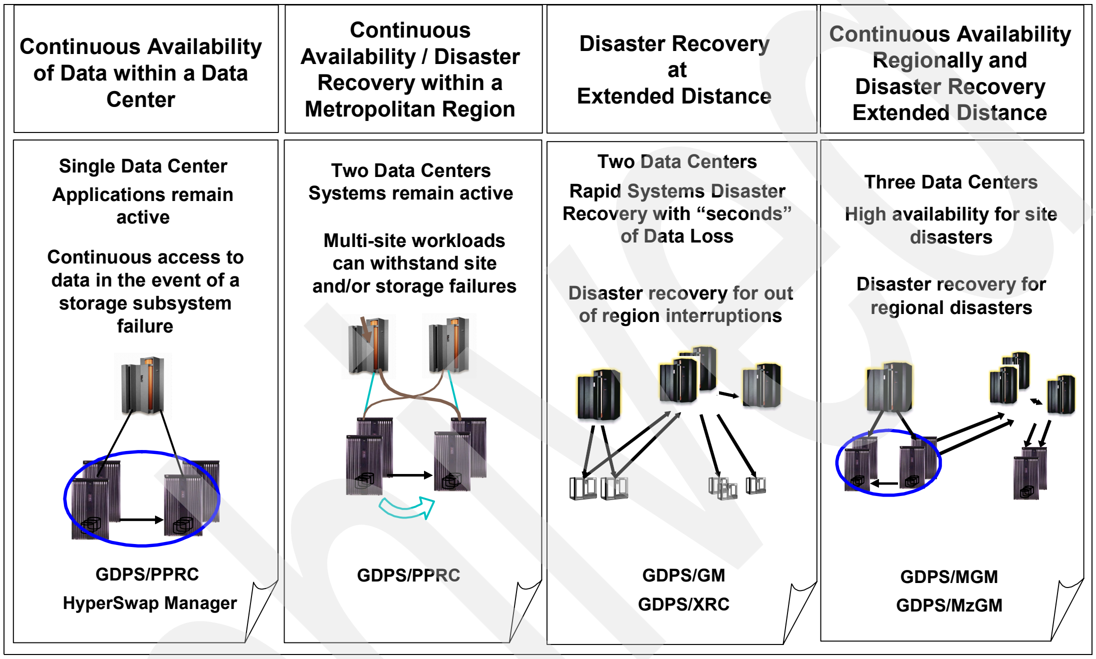
GDPS combines system automation software and hardware that are designed to work together to provide the means of managing and achieving a complete switch of all resources from one site to another automatically. This includes:

- Processors, CECs, LPARs, CF structures, Sysplex clocks
- Network resources
- Disk and tape data replication environments

Additionally, GDPS is designed to be used to manage all resources required for a set of applications in the event of a planned or unplanned requirement to:

- Switch the applications from one site to another, and
- Start the application environment as well as the applications at the remote site automatically.

Figure 3-17 provides an overview of the various System z business continuity IT infrastructure solutions that GDPS is designed to help enable for continuous availability, continuous operations, and disaster recovery.



Continuous Availability of Data within a Data Center	Continuous Availability / Disaster Recovery within a Metropolitan Region	Disaster Recovery at Extended Distance	Continuous Availability Regionally and Disaster Recovery Extended Distance
Single Data Center Applications remain active Continuous access to data in the event of a storage subsystem failure	Two Data Centers Systems remain active Multi-site workloads can withstand site and/or storage failures	Two Data Centers Rapid Systems Disaster Recovery with "seconds" of Data Loss Disaster recovery for out of region interruptions	Three Data Centers High availability for site disasters Disaster recovery for regional disasters
GDPS/PPRC HyperSwap Manager	GDPS/PPRC	GDPS/GM GDPS/XRC	GDPS/MGM GDPS/MzGM

Figure 3-17 GDPS System z Business Continuance Solutions

Managing Open Systems data replication with GDPS

IBM GDPS also provides inter-operability automation called Distributed Cluster Manager (DCM) for the Open Systems environment. DCM inter-operates with the following distributed cluster management products:

- Tivoli System Automation Application Manager (TSA Appman)
- Veritas Cluster Server (VCS) to provide an end to end Enterprise solution

Figure 3-18 on page 42 outlines these solutions. DCM is made available through an IBM Services offering called Geographically Dispersed Open Clusters (GDOC).

DCM can also inter-operate with the GDPS three-site solutions:

- GDPS/MzGM which uses z/OS Global Mirror
- GDPS/MGM which uses DS8000 Global Mirror

Storage-based data replication can be managed for the combination of GDPS and DCM when all of the data, System z CKD data and Distributed Systems Fixed Block data, resides on DS8000 subsystems. Management of data Open Systems LUNs is accomplished through an ECKD device address function that maps to the LUNs in question. Alternatively, the data replication associated with each Open Systems host can be managed within that Open System host environment.

