DS8000 Global Mirror Best Practices

Nick Clayton

Alcides Bertazi

Bert Dufrasne

Peter Klee

Robert Tondini

In partnership with
IBM **Academy of Technology**

IBM

Redpaper

IBM

International Technical Support Organization

DS8000 Global Mirror Best Practices

March 2017

REDP-5246-00

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

First Edition (March 2017)

This edition applies to the IBM DS8000® series with DS8000 LMC 8.8.21.xx.xx (bundle version 88.21.xxx.xx).

This document was created or updated on October 16, 2018.

Contents

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

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 grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

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 jurisdictions 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 websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

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

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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 actual people or business enterprises 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. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

AIX®	IBM®	Resource Measurement Facility™
DB2®	IBM Spectrum™	RMF™
DS8000®	IBM Spectrum Control™	System i®
Easy Tier®	MVS™	System Storage®
Enterprise Storage Server®	NetView®	System z®
FlashCopy®	Parallel Sysplex®	Tivoli®
GDPS®	PowerHA®	WebSphere®
Geographically Dispersed Parallel	Redbooks®	z Systems®
Sysplex™	Redpaper™	z/OS®
HyperSwap®	Redbooks (logo) ®	

The following terms are trademarks of other companies:

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

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

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

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

Preface

This IBM® Redpaper™ publication reviews the architecture and operations of the IBM DS8000® Global Mirror function. The document looks at different aspects of the solution in terms of performance, infrastructure requirements, data integrity, business continuity, and impact on production.

Hints and tips are provided on how to best configure the overall Global Mirror environment, in terms of connectivity, storage configuration, and specific parameters tuning. The guidelines that are provided are in general related to performance, which ultimately ensures a better recovery point objective (RPO). Therefore, we encourage you to follow those guidelines.

Authors

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

Nick Clayton is an Executive IT Specialist and Solution Architect for DS8000 development, IBM Systems. His specific interests include storage tiering, replication technology, and business resilience. He is a member of the IBM GDPS® design team. He also works with clients on their storage strategy and advises them on their deployment of IBM storage technology. Nick is the author or co-author of many patents, articles, white papers, and IBM Redbooks® publications on IBM storage technology. He is a regular presenter at storage conferences. Nick graduated from Trinity College Cambridge in 1994 with a degree in Mathematics and has had previous roles both within and outside IBM in the areas of parallel sysplex, performance, and enterprise storage. He is a member of the IBM Academy of Technology.

Alcides Bertazi joined IBM Brazil in 1979 as a Technical Specialist working in the IBM Hardware Support Center. In 1993, he moved to the Customer Support Center to provide mainframe customers with technical support on IBM z/OS®, Operational System, and IBM DB2® products. In 1999, he moved to IBM System Technology Group under Americas Storage Advanced Technical Support department covering storage products around Latin America region. Since then, he has provided presales technical support, initially for the IBM Enterprise Storage Server® product, and later for the IBM DS8000 Storage family. Currently, he is supporting the storage systems sales team in Brazil focusing on complex storage deals and technical support for critical situation resolution.

Bert Dufrasne is an IBM Certified IT Specialist and Project Leader for IBM System Storage® disk products at the ITSO, San Jose Center. He has worked at IBM in various IT areas. He has written many IBM publications, and has developed and taught technical workshops. Before joining the ITSO, he worked for IBM Global Services as an Application Architect. He holds a Master's degree in Electrical Engineering.

Peter Klee is an IBM Professional Certified IT specialist in IBM Germany. He has many years of experience in Open Systems platforms, SAN networks, and high-end storage systems. He formerly worked for a large bank in Germany where he was responsible for the architecture and the implementation of the disk storage environment. He joined IBM in 2003, where he worked for Strategic Outsourcing. Since July 2004, he has worked for ATS IBM System Storage Europe. His main focus is copy services, disaster recovery, and storage architectures for IBM DS8000 in open systems environments.

Robert Tondini is a Certified Senior IT Specialist based in IBM Australia, providing storage technical support. He has 20 years of experience in the mainframe and storage environments. His areas of expertise include IBM high-end disk and tape storage subsystems and disaster recovery solutions. He has co-authored several IBM publications and workshops for IBM enterprise storage systems, Advanced Copy Services, and IBM Tivoli® Storage Productivity Center for Replication.

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

Theodore (TJ) Harris
Flavio Morais
Brian Sherman
Paul Spagnolo
Gail Spear
Warren Stanley
Alexander Warmuth
IBM

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

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 email to:

   redbooks@us.ibm.com

► Mail your comments to:

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

Stay connected to IBM Redbooks

- ► Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- ► Follow us on Twitter:

  http://twitter.com/ibmredbooks

- ► Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- ► Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- ► Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Global Mirror overview and architecture

This chapter provides a general overview of the DS8000 Global Mirror function, its architecture, and details of the replication process.

This chapter includes the following sections:

► Global Mirror overview
► Setting up a Global Mirror session
► Forming consistency groups

1.1 Global Mirror overview

DS8000 Global Mirror is a two-site, unlimited distance data replication solution for both z Systems and Open Systems data.

When you replicate data over long distances, usually beyond 300 km, asynchronous data replication is the preferred approach. With asynchronous replication, the host at the local site receives acknowledgment of a successful write from the local storage instantly. The local storage system then sends the data to the remote storage later. Thus, replication is said to be done asynchronously.

In an asynchronous data replication environment, an application write I/O has the following steps:

1. Write application data to the primary storage system cache.

2. Acknowledge a successful I/O to the application so that the next I/O can be immediately scheduled.

3. Replicate the data from the primary storage system cache to the auxiliary storage system.

4. Acknowledge to the primary storage system that data has successfully arrived at the auxiliary storage system.

In an asynchronous type technique, the data transmission and the I/O completion acknowledge are independent processes, which results in virtually no application I/O impact.

When application data is spread across several volumes and eventually across multiple storage systems, asynchronous replication means that the data at the remote site does not necessarily represent the same order of writes as on the local storage. With asynchronous data replication techniques, special means are required to ensure data consistency for dependent writes at the secondary location.

To ensure consistency, Global Mirror uses the concept of *consistency groups*. With DS8000 Global Mirror data consistency is provided periodically by the following sequence:

1. The host I/Os to the primary volumes are suspended periodically for a short time.

2. This frozen data, which represents a *consistency group*, is transmitted to the remote site.

3. Consistent data is saved at the remote site.

The tradeoff of providing consistency in an asynchronous replication is that not all the most recent data can be saved in a consistency group. The reason is that data consistency can only be provided in distinct periods of time. When an incident occurs, only the data from the previous point of consistency creation can be restored. The measurement of the amount of data that is lost in such a case is called the recovery point objective (RPO), which is expressed in units of time, usually seconds or minutes.

The RPO is not a fixed number. It depends on the available bandwidth and the quality of the physical data link, and on the current workload at the local site. How to scale the Global Mirror and how to deal with the RPO is described later.

Global Mirror is based on an efficient combination of Global Copy and IBM FlashCopy® functions. It is the storage system microcode that provides, from the user perspective, a transparent and autonomic mechanism to intelligently use Global Copy with certain FlashCopy operations to attain consistent data at the secondary site.

To accomplish the necessary activities with a minimal impact on the application write I/O, Global Mirror introduces a smart bitmap approach in the primary storage system. Global Mirror uses two different types of bitmaps:

- The *Out-Of-Sync* (OOS) bitmaps that are used by the Global Copy function
- The *Change recording CR* bitmap that is allocated during the process of consistency formation

Figure 1-1 identifies the following essential components of the DS8000 Global Mirror architecture:

- Global Copy, which is used to transmit the data asynchronously from the primary (local) volumes $H1$ to the secondary (remote) volumes $H2$.

- A FlashCopy relation from the Global Copy secondary volumes $H2$ to the FlashCopy target volumes Jx.

- A Change Recording (CR) bitmap that is maintained by the Global Mirror process running on the primary storage system while the consistency group is created at the primary site.

  When the Global Mirror Process at the primary site creates a consistency group, the primary volumes $H1$ are frozen for the time it takes to allocate a new CR bitmap in the primary storage memory. All new data that is sent from the hosts is marked for each corresponding track[1] in the CR bitmap.

The newly formed consistency group is represented in the Global Copy named OOS *bitmaps*. When Global Copy sends the consistency group to the remote site, each transmitted track is checked against the OOS bitmap.

Figure 1-1 shows the architecture of DS8000 Mirror.
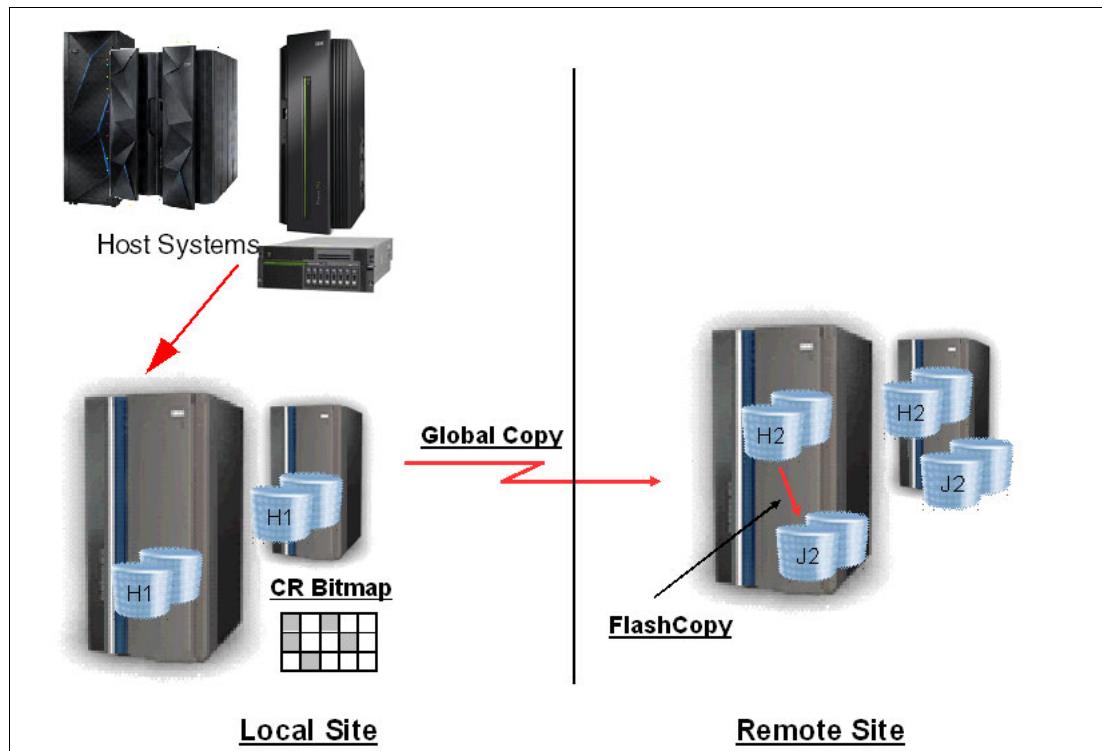
Figure 1-1   General architecture of DS8000 Global Mirror

---

[1] A track size can be 64 KB in Open Systems or around 57 KB in a mainframe architecture.

When the whole consistency group has been transmitted to the remote site, it must be saved to the Jx volumes by using the FlashCopy function. Doing so ensures that there is always a consistent image of the primary data at the secondary location.

1.1.1 Communication between primary disk systems

A Global Mirror session is a collection of volumes that are managed together when you create consistent copies of data volumes. This set of volumes can be in one or more logical storage subsystems (LSSs) and one or more storage disk systems at the primary site. Open Systems volumes and z/OS volumes can both be members of the same session.

> **Tip:** Global Mirror allows for the creation of multiple sessions on a primary disk system with different devices associated with each session. Defining separate sessions for different application hosts allows you to not only specify a different RPO for each application based on their business criticality, but also to leave specific host volumes in one session unaffected when another Global Mirror session must be failed over.

Figure 1-2 shows such a Global Mirror structure. A master coordinates all efforts within a Global Mirror environment. After the master is started and manages a Global Mirror environment, the master issues all related commands over Peer-to-Peer Remote Copy (PPRC) links to its attached subordinates at the primary site. The subordinates use inband communication to communicate with their related auxiliary storage systems at the remote site. The master also receives all acknowledgements from the subordinates, and coordinates and serializes all the activities in the Global Mirror session.
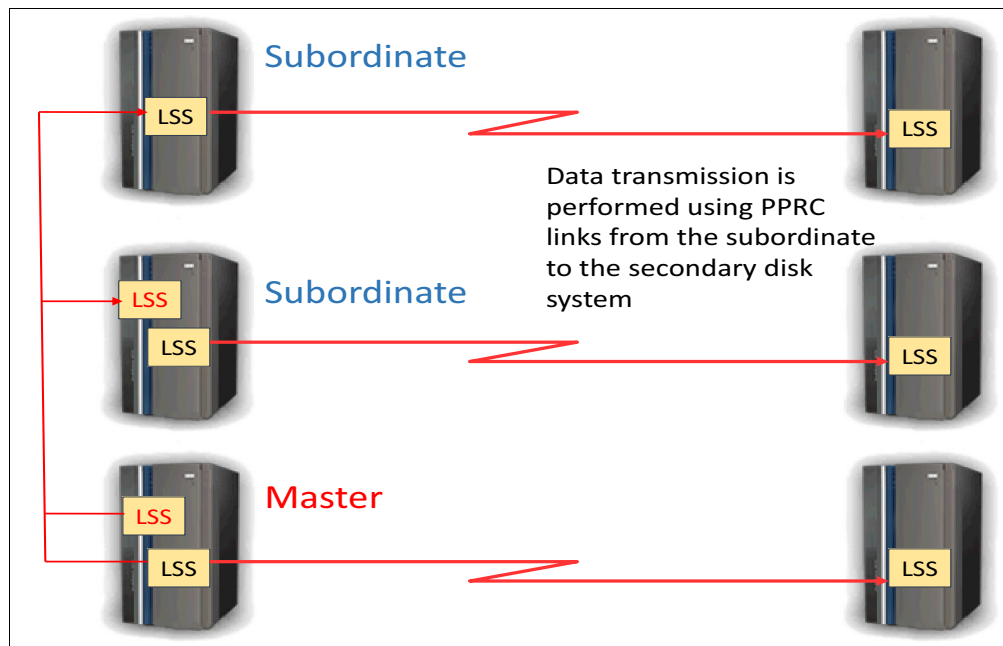
Figure 1-2   Global Mirror as a distributed application

With two or more storage systems at the primary site, which participate in a Global Mirror session, the subordinate is external and requires separate attention when you create and manage a Global Mirror session or environment.