Archived

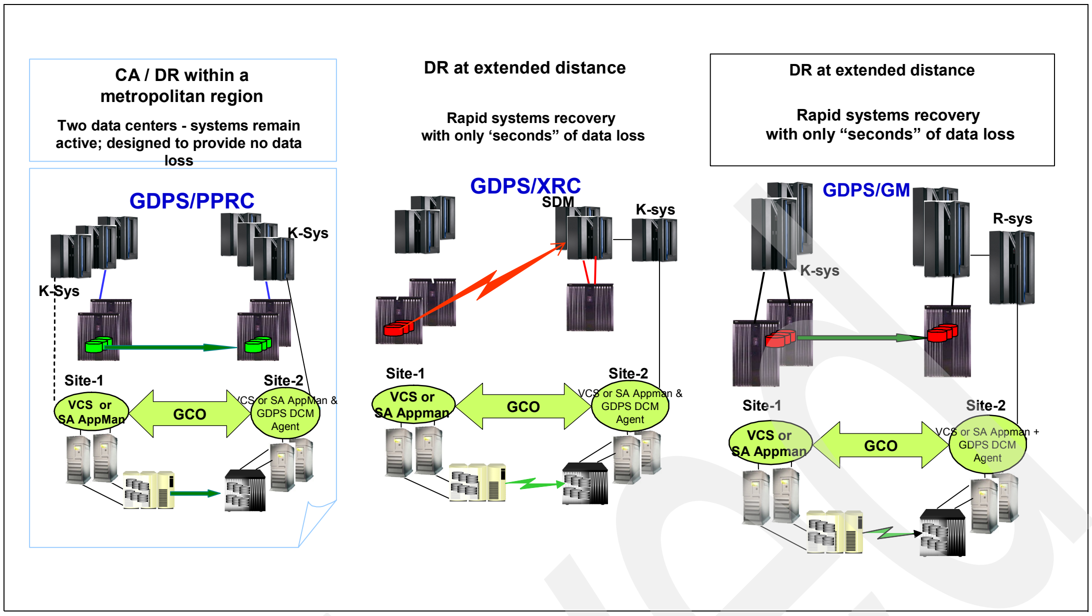

CA / DR within a metropolitan region
Two data centers - systems remain active; designed to provide no data loss
GDPS/PPRC
K-Sys
K-Sys
Site-1
Site-2
VCS or SA AppMan
GCO
VCS or SA AppMan & GDPS DCM Agent
DR at extended distance
Rapid systems recovery with only 'seconds" of data loss
GDPS/XRC
SDM
K-sys
Site-1
Site-2
VCS or SA Appman
GCO
VCS or SA Appman & GDPS DCM Agent
DR at extended distance
Rapid systems recovery with only "seconds" of data loss
GDPS/GM
K-sys
R-sys
Site-1
Site-2
VCS or SA Appman
GCO
VCS or SA Appman + GDPS DCM Agent

Figure 3-18 GDPS Distributed Cluster Management (DCM)

3.4.2 GDPS summary

GDPS automation provides the Tier 7 disaster recovery protection for all System z images as well as management of data on Open System LUNs, enabling an installation to recover both System z and associated Open System data to a consistent point in time. GDPS can scale from a single site with one CPU to three sites, helping implement BC/DR solutions that enable recovery in the event of a local or regional disaster.

GDPS is designed to help enhance the remote disk mirroring functions available with IBM's DS8000 disk storage subsystems. Installations operating remote disk mirroring have found that the time required to switch operations from one site to another without the help of automation can be greater than four hours. GDPS is designed to provide installations with a total multi-site resource management capability that can enable an installation to perform a site switch in as little as 30 minutes. In the event of an unplanned situation, GDPS, through its freeze triggers, is designed to confirm that all data at the secondary site is I/O-consistent to a single point in time, thereby giving the installation the ability to perform a database RESTART.

GDPS has been available since 1998. Since the initial GDPS/PPRC offering, a number of new offerings, features, and functions have been added addressing client requirements. As of 2009, IBM has over 500 GDPS implementations across clients in more than 34 countries. IBM is currently providing some of the highest levels of continuous operations and disaster protection available in the marketplace.

GDPS is a Netview/Systems Automation for z/OS 'application' and is made available as a GTS services product. For more information on the GDPS solutions refer to *The GDPS Family - An Introduction to Concepts and Capabilities*, SG24-6374.

Archived

3.5 Vendor independence

Proprietary vendor solutions have the potential to increase the cost of a multi-site configuration significantly as well as the overall cost of a disaster recovery solution. Typically two copies of data are required for storage mirroring whether the data resides on disk or tape. Selecting a single mirroring technique that is available from multiple vendors can provide a cost-effective situation when adding additional storage.