To form consistency groups across multiple disk systems, the different disk systems must be able to communicate. This communication path must be resilient and high performance so that it causes minimal effect to the production applications. To provide this path, Fibre Channel links that use FCP protocol are used, which can be direct connections or more typically over a SAN.

> **Communication links:** Although just one path between a master LSS and one LSS in the subordinate is required, having two redundant paths is preferred for resiliency.

1.2 Setting up a Global Mirror session

To understand how Global Mirror works, this section explains how a Global Mirror environment (a Global Mirror session) is created and started. This approach is a step-by-step one and helps you understand the Global Mirror operational aspects.

1.2.1 A simple configuration

To understand each step and to show the principles, start with a simple application environment where a host makes write I/Os to a single application volume (A) as shown in Figure 1-3.
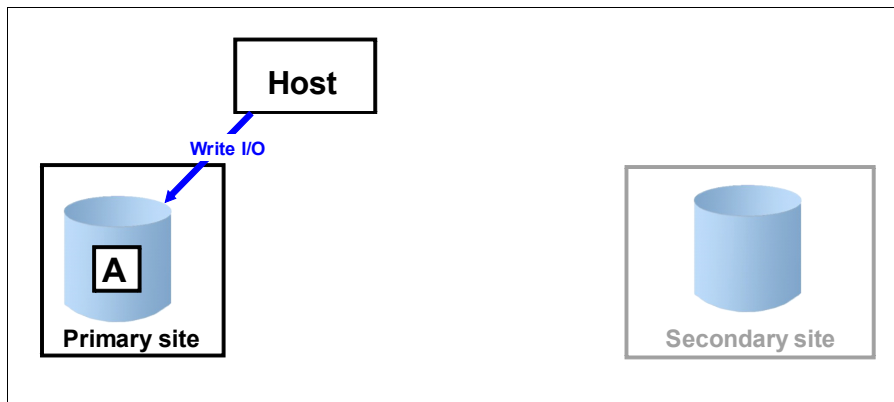
Figure 1-3   Start with a simple application environment

1.2.2 Establishing connectivity to a secondary site (PPRC paths)

Now, add a distant secondary site that has a storage system (B), and interconnect both sites (see Figure 1-4).
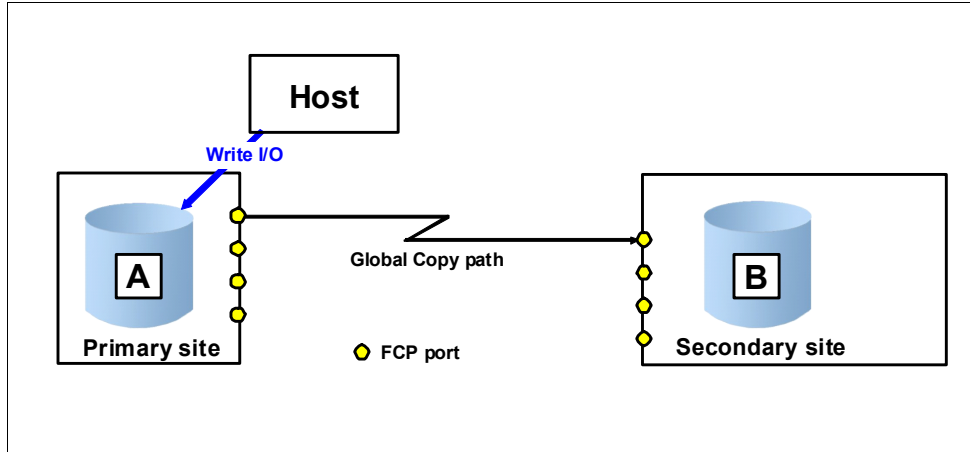
Figure 1-4   Establish Global Copy connectivity between both sites

Figure 1-4 shows how to establish Global Copy paths. Global Copy paths are logical connections that are defined over the physical links that interconnect both sites.

1.2.3 Creating a Global Copy relationship

Next, create a Global Copy relationship (establishing Global Copy pairs) between the primary volume and the secondary volume (see Figure 1-5).
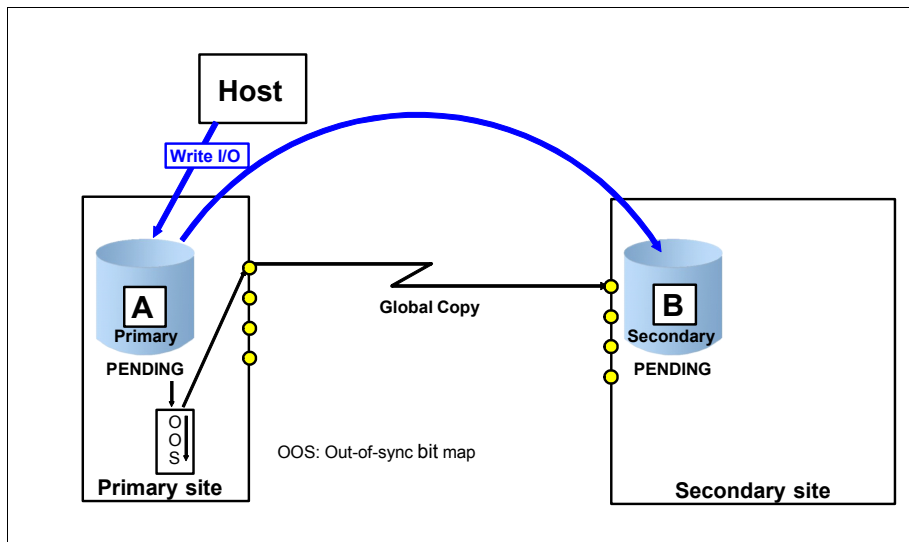
Figure 1-5   Establish a Global Copy volume pair

Creating the Global Copy relationship changes the target volume state from simplex (no relationship) to target Copy Pending. This Copy Pending state applies to both volumes: Primary Copy Pending and secondary Copy Pending.

Data is copied from the primary volume to the secondary volume. An OOS bitmap is created for the primary volume that tracks changed data as it arrives from the applications to the primary disk system. After a first complete pass through the entire A volume, Global Copy scans constantly through the OOS bitmap and replicates the data from the A volume to the B volume based on this out-of-sync bitmap.

Global Copy does not immediately copy the data as it arrives on the A volume. Instead, this process is an asynchronous one. When a track is changed by an application write I/O, it is reflected in the out-of-sync bitmap with all the other changed tracks. Several concurrent replication processes can work through this bitmap, which maximizes the usage of the high-bandwidth Fibre Channel links.

This replication process keeps running until the Global Copy volume pair A-B is explicitly or implicitly suspended or terminated.

Data consistency does not yet exist at the secondary site.

1.2.4 Introducing FlashCopy

FlashCopy is a part of the Global Mirror solution. Establishing FlashCopy pairs is the next step in establishing a Global Mirror session (see Figure 1-6).
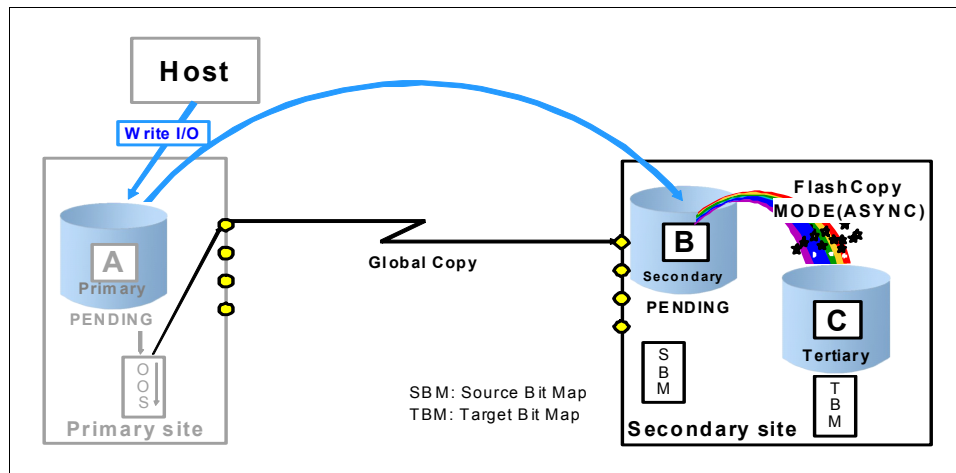
Figure 1-6   Introduce FlashCopy in to the Global Mirror solution

Figure 1-6 shows a FlashCopy relationship with a Global Copy secondary volume as the FlashCopy source volume. Volume B is now both a Global Copy secondary volume and a FlashCopy source volume at the same time. In the same storage server is the corresponding FlashCopy target volume.

This FlashCopy relationship has certain attributes that are required when you create a Global Mirror session:

► Inhibit target write: Protect the FlashCopy target volume from being modified by anything other than Global-Mirror-related actions.

► Start change recording: Apply changes only from the source volume to the target volume that occur to the source volume between FlashCopy establish operations, except for the first time that FlashCopy is established.

► Persist: Keep the FlashCopy relationship until explicitly or implicitly terminated. This parameter is automatic because of the change recording property.

► Nocopy: Do not start background copy from source to target, but keep the set of FlashCopy bitmaps (source bitmap and target bitmap) required for tracking the source and target volumes. These bitmaps are established when a FlashCopy relationship is created. Before a track in the source volume B is modified, between consistency group creations, the track is copied to the target volume C to preserve the previous point-in-time copy. This copy includes updates to the corresponding bitmaps to reflect the new location of the track that belongs to the point-in-time copy. The first Global Copy write to its secondary volume track with the window of two adjacent consistency groups causes FlashCopy to perform copy on write operations.

Some interfaces, such as Copy Services Manager, IBM Geographically Dispersed Parallel Sysplex™ (GDPS), and TSO, have these FlashCopy parameters embedded in their copy-services-related commands.

1.2.5 Defining a Global Mirror session

Creating a Global Mirror session does not involve any volume within the primary or secondary sites. The focus is on the primary site (see Figure 1-7).
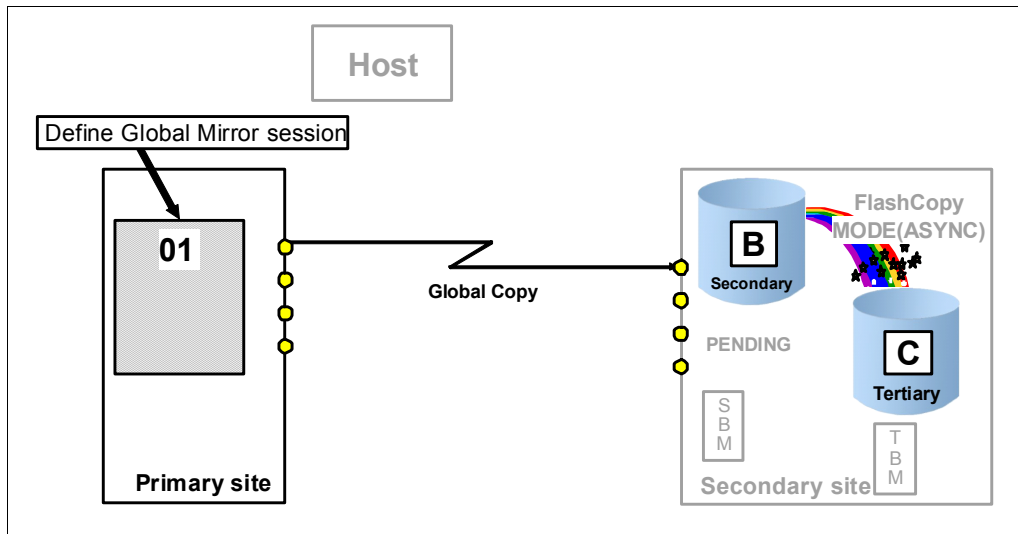
Figure 1-7   Define a Global Mirror session

Defining a Global Mirror session creates a token, which is a number between 1 and 255. This number represents the Global Mirror session.

This session number is defined at the LSS level. Each LSS that has volumes that are part of the session needs a corresponding session defined. Up to 32 Global Mirror hardware sessions can be supported within the same primary DS8000. *Session* means a hardware/firmware-based session in the DS8000, which is managed by the DS8000 in an autonomic fashion.

1.2.6 Populating a Global Mirror session with volumes

The next step is the definition of volumes in the Global Mirror session. The focus is still on the primary site (see Figure 1-8). Only Global Copy primary volumes are meaningful candidates to become members of a Global Mirror session.

This process adds primary volumes to a list of volumes in the Global Mirror session. However, it does not perform consistency group formation yet because the Global Mirror session has not started yet. When more volumes are added to the session, they are initially placed in a Join Pending state until they have performed their initial copy. After this process has completed, they join the session when the next consistency group is formed.