Several vendors have their own proprietary disk mirroring techniques, for example:

- EMC SRDF, SRDF/A
- Hitachi TrueCopy and Universal Replicator

In our view, the disadvantage of proprietary disk mirroring solutions in this situation is that all disk subsystems involved in mirroring generally must be from that vendor, including any storage required for future growth.

Further, if one considers a situation in which one must execute an unplanned site switch for a disaster and the subsequent disaster recovery considerations, including perhaps failing over to a third party business recovery house, being constrained by the requirement of having to use a specific storage subsystem can add to the total cost of a subsequent contract. Selecting a widely-used architecture can help maintain focus on competitive prices when adding additional storage.

3.5.1 Business continuance automation with synchronous mirroring

The IBM Metro Mirror (PPRC) architecture requires that both the primary and secondary disk subsystem be peers, but when combined in a GDPS/PPRC implementation, multiple pairs of disk subsystems from multiple vendors can exist side-by-side. The normal customer GDPS/PPRC environment is multi-vendor. The IBM Metro Mirror architecture is implemented by most major vendors including IBM, Hitachi, and EMC.

The GDPS/PPRC[16] and GDPS/PPRC HyperSwap Manager BC/DR automation functions provide the ability for many of the high-end disk vendors that have implemented IBM's Metro Mirror Architecture to work with z/OS, zLinux and z/VM® data. GDPS automation of the PPRC Architecture provides a system environment where each vendor's peer to peer storage subsystems running Metro Mirror can run side by side. All high-end vendors' synchronous remote copy Metro Mirror configurations can be placed in a GDPS/PPRC environment to provide a vendor-independent end-to-end solution with full cross-volume and cross-storage subsystem data integrity and data consistency. These vendors also have implemented the data freeze commands and triggers. This interoperability capability has became the defacto standard.

Many customers want to make new disk purchase decisions independent of previous disk storage subsystem purchases. The GDPS/PPRC and GDPS/PPRC HyperSwap Manager solutions are designed to promote multi-vendor inter-operability by supporting the defacto standard, that is, the IBM Metro Mirror architecture. It is important to note that each vendor's synchronous remote copy implementation requires that the primary and secondary be from the same vendor. For example, an IBM DS8000 primary can not be used with an EMC Symmetrix DMX secondary. However, because GDPS/PPRC and GDPS/PPRC HyperSwap Manager support multi-vendor environments, a database which spans, for example, an IBM and an EMC Symmetrix at the primary site can be replicated consistently to a corresponding

[16] EMC's remote copy function is called "Symmetrix Remote Data Facility" or SRDF. EMC has also implemented a synchronous remote copy function which supports GDPS, known as GDPS/SRDF. For purposes of this paper the solution from all vendors is referred to as: GDPS/Metro Mirror (PPRC).

Archived

and Symmetrix at the secondary through the use of GDPS/PPRC or GDPS/PPRC HyperSwap Manager.

Customers exploit this vendor independence to provide flexibility in their IT infrastructure. This flexibility enables installations to exploit their storage investments fully and depreciate their existing disk storage subsystems after which the next storage subsystem procurement can be made.

Business continuance automation with asynchronous mirroring

IBM's z/OS Global Mirror (XRC) architecture provides even more flexibility, as only the primary disk subsystem must be XRC-capable. The target storage subsystem at the secondary site does not have to be XRC-capable, but only a compatible disk image. IBM, Hitachi, and EMC support the z/OS Global Mirror (XRC) architecture. Storage from IBM, Hitachi, EMC, and Sun Microsystems can be used as z/OS Global Mirror (XRC) target subsystems.

3.6 Business continuity next steps

As the needs of 24x7x365 e-Business grows, so will the requirements for better, faster, and more flexible business continuity solutions. Here are some of the latest areas of research and development for business continuity solutions focused at improving the end-to-end backup process to provide data resiliency and maintain the quality of the data.

- Making local copies of data easier to create, i.e. improving FlashCopy usability and improving integration with the Image Copy and backup processes.
- Making creation of local copies less expensive to help reduce the impact of data corruption events and integrating CDP technology and methodology fully with the backup process. Traditionally, data corruption events have included events surrounding hardware, operating system software, middleware (file system, database), application software and operational scenarios. Increasing customer concern relative to malicious intent and computer virus' has spawned this new research/development focus.