Adding a disk system to a Global Mirror session follows the same process except that a brief pause/resume of the consistency group formation process must be performed. This process can take a few seconds and is done to define the new disk system and its control paths to the master process. This process has minimal impact to the environment and only results in a small increase to the RPO for a brief period.
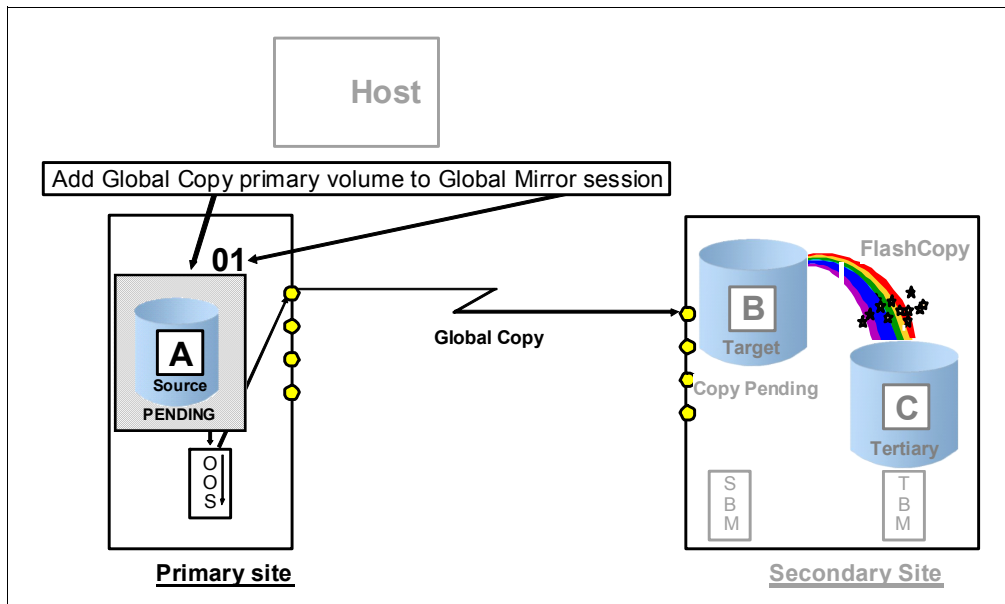
Figure 1-8   Add a Global Copy primary volume to a Global Mirror session

1.2.7 Starting a Global Mirror session

Global Mirror forms consistency groups at the secondary site. As Figure 1-9 indicates, the focus here is on the primary site, with the **start** command issued to an LSS in the primary storage system. With this **start** command, you set the master storage system and the master LSS. From now on, session-related commands must go through this master LSS.
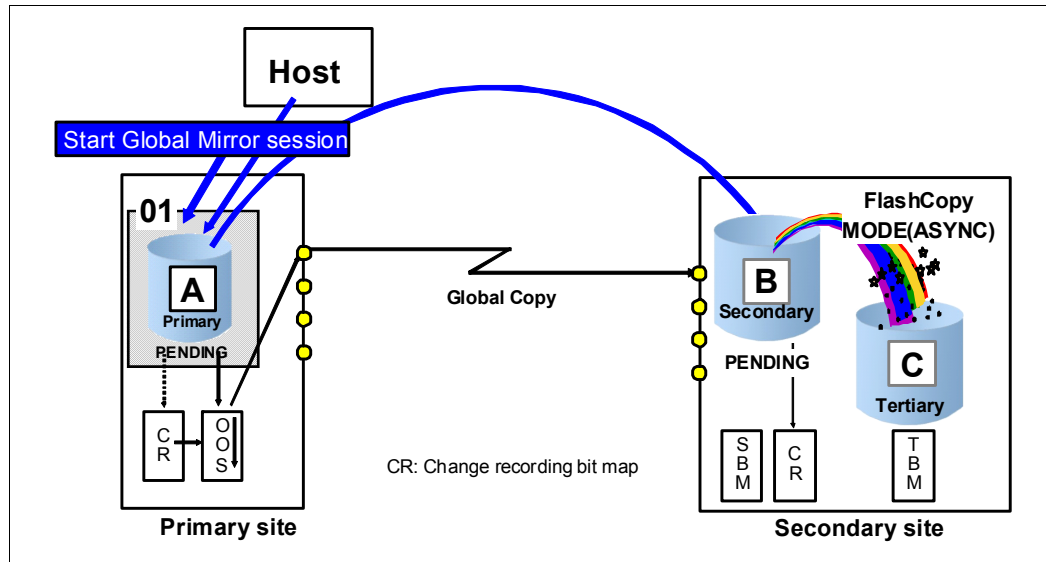
Figure 1-9   Start Global Mirror

This **start** command triggers events that involve all the volumes within the session. These events include fast bitmap management on the primary storage system, issuing inband FlashCopy commands from the primary site to the secondary site, and verifying that the corresponding FlashCopy operations finished successfully. This process happens at the microcode level of the related storage systems that are part of the session, and is fully transparent and autonomic from the user's perspective.

All B and C volumes that belong to the Global Mirror session comprise the consistency group.

1.3 Forming consistency groups

As described in 1.1, "Global Mirror overview" on page 2, the process for forming consistency groups can be broken down into these steps:

1. Create consistency group on primary disk system
2. Send consistency group to secondary disk system
3. Save consistency group on secondary disk system

1.3.1 The different phases of consistency formation

The numbers in Figure 1-10 illustrate the sequence of the events that are involved in the creation of a consistency group. This illustration provides only a high-level view that is sufficient to understand how this process works.
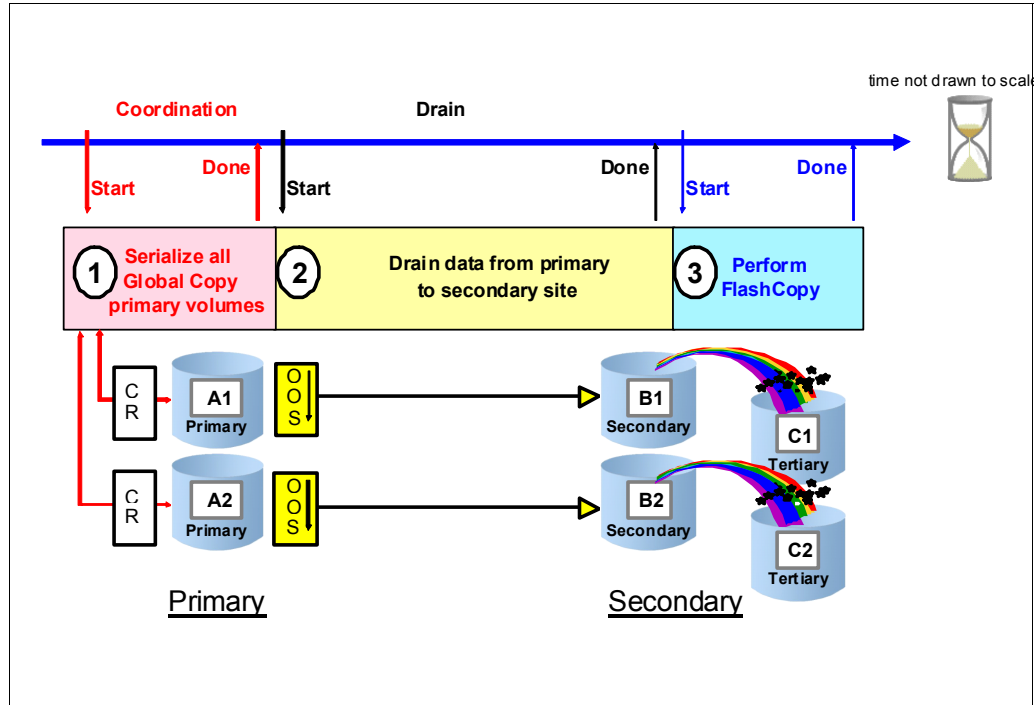
Figure 1-10   Formation of a consistent set of volumes at the secondary site

Note that before step 1 and after step 3, Global Copy constantly scans through the OOS bitmaps and replicates data from A volumes to B volumes, as described in 1.2.3, "Creating a Global Copy relationship" on page 6.

Consistency group formation

When the creation of a consistency group is triggered by the Global Mirror master, the following steps occur:

1. Coordination phase

   All Global Copy primary volumes are serialized. This serialization imposes a brief hold (freeze) on all incoming write I/Os to all involved Global Copy primary volumes. After all primary volumes are serialized across all involved primary DS8000s, the freeze on the incoming write I/Os is released. All further write I/Os are now noted in the CR bitmap of each volume. They are not replicated until step 3 on page 12 (Perform FlashCopy) is completed, but application write I/Os can immediately continue.

   Maximum coordination time: The maximum coordination time can be modified when the Global Mirror is established. The default value is 50 milliseconds. The maximum value is 65535 milliseconds. In some situations with a master/subordinate configuration, you might need to modify the default value.

The default for the maximum co-ordination time of 50 ms is a small value compared to other I/O timeout values like missing interrupt handler for devices in mainframe environment or SCSI I/O timeouts on Distributed platforms. Therefore, even in error situations when this timeout would be triggered, Global Mirror protects production performance rather than affecting the primary site in an attempt to form consistency groups in a time when error recovery or other problems are occurring.

2. Draining phase

In the draining phase, all tracks that were noted in the OOS bitmaps are transmitted to the remote site by using the Global Copy function. After all out-of-sync bitmaps have been processed, step 3, Perform FlashCopy, is triggered by the Master storage system microcode at the primary site.

> **Maximum drain time:** The maximum drain time is the maximum amount of time that Global Mirror will spend draining all data still at the primary site and belonging to a consistency group before failing that consistency group formation.
>
> The *maximum drain time* can also be modified when the Global Mirror is established. The default value is 30 seconds before microcode release 8.1. Starting with microcode release 8.1, the default value was changed to 240 seconds. The default value works well for most of implementations, so leave this parameter at the default.

If the maximum drain time is exceeded, then Global Mirror changes to Global Copy mode for a period to catch up in the most efficient manner. While in Global Copy mode, the overhead is lower than continually trying and failing to create consistency groups.

The previous consistency group will still be available on the C devices, so the effect of this process is simply that the RPO increases for a short period. The primary disk system evaluates when it is possible to continue to form consistency groups and restarts consistency group formation then.

The default for the maximum drain time allows a reasonable time to send a consistency group while ensuring that if some non-fatal network or communications issue occurs, then the system does not wait too long before evaluating the situation and potentially dropping into Global Copy mode until the situation is resolved. In this way, production performance is protected rather than attempting (and possibly failing) to form consistency groups at a time when this process might not be appropriate.

In situations with reduced bandwidth, it might be indicated to adapt this value.

If the system is unable to form consistency groups for 30 minutes, by default Global Mirror forms a consistency group without regard to the maximum drain time. It is possible to change this time if this behavior is not desirable for your particular environment.

3. Perform FlashCopy

Now the B volumes contain all data as a quasi point-in-time copy and are consistent. A FlashCopy is triggered by the primary system's microcode as an inband FlashCopy command. This FlashCopy is a two-phase process:

a. FlashCopy operation

First, the FlashCopy command is issued to all involved FlashCopy pairs in the Global Mirror session. A FlashCopy is done for each pair relation, individually grouped per LSS with each B volume, as FlashCopy source, and each C volume as a FlashCopy target volume. A consistency group is considered to be successfully created when all FlashCopy operations are successfully completed.

Before the copy operation starts, an internal attribute named *revertible bit* is set for each FlashCopy source volume. With this bit, the FlashCopy operation is protected

against updates from the local site. In this phase, the data is transmitted from the Global copy target volumes to the FlashCopy target volumes. When the transmission is completed, the FlashCopy sequence number of the pair is increased. The sequence number can be obtained for each pair with the DSCLI command `lspprc -l`.

b. FlashCopy termination

In this phase, the FlashCopy operation is finalized by terminating the copy process and increasing the FlashCopy sequence number. When all FlashCopy operations have completed, the revertible bits of the FlashCopy primary volumes are reset and the whole FlashCopy operation is committed. Finally, the revertible bit of the FlashCopy pair relation is reset.

When the FlashCopy is complete, a consistent set of volumes is created at the secondary site. This set of volumes, the B and C volumes, represents the consistency group.

For this brief moment only, the B volumes and the C volumes are equal in their content. Immediately after the FlashCopy process is logically complete, the primary systems' microcode is notified to continue with the Global Copy process from A to B. To replicate the changes to the A volumes that occurred during the step 1 to step 3 window, the change recording bitmap is mapped against the empty out-of-sync bitmap. From now on, all arriving write I/Os end up again in the out-of-sync bitmap. The conventional Global Copy process, as outlined in 1.2.3, "Creating a Global Copy relationship" on page 6, continues until the next consistency group creation process is started.

Consistency group interval time

The time between consistency groups formation is called the consistency group interval time. It is possible to specify a time period when the data transmission is continued in normal Global Copy mode, without forming consistency groups. As a default, the consistency group interval time is set to zero, which means the Global Mirror will form the next consistency group immediately. Valid values are round numbers between 0 and 65,535 seconds.