An end-to-end focus on data corruption involves the major categories:

- Detection: Detecting when data gets corrupted. Solution areas include:
  - Running data health checks before data is backed up
  - New monitors to detect various hardware and software messages that might be indications of various corruption events
  - Application specific health checks
  - Virus scans run on a regular basis before data is backed up
  - Improved customer operations awareness and procedures
- Protection: Against data corruption events. Solution areas include:
  - More frequent backups via the use of technologies like point-in-time FlashCopies
  - CDP technology/methodology
  - Disk and Tape Subsystem Encryption
- Isolation: When data corruption occurs, understand what data became corrupted and how it happened, and isolate further data corruption once detected. Solution areas involve emerging on-demand technology that treats data as a service.

Archived

- Repair: Real time repair of corrupted data. The challenge here is how a company fixes the corrupted data in real time, while continuing to let transactions occur against the good data in the database or file system. Traditional, restore from backup tape and forward recovery of databases take too much time in an on-demand world, as the entire database is unavailable until database log forward recovery is completed.

Addressing the areas concerned with the various aspects of data resiliency will be an evolving area focused on understanding the many different ways data actually can get corrupted. Data corruption events are rare, but their impact, depending on the type of corruption can take hours to days to recover. The challenge is to develop the right set of cost-effective solutions.

3.7 Summary

This paper has presented some of the lessons learned as a consequence of the WTC disaster and offers guidelines to assist organizations with addressing the initial questions that should be answered as their requirements for higher availability and disaster recovery evolve over time.

In summary:

- What disaster recovery tier are you at today?
- What are your current RPO and RTO?
- What is your future requirement for RPO and RTO?
- What is the distance between your two sites?
- What host platform(s) do you use: z/OS, Open Systems or mixed data?
- How sensitive is your primary application's write I/O activity to various remote copy overheads?

The IBM DS8000 and DS6000 are uniquely positioned along with the IBM TS7700 Grid to support customers at all disaster recovery tiers, while enabling them to move from one tier to the next as business and technical requirements change.

The disk subsystem Metro Mirror architecture and the IBM FlashCopy architecture can be used in conjunction with z/OS Global Mirror or Global Mirror. These solutions can exploit other DS8000 advanced features and functions like PAVs, HyperPAVs, Multiple Allegiance and I/O Priority Queuing to support the IT infrastructure flexibility needed to meet the demands of today's e-Business. All of these features are, in turn, exploited by IBM's GDPS service offering to provide one of the industry's best multi-site resource management solutions available today.

Archived

Archived

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper.

IBM Redbooks

For information about ordering these publications, see "How to get Redbooks" on page 47. Note that some of the documents referenced here might be available in softcopy only.

- *IBM System Storage Business Continuity: Part 1 Planning Guide, SG24-6547*
- *IBM TotalStorage Enterprise Storage Server PPRC Extended Distance*, SG24-6568
- *Planning for IBM Remote Copy,* SG24-2595
- *Backup Recovery and Media Services for OS/400: A Practical Approach*, SG24-4840
- *Disaster Recovery Using HAGEO and GeoRM*, SG24-2018

Other publications

This publication is also relevant as further information sources:

- *Geographically Dispersed Parallel Sysplex: The Ultimate e-business Availability Solution,* GF22-5114

How to get Redbooks

You can search for, view, or download Redbooks, Redpapers, Technotes, draft publications and Additional materials, as well as order hardcopy Redbooks publications, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Archived

© Copyright IBM Corp. 2010. All rights reserved.

Archived

Index

Archived

© Copyright IBM Corp. 2010. All rights reserved.

Archived

Archived

IBM Storage Infrastructure for Business Continuity

Archived

Describes IBM storage replication and automation functions for each BC Tier

Describes the lessons learned from two recent major disasters

Provides disaster recovery metrics and IBM storage solutions

The importance of business continuity and disaster recovery remains at the forefront of thought for many executives and IT technical professionals. This IBM Redpaper describes the lessons learned from recent disasters and how IBM storage technology can help businesses address many of the issues related to protecting their storage infrastructures and business-critical IT applications.

Two principal disaster recovery metrics, Recovery Time Objective and Recovery Point Objective, are defined and, along with the associated cost tradeoffs, are discussed from the vantage point of various IBM storage technology solutions.

Two IBM Business Continuance/Disaster Recovery (BC/DR) automation solutions, known as GDPS/PPRC with HyperSwap and GDPS/PPRC HyperSwap Manager, are described and shown how they can help an installation move closer to attaining a goal of continuous operation. For z/OS installations operating two or more sites, in the event of a storage subsystem, host, network or communications facility failure, a switch to processing at an alternate site can be made in almost real time by using GDPS/PPRC with HyperSwap.

Additionally, many Clustered Open Systems that are integrated with IBM Remote Copy technology can be configured to switch to a second site in almost real time. In these situations, when a site switch is executed, applications that have been cloned at both sites can continue running with minimal impact to the user.

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

REDP-4605-00