Figure 1-11 shows a more detailed breakdown of the consistency group formation process.
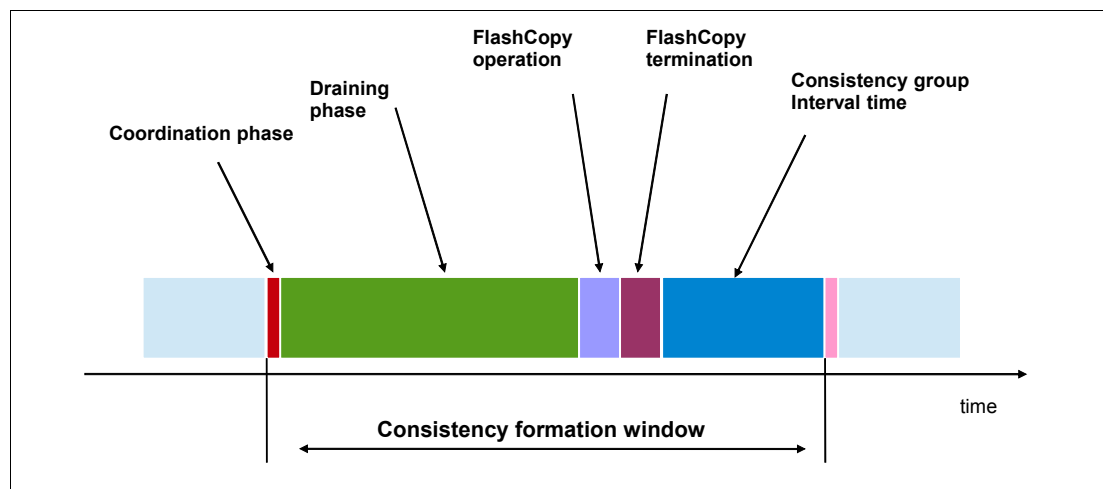

Figure 1-11   The consistency formation process

Important: Although the coordination time, the maximum drain time, and the consistency group interval time are tunable parameters, do not modify the default values before performing a careful analysis of the current behavior of the Global Mirror.

1.3.2 Impact of the consistency formation process

One of the key design objectives for Global Mirror is not to affect the production applications. The consistency group formation process involves the holding of production write activity to create dependant write consistency across multiple devices and multiple disk systems.

This process must be fast enough that an impact is negligible and is not perceived by applications. With Global Mirror, the process of forming a consistency group is designed to take 1 - 3 milliseconds. If you form consistency groups every 1 - 3 seconds, then the percentage of production writes impacted and the degree of impact is very small.

The example below shows the type of impact that might be seen from consistency group formation in a Global Mirror environment. Assume the following key data for a high performing storage system:

IO = 100,000 [IO/s]	A high load value in many environments
R/W ratio = 1:3	Typical read/write ratio
IO_{write} = 25,000 [IO/s]	Resulting write operations
RT_{min}= 0.2 [ms]	Assumed to be the DS8880 response time
$CGtime_{min}$ = 1 [ms]	Minimum time to form a consistency group at the GM primary site
$CGform_{min}$ = 3 [1/s]	The fastest time for Global Mirror to form consistency, by assuming unlimited bandwidth to the remote site
CGinterval = 0 [s]	Assume that you form consistency groups as fast as possible

First, calculate the number of write operations that might be affected during the consistency group coordination time:

$$N_{writes} = IO_{write} \times (RT_{min} + CGtime_{min})$$

Note: In reality, not all writes experience a delay during the freeze. To account for the worst possible case, add the storage system response time.

With the specific data, the number of impacted writes can be calculated as follows:

$$25,000 \text{ IOwrites/s} \times (0.0002 \text{ s} + 0.001 \text{ s}) = \textbf{30 } IO_{\textbf{writes}}$$

These 30 write operations can occur during the freeze of the primary volumes are delayed by the time it takes to create the consistency group, which is $CGtime_{min}$ = 1 ms. In the worst case, the consistency formation happens every CGformmax = 3 s.

$$CG_{delay} = N_{writes} \textbf{ / IOwrites} \times CGtime_{min} \textbf{ / CGform}_{min}$$

which is:

$$30 \text{ IOwrites} / 25,000 \text{ IOwrites/s} \times 0.001s / 3s = 0.000\ 000\ 04 \text{ s} = \textbf{0.0004 ms}$$

So, taking the host response time of RT_{min}= 0.2 ms, the average delay of 0.0004 ms due to consistency formation would cause 0.2% increase in response time.

As a conclusion, the impact is insignificant for normal performance monitoring tools.

1.3.3 Collisions

Collisions are situations where a host write operation to a certain track is part of the current consistency group that has not yet been sent completely to the remote site. To resolve this situation, the old track, which is presumably still in flight, is stored in an internal sidefile of the DS8000 storage system. This feature is referred to as *collision avoidance*. When this track is saved in the sidefile, the new track can be written as normal.

For the transmission of the old track, the source of this track is linked to the side file. This configuration ensures that the current consistency is completed in order. The new track will be transmitted with the next consistency group.

2

Planning and implementation

This chapter presents considerations and recommendations when planning for Global Mirror implementation. Topics involving sizing tools, primary and secondary disk system physical and logical configurations, replication at longer distances, volume placement, and EasyTier along with recommendations.

This chapter includes the following sections:

- ► Sizing tools for Global Mirror
- ► Global Mirror implementation planning
- ► Configuration guidelines for primary and secondary

17

2.1 Sizing tools for Global Mirror

One of the main Global Mirror objectives is the capability to achieve a recovery point objective (RPO) of 3 to 5 seconds with sufficient network bandwidth and resources. Moreover, it is important to achieve low RPO values without affecting the production applications that run on the Global Mirror primary disk system.

Figure 2-1 is a simplified Global Mirror diagram that indicates some sizing aspects and requirements. To achieve the required RPO, you need to understand the production workload. At least 24 hours, ideally several days of performance data are required to get an accurate understanding of the workload pattern. In particular, look for MBps write statistics.

Another important task is to model primary and secondary disk systems based on the workload profile. Disk Magic is the preferred tool to size disk systems based on workload profile. This tool can estimate disk systems behavior by creating different models with different combinations of DS8000 internal resources. Workload growth projections can also be done to understand internal storage system limits such as processors, host adapters, ports, and physical disks, among others.

Besides regular modeling tasks, Global Mirror secondary disk systems might require more evaluation due to Global Mirror journaling overhead. This overhead is FlashCopy activity that is going on along with production application write streams coming from the primary disk systems.
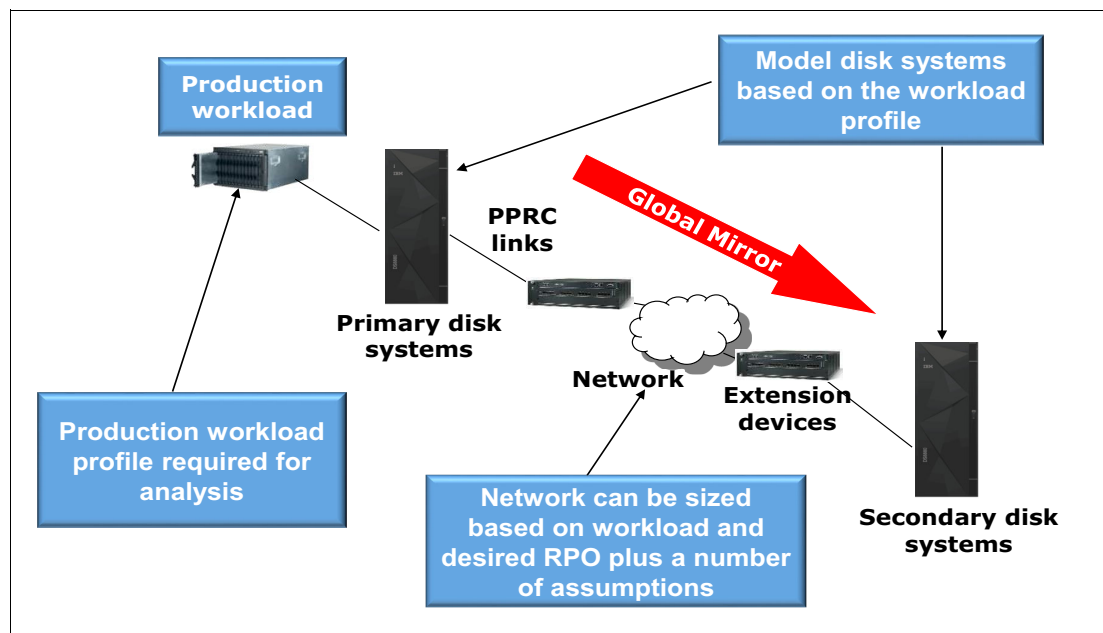
Figure 2-1   Global Mirror sizing considerations

Another key item to consider when sizing Global Mirror is connectivity between local and remote sites. It is important to determine how much bandwidth on a wide area network, or how many inter-switch links (ISLs) in a SAN fabric environment are required between the two sites. Network sizing can be quite challenging depending on the available infrastructure complexity.

Different tools can be used to collect information and size Global Mirror implementations. Some are more suitable than others, depending on environmental variables like Operational

System type, storage disk system brand already in place, and the performance management tools that are available for use. The following is a list of popular IBM tools:

- z/OS IBM Resource Measurement Facility™ (IBM RMF™) and RMF Magic
- IBM Spectrum™ Control (formerly known as Tivoli Storage Productivity Center)
- Global Mirror Bandwidth/RPO estimation tool
- IBM Disk Magic

2.1.1 RMF and RMF Magic

Resource Measurement Facility (RMF) is an IBM strategic product for z/OS performance measurement and management. It is the base product to collect performance data from z/OS and Sysplex environments, and to monitor systems performance behavior. It allows you to optimally tune and configure your system according to your business needs. RMF data collection provides workload profiles that are useful in determining the write MBps pattern that is required for Global Mirror sizing. RMF can be licensed for clients.

RMF Magic is a convenient tool to analyze RMF data. This tool can generate different charts and tables with the information you need as data input for Global Mirror Bandwidth and the RPO estimation tool. RMF Magic for IBM products is an internal IBM tool.

2.1.2 IBM Spectrum Control

IBM Spectrum Control™ storage management tool provides comprehensive IBM disk systems historical performance statistics and reports by capturing performance statistics directly from storage disk subsystems. Therefore, IBM Spectrum Control is a tool that can run on any operating system and can be used in mainframe, distributed systems, and IBM i platforms.

For Global Mirror sizing, several reports can be generated by IBM Spectrum Control such as reports by subsystem, by controller, by arrays, and by volumes. Different reports can be used for each specific case. For example, by volumes reports are useful to analyze specific volume groups belonging to applications or systems that need to be in a Global Mirror session. IBM Spectrum Control reports can be used as input for the Disk Magic tool. The IBM Spectrum Control license is based on managed usable capacity.

2.1.3 Global Mirror Bandwidth/RPO estimation tool

The Global Mirror Bandwidth/RPO estimation tool is a spreadsheet-based tool. It is designed to help size RPO and bandwidth, and the relationship between those two factors based on a particular workload profile. It uses a workload profile for at least 24 hours (ideally a few days). This profile helps define the parameters and assumptions of the environment to generate an RPO and bandwidth estimations.

It is important to have a workload profile that represents the heaviest write activity peak period and to input realistic assumptions for your environment.

The Global Mirror Bandwidth and RPO estimation tool can help demonstrate the effect of different communication link technologies and bandwidth on RPO. For instance, different estimations can be created by varying the number of links and their technology to understand the RPO behavior. Alternately, you can start from a wanted RPO value and find the number of required links of a specific technology that is needed to achieve that RPO.

The Global Mirror Bandwidth/RPO estimation tool is an internal IBM tool, and is for internal IBM use only.

2.1.4 Disk Magic

Disk Magic is used to model and estimate the future performance of storage systems that are attached to IBM z Systems®, distributed systems, and IBM i platforms. This tool can also size IBM storage disk system internal resources to accomplish performance requirements for a specific workload profile. Disk Magic is an IntelliMagic product that uses advanced algorithms for specific IBM storage systems, which are developed in close cooperation with the IBM performance teams.

Disk Magic supports manual and automated input of performance statistics. Performance statistics can be gathered through specific host-based tools:

► SMF/RMF data format from z/OS mainframe environment
► IOSTAT from IBM AIX®, Linux, Linux on IBM System z®, and other UNIX platforms
► Performance Tool reports from IBM i
► PERFMON statistics for Microsoft Windows
► ESXTOP of RESXTOP tools from VMware

Disk Magic supports IBM Spectrum Control reports for disk systems as well. Collect and format all these reports according to the specific Disk Magic instructions. Disk Magic for solving and modeling IBM disk products is licensed only to IBM.

> **Tip:** When modeling z/OS workloads, use the RMFPACK tool. This tool puts together all the RMF data that is needed to size a z/OS environment in a single file. This file is then used as automated input by the Disk Magic tool. RMAFPACK is a useful tool when sizing z/OS environments. It can be downloaded from the Disk Magic download page.

2.2 Global Mirror implementation planning

Global Mirror solution design depends on client disaster recovery requirements, mainly RPO. These requirements influence the overall infrastructure sizing and management, such as network and cross site connectivity selection, primary and secondary disk system configurations, and how to manage and control the environment.

Using appropriate tools and extrapolating performance statistics allows you to model and configure the required infrastructure. Even though available sizing tools have good levels of accuracy for Global Mirror sizing, chances are that variables out of your control, such as unexpected scenario changes, poor or incomplete performance statistics, and unexpected workload growth, can cause unexpected results. Therefore, use Global Mirror in the following two stages for production workloads. Management interfaces, including Copy Services Manager and GDPS, support implementing Global Mirror in this way.

1. Configure and start Global Copy, and monitor the environment.
2. Start the Global Mirror session.

Figure 2-2 shows all the required steps to set up and start the Global Copy environment. For Global Mirror setup command details, see *DS8000 Copy Services*, SG24-8367.
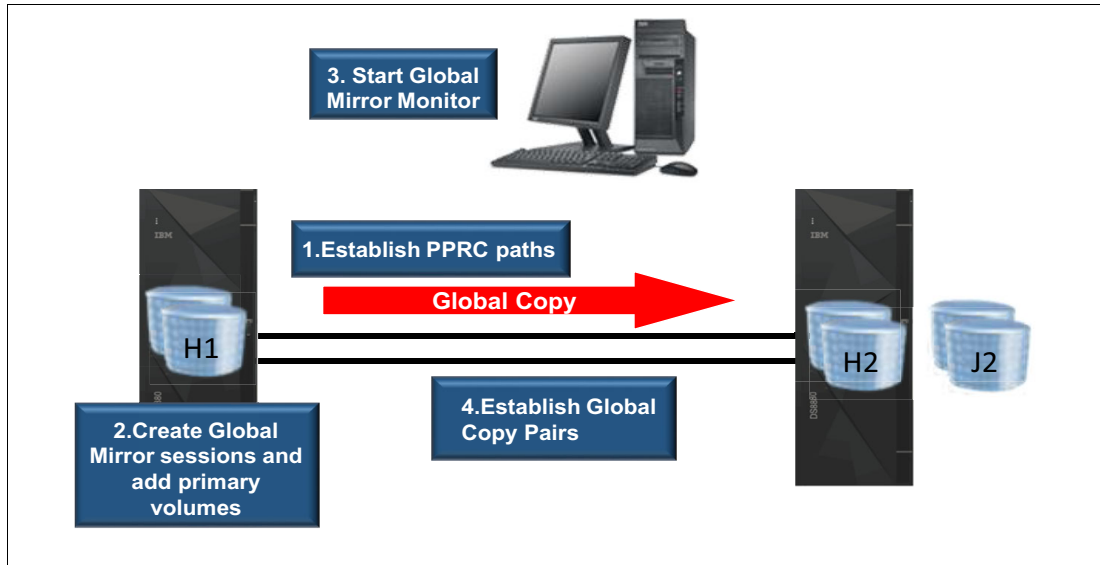
Figure 2-2   Configure and start Global Copy

If you have many volumes, consider starting Global Copy pairs for a certain volume's subset and gradually add more volume pairs while monitoring network utilization.

Wait until initial bulk data transfer is complete, and then run the Global Copy for 24 hours or more. Check out-of-sync tracks and the Global Copy behavior to ensure that they are working as expected. Moreover, investigate network utilization and statistics to discover and fix high rates of packet retransmissions and other errors if they exist.

With the network sizing and performance assumptions validated, continue setting up the Global Mirror session as shown in Figure 2-3.
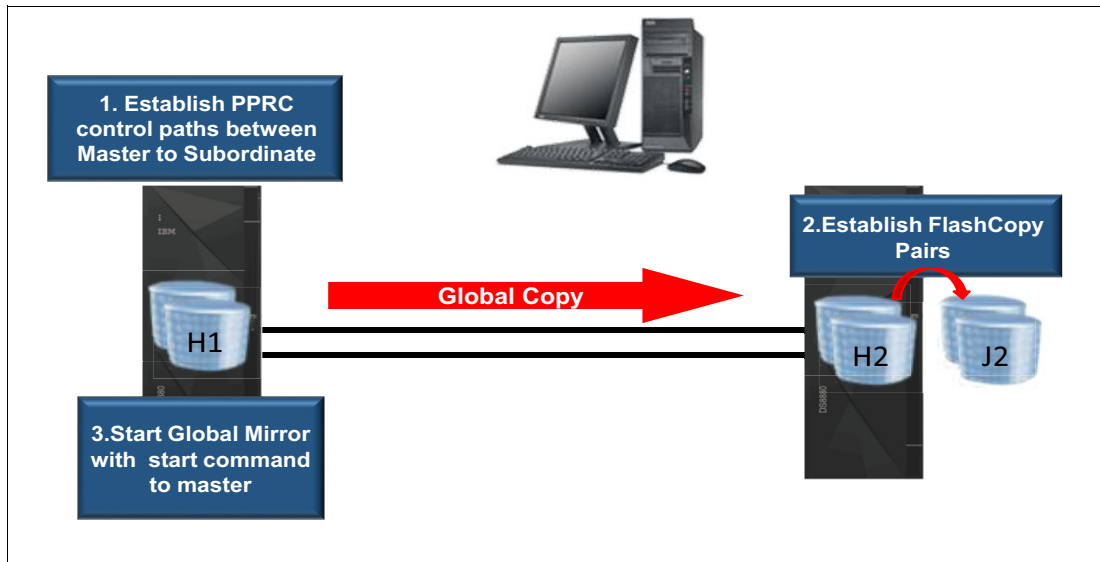
Figure 2-3   Configure and start Global Mirror session

Global Mirror should work as expected, assuming that a correct sizing was done and all resources, including network infrastructure, have been configured and correctly implemented.

Analyze out-of-synch tracks to ensure that they behave as expected and that RPO values meet the requirements. This activity can easily be done with GDPS and CSM management tools in place. Without these management applications, this analysis can be complicated.

Scripts should be created to continuously collect out-of-synch tracks measurements and the number of successful consistency groups created on a time interval basis. The interval time between each data collection operation should be adjusted to a multiple of the required RPO. Scripts must run for at least 24 hours, and preferably more. A large amount of time-interval-based information is created. An average RPO value must then be calculated for each time interval.

Another important point about management is that recovering a consistency group at a remote site after a disaster is not practicable without management applications specifically designed for that purpose like GDPS or CSM. It is not feasible and even too risky to manually deal with certain combinations of FlashCopy states for each logical volume participating in a Global Mirror session to make the correct decision while the consistency group is being restored.

> **Important:** Managing Global Mirror with scripting technics or any other kind of executable files like Command lists or REXX programming in mainframe platform and scripts or batch files with embedded DSCLI commands is complicated and risky. This limitation is especially significant when hundreds or thousands of volumes are managed by a single Global Mirror session.
>
> Generally, manage Global Mirror environment through management interfaces specifically designed for that purpose like Geographic Parallel Sysplex (GDPS) or Copy Services Manager (CSM). For more information, see 7.1, "Copy Services Manager" on page 70 and 7.2, "GDPS Global Mirror (GDPS/GM)" on page 72.

2.2.1 Growth within Global Mirror configurations

Growing Global Mirror configurations means adding resources to the environment or, in some cases, replacing existing resources with more powerful ones. That is, volumes, physical links, or even disk subsystems can be added to Global Mirror environments with minimum or no impact. In the same manner, resources can also be replaced to grow the environment, for instance, an entire disk subsystem being replaced by a more powerful one during technology refreshing or product decommissioning.

Resources can also be removed from the Global Mirror environment with minimal impact to the solution. Adding or removing an entire disk system is considered a topology change and requires that the Global Mirror session be stopped and then started again with the new topology in place. In this case, the RPO is higher than the amount of time that the Global Mirror session was stopped.

Adding or removing other resources like logical volumes or physical links or ports do not require stopping the Global Mirror session. Adding or removing links or ports can be done with no impact. This advantage is because the link or port being removed is not the last active one with the last logical path for any logical subsystem, and the remaining bandwidth is enough to support the environment.

Higher RPO and Global Mirror session status changes are observed when adding volumes to a Global Mirror session until new volumes are fully copied by Global Copy. After new volumes are fully copied, Global Mirror status and RPO will automatically come back to their normal states.

Removing volumes has no impact to Global Mirror environment.

Considering the following guidelines when changing Global Mirror configurations:

► When adding volumes to a Global Mirror session, evaluate link bandwidth requirements based on the expected workload on new volumes.

► Add volumes in small subsets and during low write activity periods to prevent replication link saturation and minimize RPO impacts.

► Removing physical links, DS8000 physical ports and related logical paths means less available bandwidth. Ensure that there will be enough bandwidth after removing any of those resources to keep Global Mirror working correctly.

> **Tip:** Generally, evaluate the infrastructure before you apply changes, and monitor the environment while changes are being implemented and after changes have been activated. Previous analysis might discover bottlenecks that should be eliminated before applying changes in a Global Mirror implementation. Monitoring the environment while changes are being implemented gives you the chance to take actions to minimize unexpected impacts. And change results must be evaluated as well.

2.3 Configuration guidelines for primary and secondary

The following are generic guidelines to be used during the design phase of a Global Mirror solution:

► Balance primary and secondary resources according to all solution requirements.

When possible, configure primary disk systems and their respective secondaries with equivalent internal resources. Besides helping meet required RPO, expected performance levels can also be reached when running production systems in the secondary site.

► One to One or Two to One configurations tend to be easier to implement and manage.

One primary disk system replicating data to a single disk system in a remote location is the simplest and easiest topology to be implemented and managed. A "two-to-one" or "one-to-two" topology adds a little more challenge for management, but it still practicable. However, as topology becomes more complex, most notably in a many-to-many design, the complexity dramatically increases. Therefore, keep Global Mirror simple from the implementation and management standpoints by avoiding complex topologies.

► Take special care when reusing old technology as Global Mirror secondary disk systems.

Older DS8000 models like DS8700, DS8800, and DS8870 can be used as Global Mirror secondaries and they work well in either Global Mirror site. However, make sure that available resources within these products meet all expectations. Old technology products need to be evaluated not only for performance, but also for capabilities that might not be supported by these disk systems. These capabilities might include thin provisioning for CKD volumes, multi-target PPRC, more efficient management against intermittent link problems, scalability limitations, and Flash cards.

- Keep your disaster/recovery solution active while accessing data at the remote site.

  It is a good practice to keep the Global Mirror solution active when data at the remote site needs to be accessed for tests or any other purpose to keep the data protected. An extra set of volumes called Global Mirror practice volumes can be used for any purpose in the secondary site while keeping Global Mirror active. This goal is accomplished by FlashCopy multi-target functions that create a consistent copy of the data on the practice volumes.

- Use the Global Copy failover failback capability:

  - Failover is a function that makes Global Mirror target volumes available to be read and written. It also creates a bitmap for each volume that is used to track write updates that occur in secondary disk systems. Then volumes in both sites are accessible for read and write operations, but with write operations being controlled by Global Copy.

    Failover capability is commonly used when doing tests in the secondary site with production systems running at primary site and there are no practice volumes in the solution. In this case, Global Mirror stops consistency group formation, but a consistent copy from the last created consistency group can be restored if a disaster strikes the primary site. Obviously this last consistent copy might fall behind depending on the elapsed time between the time Global Mirror stopped and the time that the disaster struck the primary site.

  - Failback is the function performed during Global Mirror restoration procedure after a failover function has been issued. For example, after all activity in the secondary site has finished, failback is used to reestablish the environment. A failback resynchronizes the primary and secondary volume pairs. Failback allows you to decide in which direction data resynchronization is run. If hot data is at the primary site, then select primary-to-secondary. By selecting the primary-to-secondary direction, all modified data at both sites is copied from the primary volumes to the secondary volumes. If you choose secondary-to-primary direction, the same process resynchronizes the data, but in the reverse direction. Without the failback capability, a full copy would be required to restore the Global Mirror environment every time data in both sites needs to be accessed by application hosts.

2.3.1 Primary disk system performance

Global Mirror protects production application performance at the expense of how current the consistent copy at the remote site is. Global Mirror ensures that during periods with insufficient bandwidth such as unexpected workload peaks, link problems, or secondary disk system issues, production performance is protected. In this case, Global Mirror stops consistency group formation and data is transmitted to the secondary site in the most efficient manner by Global Copy. When the system returns to normal, consistency group formation resumes in a timely fashion.

That is the case when Global Mirror falls behind and needs to catch up. Because the data to be transmitted is not in the primary system cache anymore, it needs to be read from the back end to read cache of the primary system, and then sent to the secondary site. Global Mirror stops consistency group creation and allows Global Copy to send data to secondary disk system. This process minimizes performance impacts because data is sent at a lower priority, so production is protected at the expense of a lower speed of data transmission to the secondary disk system.

A balanced production workload means that the I/O activity is spread among DS8000 main controllers (CECs), extent pools, and host adapters. Generally, configure an even number of extent pools, and have each one have the same number of ranks with the same

characteristics as much as possible. This configuration ensures that all extent pools provide the same performance levels.

Usually two extent pools provide performance and easy management. Spread application servers connectivity across all DS8000 host adapters (HAs). If possible, have dedicated DS8000 HAs for remote replication. These recommendations are valid to improve performance on secondary disk systems as well.

> **Tip:** For DS8000 Host Adapter cabling recommendations, see *DS8000 Host Adapter Configuration Guidelines*, TD105671, which can be found at this website.

Use performance monitoring tools such as IBM Spectrum Control and RMF reports for performance management. Historical performance information can help to understand what the baseline behavior should be, which helps during the troubleshooting process.

2.3.2 Performance at distance

As the cost of telecommunications decreases, businesses are looking to implement disaster recovery solutions at longer distances. Intercontinental distances are now more common and replication solutions must be able to support long distances

Distance can affect replication solutions, both by increasing the RPO and by decreasing the throughput. As an asynchronous replication solution, Global Mirror is designed to operate at long distances. However, as distances grow, more impact is experienced.

To send large amounts of data at long distances, a significant degree of parallelism is required to ensure that the bandwidth is fully used. With poor or little parallelism, throughput is reduced and a large amount of time is spent waiting for acknowledgements that data has been received at the remote location. However, at shorter distances, the same degree of parallelism might be counter-productive.

Another point related to parallelism and long distances is that write I/Os with small block size being replicated with poor parallelism at a very long distance are more affected than write operations carrying out large block sizes.

> **Notes:**
>
> Global Mirror sends multiple updates to the same track in a single operation to minimize the effect in small block write operations. That configuration means, for write streams such as in database log files, the number of write operations coming from the application server is significantly higher than Global Mirror operations that are required to transmit all application updates to the remote site. For example, in the Open Systems platform, 16 write operations with 4 KB block size of a database log file, which is 64 KB or one Fixed Block logical track in size, can potentially be sent in a single Global Mirror operation.
>
> DS8000 has a set of internal parameters known as *pokeables,* sometimes referred to as *product switches*. These internal parameters are set to provide the best behavior in most typical environments. In special cases, like intercontinental distances or when bandwidth is very low, some internal tuning might be required to adjust those internal controls to keep Global Mirror as efficient as it is in more common environments. Pokeable values can be displayed by a graphical user interface (GUI) or by Copy Services Manager, but they can only be changed by IBM technical support teams.

2.3.3 Secondary disk system performance

As a general recommendation, ensure that a Global Mirror secondary disk system can provide the same or better performance than its counterpart disk system at the primary site. Global Mirror puts more stress on the secondary disk system because of FlashCopy processing, and so more resources might be required to achieve good performance.

FlashCopy with the **nocopy** parameter is used by Global Mirror to save consistency groups. The **nocopy** parameter generates activity in the disk system's back end only when a piece of data that was updated needs to be destaged from the read cache. Before destaging the updated data of a FlashCopy source volume, the *Copy-on-Write* approach used by FlashCopy saves the original copy in the back end by copying it to the corresponding FlashCopy target volume.

The FlashCopy *Copy-on-Write* approach works on a logical track basis, meaning that if just a small block of data is written to logical track, that entire track is copied to its correspondent FlashCopy target volume. With the **nocopy** parameter, if a logical track is not updated, then it does not need to be copied to the related FlashCopy target volume. To preserve the point-in-time copy, this process runs only once during the FlashCopy relationship lifetime.

When a FlashCopy relationship is established or if it is refreshed, which is the case for Global Mirror, then FlashCopy bitmaps are reset and the whole process starts again.

Figure 2-4 illustrates the *Copy-on-Write* process as it occurs between Global Mirror target volumes and Global Mirror Journal volumes. When a write operation (1) is sent to a Global Mirror target volume (H2), that write operation is acknowledged immediately (2) if there is space in the DS8000 nonvolatile storage (NVS). Before allowing the update to be destaged (4), the previous track on the Global Mirror target volume (H2) must be copied to its corresponding Global Mirror Journal volume (J2) in (3). After the previous copy is saved in the Global Mirror journal volume (J2), then the updated data is destaged (4).
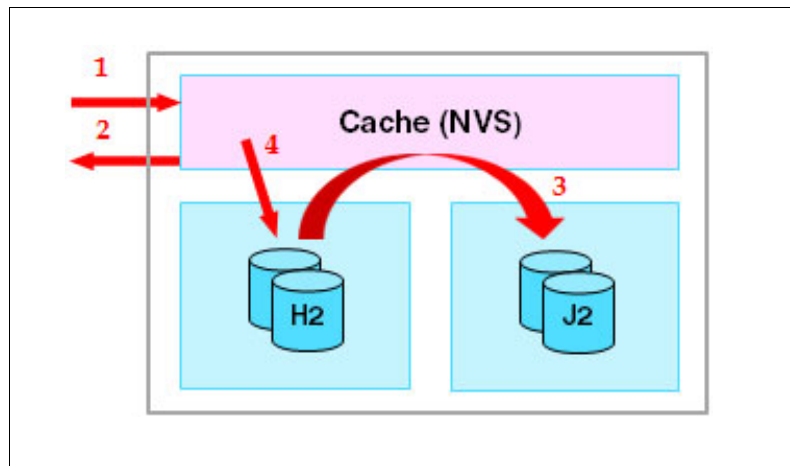
Figure 2-4   Secondary disk system performance

During normal operations, with consistency groups being created every 3 - 5 seconds, the FlashCopy overhead generated by the *Copy-on-write* process tends to be low in comparison with a situation in which Global Mirror stops creating consistency groups for any reason or falls behind and needs to catch up. That is the case where the back end is challenged by write streams coming from primary disk systems plus a huge FlashCopy workload caused by the journaling activity to preserve data consistency.

If using fully provisioned volumes for Global Mirror journaling, the required additional space should provide good performance by doubling the number of physical disks. However, when physical disks with larger capacity are used instead, then a deeper analysis might be required.

Thin provisioning for Global Mirror journal volumes requires less additional space and obviously, it is expected to be the most popular configuration because of capacity savings. When sizing secondary disks with thin provisioning volumes, additional variables must be considered. In particular, consider the write activity intensity and its distribution among extent pools and space allocation granularity, which depends on the extent size being used.

Work with IBM support teams when using thin provisioning for Global Mirror journal volumes.

A well-configured secondary disk system that has all resources that were accounted during the sizing process might not work well if its logical configuration does not follow the recommendations for best performance. Spreading the workload across all extent pools helps avoid back-end constraints. Moreover, flash technology combined with DS8000 *EasyTier* functionality can reduce even more back-end saturation events.

EasyTier can efficiently distribute workload across all ranks in homogeneous and in heterogeneous or hybrid extent pools. *EasyTier* is preferred for almost all kinds of workload and therefore it is preferred for Global Mirror implementations too.

2.3.4 Volume placement on secondary disk system

Figure 2-5 illustrates how extent pools and volumes should be configured as a general rule for performance and management when planning volume placement in Global Mirror secondary disk systems.
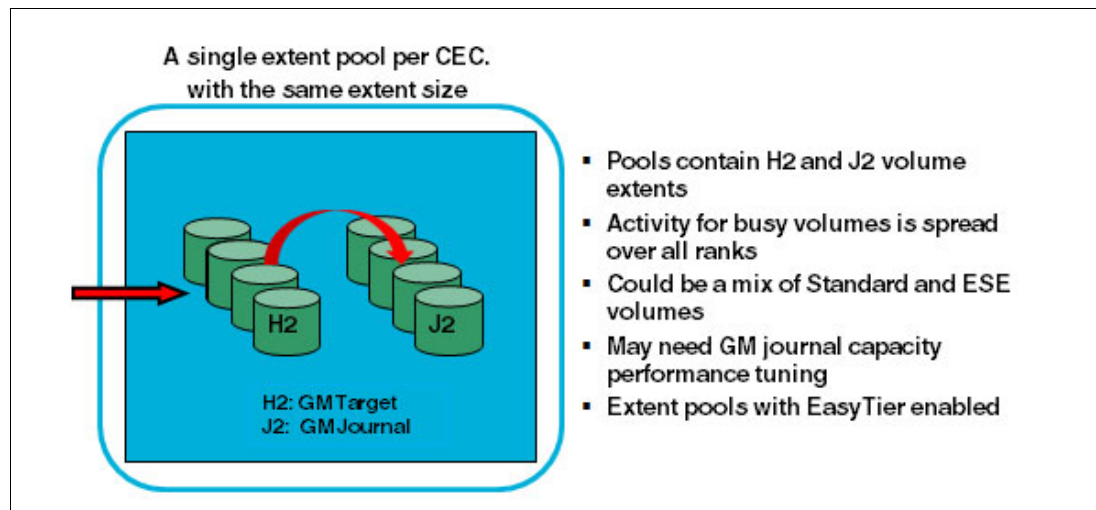
Figure 2-5   Secondary disk system volume placement without practice volumes.

As can be seen in Figure 2-5, with a single extent pool per DS8000 server (CEC), it is preferred to have each extent pool holding both Global Mirror target volumes (H2) and their correspondent Global Mirror Journal volumes (J2).

Mixing extents with different extent sizes in one extent pool is not allowed. Although extent pools can only have extents of the same size, fully provisioning and thin provisioned volumes can coexist in the same extent pool independently of what the extent size is.

Some internal tuning might be required to improve FlashCopy efficiency, depending on the initial space allocation percentage in the extent pools. This internal tuning, when it is required, is done by IBM technical support team during the Global Mirror implementation process.

As stated in Figure 2-5, EasyTier is preferred to manage data access pattern no matter if extent pools are homogeneous or hybrid. For more EasyTier information and specific recommendations for Global Mirror, see "Taking advantage of Easy Tier" on page 31.

With practice volumes in place, as showed in Figure 2-6, the same rules apply. Note that now Global Mirror practice volumes (H2) are to be accessed by application hosts in the secondary site while Global Mirror target volumes (I2) become intermediate volumes with no access to the application servers.
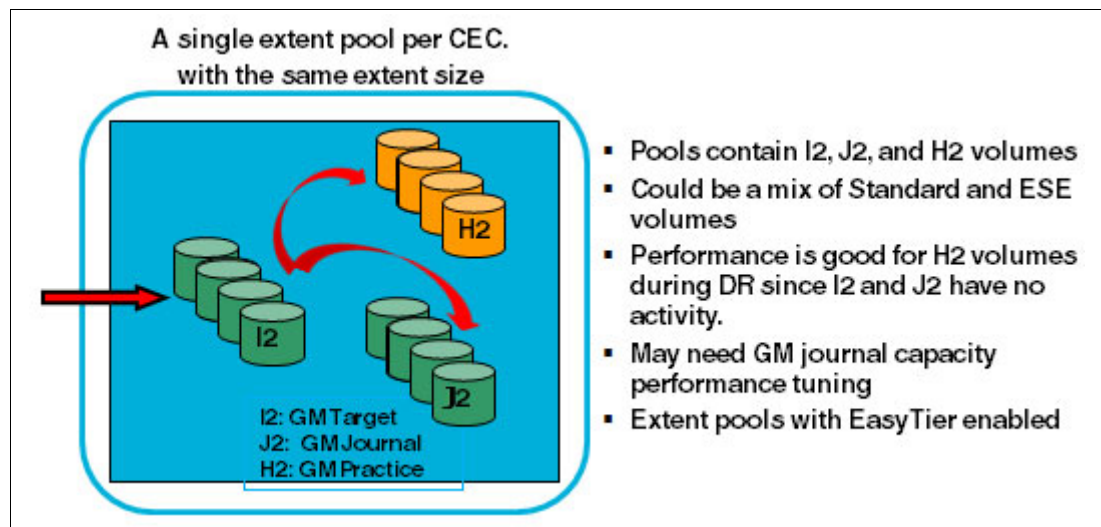
Figure 2-6   Secondary disk system volume data placement with Global Mirror practice volumes

Now Global Mirror targets and their related Global Mirror journal and Global Mirror practice volumes should all be in the same extent pool.

As previously mentioned, the DS8000 has a set of internal parameters called *pokeables* or *product switches.* By default, these parameters are set to provide the best behavior in most typical environments. However, due to the nature of the thin provisioning technology, when thin provisioned volumes are used for Global Mirror journaling, some tuning might be needed by changing some pokeable default values. Pokeable values can be displayed by using a GUI or Copy Services Manager, but changing a pokeable value can be done by IBM technical support only.

2.3.5 Global Mirror secondary disk system recommendations

Global Mirror imposes more workload on secondary disk systems than on primary disk systems because of the journaling process continuously running in secondary disk systems back end and cache.

During write workload peaks with configured secondary disks systems, internal resources reaching their saturation limits causes RPO to skyrocket to the point that RPO values can be measured in hours instead of seconds or minutes.

Preferred configuration: Sizing is a key point. Global Mirror secondary disk systems must be evaluated with a focus on write activity peaks when production is running at the primary site. Also, and not less important, analysis must be done to size secondary disk systems when the production workload is running at the secondary site. Use the Disk Magic tool to model secondary disk systems for both scenarios.

Evaluate external links bandwidth to make RPO estimations. Work with performance statistics information containing write workload peak data for at least a 24-hour period, and ideally for an entire week. Use the Bandwidth/RPO estimation tool to determine the replication link infrastructure that is required for a specific RPO.

Global Mirror secondary disk systems deal with specific workloads such as journaling activity. Consider the fact that modeling tools cannot deal with all aspects that are involved in a Global Mirror environment. For instance, the number of FlashCopy relationships, the space required for internal controls, and the journaling workload cannot be directly modeled by sizing tools. Therefore, more cache size guidelines are required when sizing Global Mirror secondary disk systems.

Recommendation: Cache size should be evaluated according to IBM modeling tools. However, there are also cache size guidelines for Global Mirror secondary disk systems. Based on the number of Global Mirror primary volumes being implemented, find the secondary disk system preferred cache size in the lists below. Compare that cache size with the cache size given by the modeling tool and select the higher value as the secondary disk system cache size.

For all members of the DS8000 family before the DS8880 models, use the preferred secondary disk system cache size based on the maximum number of related primary volumes:

- ► DS8000 16-GB cache is preferred for 750 primary volumes.
- ► DS8000 32-GB cache is preferred for 1,500 primary volumes.
- ► DS8000 64-GB cache is preferred for 3,000 primary volumes.
- ► DS8000 128-GB cache is preferred for 6,000 primary volumes.
- ► DS8000 256-GB cache is preferred for 12,000 primary volumes.
- ► DS8000 384-GB cache is preferred for 18,000 primary volumes.
- ► DS8000 512-GB cache is preferred for 24,000 primary volumes.
- ► DS8000 1024-GB cache is preferred for 48,000 primary volumes.

Preferences for all DS8880 models:

- ► DS8880 64-GB cache is preferred for 6,000 primary volumes
- ► DS8880 128-GB cache is preferred for 12,000 primary volumes.
- ► DS8880 256-GB cache is preferred for 24,000 primary volumes.
- ► DS8880 512-GB cache is preferred for 48,000 primary volumes.
- ► DS8880 1024-GB cache or larger are preferred for more than 48,000 primary volumes.

Secondary disk systems back-end resources are important because they influence RPO behavior and affect performance levels when production is running at the secondary site.

Because secondary disk systems must keep up with a heavier workload generated by FlashCopy activity along with write workload coming from primary disk systems, physical disks with lesser performance are not recommended when configuring the secondary disk system. Furthermore, in some cases, flash cards, solid-state disks, or both might be necessary even if these cards or disks are not configured in primary disk systems.

Guidelines:

► Do not use Nearline drives in Global Mirror secondary disk systems. As a suggestion, use Enterprise class spinning disks with large capacity instead. For instance, when there are Nearline drives in the primary disk system, equivalent storage capacity can be configured in secondary disk systems by using Enterprise class drives with large capacity.

► Avoid using spinning disks larger than twice the size of disk drives in the primary disk system.

► Have disk drives with the same rotational speed for Global Mirror target, Global Mirror journal, and Global Mirror practice volumes, if present.

► Global Mirror secondary disk systems using thin provisioned volumes as Global Mirror targets, Global Mirror journals, or both should have at least 20% of each extent pool capacity configured with Flash cards, solid-state disks, or both.

DS8000 most recent improvements like small extents support and thin provisioning capability for CKD volumes give multiple configuration options by combining large and small extents with fully or thin provisioned volumes for both mainframe and open systems platforms. Multiple options bring flexibility and efficiency improvements, but must be evaluated.

Guidelines:

► Configure two extent pools in secondary disk systems so that each extent pool is managed by each one of the two DS8000 main controllers (CECs). Global Mirror target volumes and their corresponding journal and practice volumes are in the same extent pool.

► When using fully provisioned volumes and you have no intention to use thin provisioning, have extent pools with large extents.

► If planning to use thin provisioning, use extent pools with small extents. Even when mixing fully provisioned volumes and thin provisioned volumes in the same extent pool, use extent pools with small extents.

► When the Global Mirror primary site already has existing fully provisioned volumes, regardless of the extent size being used, with the microcode release 8.0 or later, secondary disk systems can have fully provisioned or thin provisioned volumes as Global Mirror targets. In that case, use extent pools with small extents if mixing thin provisioned and fully provisioned volumes in the same extent pool.

   If secondary disk systems have already existent volumes in extent pools with large extents that cannot be reconfigured to use small extents, thin provisioned volumes are also preferred. However, in this case, thin provisioned volumes will not benefit from the thinner granularity small extents provided when capacity needs to be allocated.

► Global Mirror practice volumes should be fully provisioned except when Global Mirror target volumes are configured as thin provisioning. The reason is that practice volumes are made ready by a FlashCopy with the *copy* attribute operation and if Global Mirror target volumes are fully provisioned, the FlashCopy with the *copy* attribute causes Global Mirror practice volumes to have their extents fully allocated.

Thin provisioning technology provides efficient storage capacity utilization. With thin provisioned volumes, application servers see only the volumes' virtual capacity, which is bigger than the available correspondent real capacity in a disk system.

Note: Starting with DS8000 Release 8.1, end-to-end thin provisioned volumes are supported by using small extents in Peer-to-Peer Remote Copy implementations. Global Mirror for both mainframe and distributed systems platforms can be implemented with all logical volumes in both primary and secondary sites, practice volumes included, configured as thin provisioned in extent pools with small extents.

If real space is lacking, application servers are not able to write onto thin provisioned volumes. Global Mirror behavior is the same. If any extent pool in Global Mirror secondary disk system runs out of capacity, Global Mirror stops creating consistency groups and the primary-to-secondary data transmission that is done by Global Copy is suspended.

Guidelines:

When using thin provisioning, capacity management is strongly advised. Alerts based on the extent pool capacity utilization threshold must be made active so warning messages that use SNMP protocol or z/OS messages that use syslog can be received and treated.

The extent pool threshold is set as a percentage of the remaining extents in the extent pool. The default value is 15 percent. Generally, change the extent pools threshold to 20 percent for better protection.

In addition, configure extents pools with reserve capacity. Generally, reserve at least 10 percent of total extents in each extent pool for future use. Doing so gives you 10 percent additional capacity that can be made available if a storage capacity upgrade cannot be provisioned quickly. This 10 percent reserved capacity can be added to the extent pool after the first warning message indicating that the extent pool utilization threshold is reached and its available capacity is low.

2.3.6 Taking advantage of Easy Tier

Most environments will usually experience I/O skews and hot spots. IBM Easy Tier® can help alleviate those situations in hybrid or homogeneous extent pools in both local and remote storage systems.

Figure 2-7 on page 32 shows an overview of two DS8000 systems in a Global Mirror session with Easy Tier enabled on both storage systems. The performance is gradually adjusted as Easy Tier analyzes the characteristics of the workload on each disk system and moves the data to the appropriate tier.

However, EasyTier in the Global Mirror secondary disk system creates a different data movement than its counterpart in the primary site. It receives only write activity coming from primary disk systems and also deals with journal activity. From that perspective, performance in the secondary site is different from the primary site levels after a site switch to move the production workload to the secondary site. Easy Tier then requires some time to adapt to the new workload, create migration plans, and migrate data across tiers.

Performance levels in the secondary site will become equivalent to performance in primary site only after Easy Tier has finished migrating all extents, which might take more than 24 hours, based on the production workload profile.

As previously stated, nearline class disks are not recommended for Global Mirror secondary disk systems. This recommendation is why the secondary disk system in Figure 2-7 on page 32 is configured with a Flash tier and an Enterprise tier. Following the recommendation,

the corresponding nearline storage capacity of the primary disk system was configured with enterprise class drives in the secondary system.

You can also notice that data placement in both sites is similar despite different workload profiles for each site. This equivalent data placement is accomplished by a DS8000 function named Heat Map Transfer (HMT) combined with an application called Heat Map Transfer Utility.

The HMT function and utility allows the secondary disk system to have a data distribution across its tiers that is similar to that of the primary. This configuration enables similar levels of performance immediately after a site switch operation.
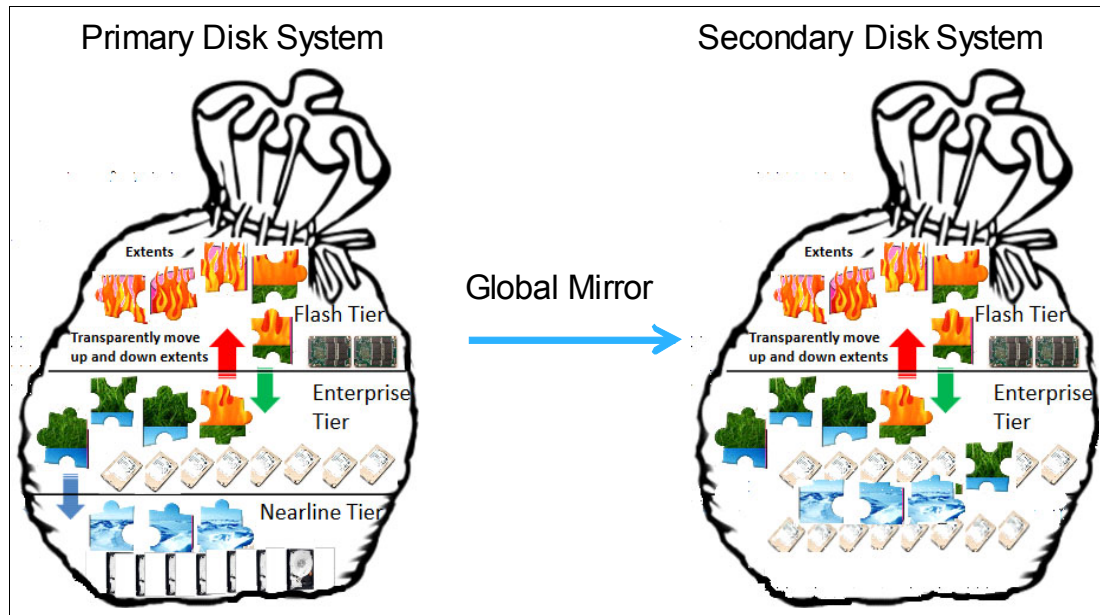

Figure 2-7   Logical configuration demonstration with Easy Tier functions

To provide equivalent performance levels after a site switch event, the Heat Map Transfer Utility, periodically or on-demand, transfers the heat map information from the primary to the secondary disk system. This operation is done twice a day in 12-hour intervals.

The secondary disk system then replaces its original heat map with the primary disk system heat map and creates a migration plan to be eventually run. This process ensures similar performance at both sites, as soon as application servers are made active at the secondary site.

> **Tip:** For secondary disk systems, have the same physical configuration as the primary system except for nearline disks. However, you can use additional flash cards or solid-state disks with thin provisioning in place to achieve good performance and low RPO.
>
> Heat map transfer capability is integrated in GDPS and CSM management interfaces. It can also be implemented as a stand-alone application running in Windows or Linux servers that are connected to both primary and secondary disk systems through the IP protocol. For more information about EasyTier functions and the EasyTier Heat Map Transfer Utility, see the following documents:
>
> ► *IBM DS8000 EasyTier*, REDP-4667.
> ► *IBM DS8870 Easy Tier Heat Map Transfer*, REDP-5015.

3

Connectivity

This chapter describes how to set up the connectivity to the remote site. It also gives an overview of the required components and how to set up the different options and functions.

This chapter includes the following sections:

- ► Inter-site connectivity
- ► Fibre Channel to IP conversion
- ► Bandwidth estimation
- ► Long-distance link considerations
- ► DS8000 configuration considerations

3.1 Inter-site connectivity

When Global Mirror is used, there are typically two data center sites, characterized by either a reasonable distance between them, or connected with a limited bandwidth. The data connection between both sites is usually implemented with high-speed WAN connections provided by a third-party network provider. The providers typically offer IP-based links with dedicated bandwidth. Today technologies are often Synchronous Optical Network (SONET) and Synchronous Digital Hierarchy (SDH) based connections. Depending on the global region, other standards might be available.

The available different bandwidths are defined by the Open Carrier Transmission Rate standard. They are named by the acronym OC-*n* where *n* is an integer number that represents the bit stream transmission rate. Table 3-1 gives an overview of the available transmission rates.

Table 3-1   SONET / SDH transmission rates

OC - Level	Data rate	SONET / SDH
OC-1	51.84 kbps	STS-1 / STM-0
OC-3	155.52 kbps	STS-3 / STM-1
OC-12	622.08 kbps	STS-12 / ST-4
OC-24	1,244.16 kbps	STS-24
OC-48	2,488.32 kbps	STS-48 / STM-16
OC-192	9,953.28 kbps	STS-192 / STM-64
OC-768	39,813.12 kbps	STS-786 / STM-256

These connections are typically IP-based and are provided to a data center as a so-called *access point*. The provider installs a communication device in the data center to provide IP-ports with one of these data rates.

3.2 Fibre Channel to IP conversion

The DS8880 only provides Fibre Channel based I/O ports for PPRC connections. Therefore, a gateway that converts Fibre Channel to IP must be implemented in each data center that hosts a DS8000. Figure 3-1 on page 35 shows a possible setup for a Fibre Channel to IP conversion.

In this setup, the Fibre Channel connections are connected directly to the FC/IP Gateway. However it is possible to use Fibre Channel switches in between. For more information about such a configuration, see "Fibre Channel flow control" on page 39.

To provide sufficient redundancy for the inter-site communication, at least two Fibre Channel connections should be provided to connect the FC/IP gateway. In addition, additional Fibre Channel connections can be considered to provide sufficient bandwidth. To meet the redundancy requirement, select an even number of connections. The most common environments use two, four, six, or eight Fibre Channel connections at each site.

On the IP-side of the gateway, the number of connections must be sufficient to at least match the total bandwidth on the Fibre Channel side of the gateway. For example, with four 8 Gbps FC connections that amount to a total bandwidth of 32 Gbps, four 10 Gbps IP links must be connected to the access point of the service provider. To realize a 32 Gbps link between the two sites, four OC-192 links are required.

For additional information about bandwidth requirements, see 3.3, "Bandwidth estimation".
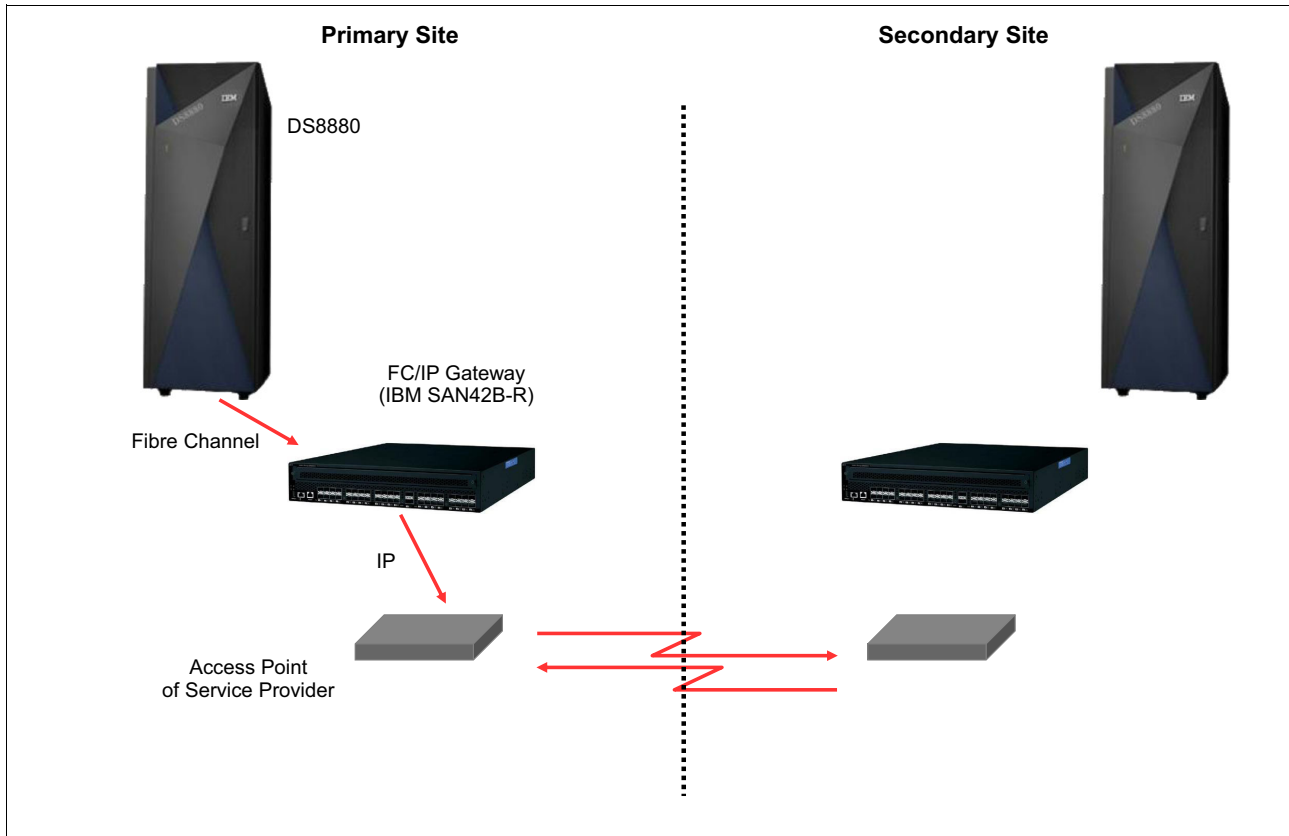

Figure 3-1   Converting Fibre Channel to IP traffic

3.3 Bandwidth estimation

Estimating the bandwidth requirement for asynchronous replication requires a little more attention than for synchronous replication. For synchronous replication, it is easy to determine the bandwidth with these steps:

1. Get a representative measurement of the write performance of the primary storage system.

2. Take the peak load.

3. Size the bandwidth of the replication link according to this peak.

For asynchronous replication, another important parameter comes into play as introduced in 1.1, "Global Mirror overview" on page 2: The recovery point objective (RPO). The reason is simple: With asynchronous replication, the data is transmitted to the remote site block by block, without any regard for the write order from the host. In other words, the data is inconsistent for some time, and a portion of data would be lost in a site disaster. This portion is measured by the RPO given in time frames like minutes or hours.

It is important that you understand this circumstance in asynchronous replication and that you specify how much data you are willing to lose in a disaster case. This decision might take some discussions because most companies do not want to lose any data. The correct approach to minimizing the data loss is to understand that either you need to provide more bandwidth, which costs money, or to reduce the distance to the secondary site to reduce signal latency.

This discussion produces an number of minutes or hours for the RPO. This value is the main input parameter for the bandwidth estimation. The required bandwidth calculation can now be based on the following parameters:

► The write load of all volumes that will participate in the data replication.

► The latency of the physical link between the data centers. This value can be requested from the link provider.

► The number of participating DS8000 volumes. As you can see in Figure 1-1 on page 3, one of the essential components of Global Mirror is the FlashCopy at the secondary site. This FlashCopy is a major component for Global Mirror to provide consistency at the remote site. FlashCopy operations are time consuming and thus must be taken into consideration.

► Expected compression rate. The FC/IP gateways offer hardware and software data compression. See 3.4.3, "Compression" on page 41.

One way to calculate the required bandwidth is to use DiskMagic. However, DiskMagic does not take a specific RPO into account.

The better approach is to contact your IBM representative. IBM can do bandwidth studies by using its own *Global Mirror RPO and Bandwidth Calculator*. The main input for this tool is a set of write performance data that can be provided by using IBM Spectrum Control data or RMF data. The result consists of two corresponding graphs that show the expected RPO for a particular bandwidth. The bandwidth can be changed interactively so that the required RPO can be adjusted.

For example, as illustrated in Figure 3-2, a calculation of the RPO was done based on a rough assumption of 20 Mbps. The first graph shows the following data:

► The black line represents the data transfer pipe.
► The green line shows the data to be sent to the secondary line.
► The blue line shows the emerging write rate at the primary storage system.
► The red line shows the accumulated writes that could not be sent to the secondary site because the bandwidth was not sufficient.

The second graph shows the RPO profile. In every interval where data was falling behind because there was more written than the link could transmit, the RPO ramps up until all accumulated writes could be transmitted to the secondary site.

In conclusion, with this write profile and a bandwidth of only 20 Mbps, the RPO can raise up to 9602 seconds and the maximum backlog is 13.01 GB.
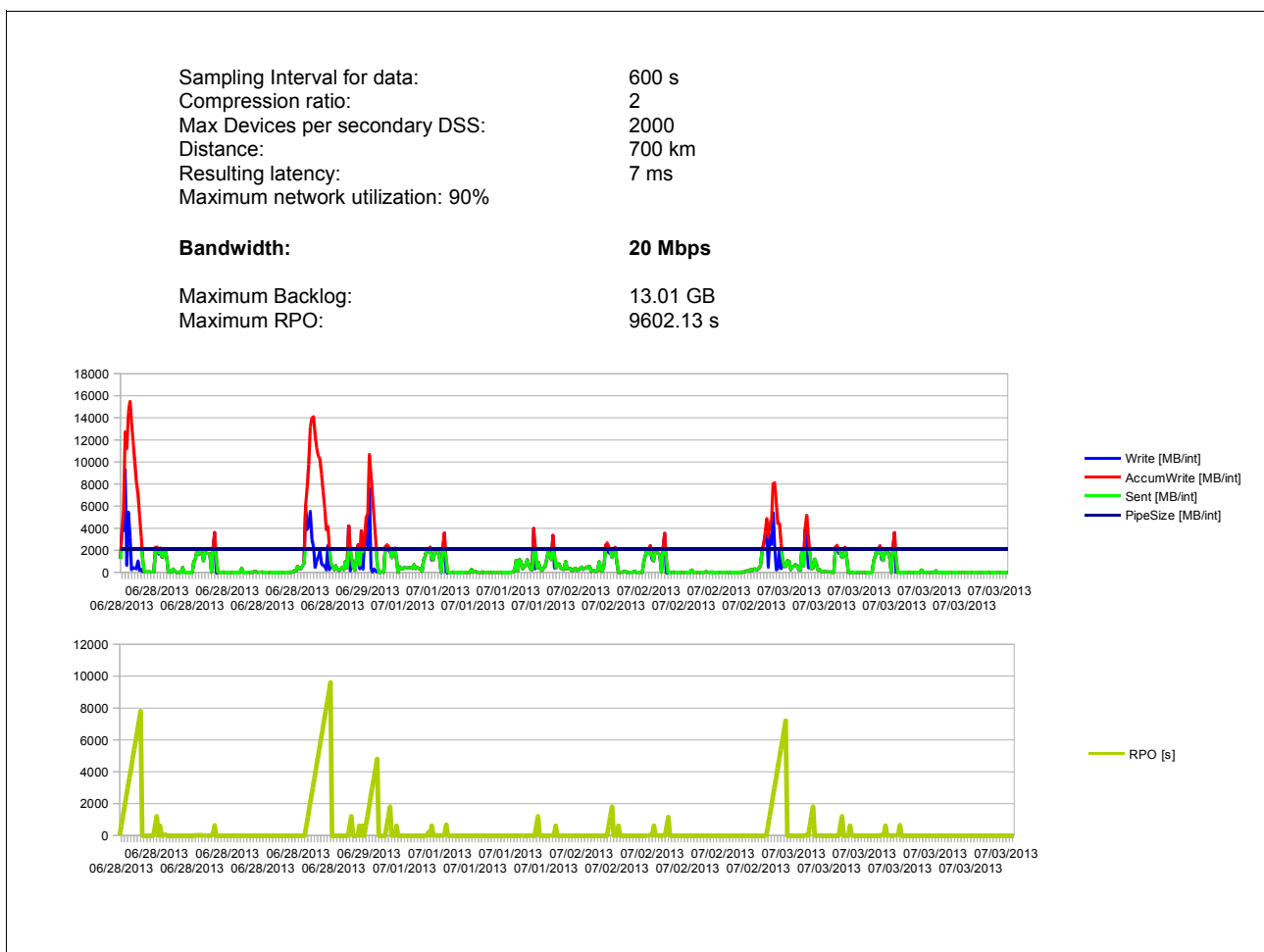

Figure 3-2   Bandwidth and RPO estimation with 20 Mbps

In Figure 3-3, the bandwidth was adjusted in such a way that the RPO will not be higher than the data sampling rate of the measurement. The RPO will be close to 10 minutes, or 600 seconds. The data backlog will be at roughly 4 GB. All of these goals can be achieved with a bandwidth of 47 Mbps.
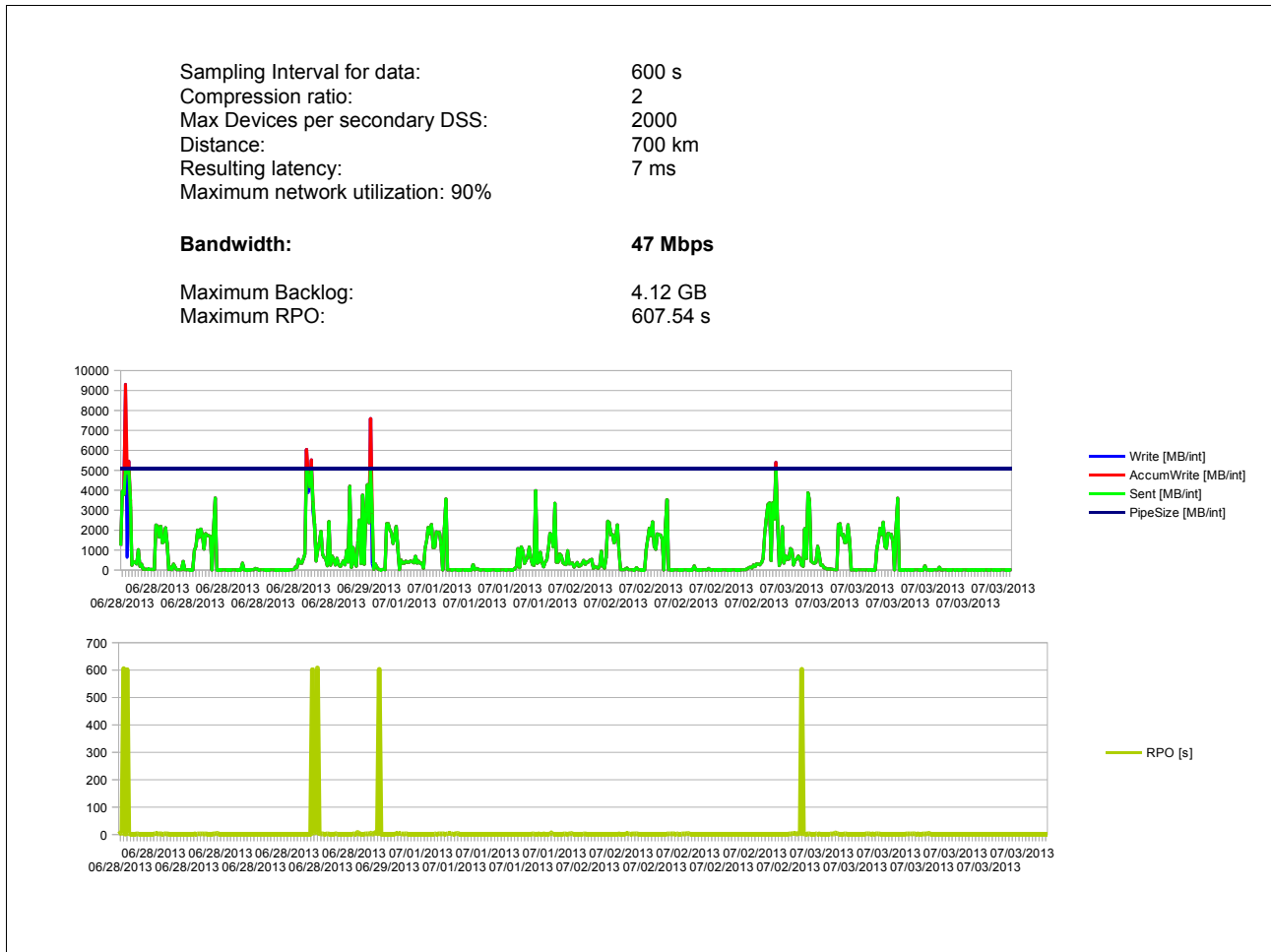

Figure 3-3   Adjusted Bandwidth estimation with optimized RPO achievement

As you can see, the Bandwidth Calculator is a great tool to facilitate the discussion of RPO and bandwidth.

3.4 Long-distance link considerations

When data is transmitted over longer distances, the data might encounter on its way to the target site changes like sections with different bandwidth or different transmission protocols. These circumstances influence attributes like throughput and response time, which are recognized by the participating DS8000 disk systems. It can even influence the host I/O when the data flow coordination or the sizing of the different link components is not done properly.

This section provides an overview of the most common attributes that require attention.

3.4.1 Fibre Channel flow control

A simple inter-site connection is shown in Figure 3-1 on page 35. However, additional Fibre Channel switches might be needed between the DS8000 and the FCIP gateway. If so, the resulting inter-switch links (ISLs) need special attention.

First, the total bandwidth of the ISLs must be sufficient. For example, if there are four links from the DS8000 in use for the replication, the ISLs should consist also of four links of the same speed. If the ISL ports have different speeds, as many ISLs as the sum of the DS8000 link speed must be supplied.

Especially when multiple Fibre Channel hops are in place, it becomes important to take a closer look at the Fibre Channel flow control. As you might know, the basic unit of data transmission is a frame. The Fibre Channel protocol consists of a control mechanism named *Buffer Credits* that allows a maximum utilization of ISL, and therefore a contiguous data flow from source to target.

Buffer credits are portions of memory in the target communication device in which the received frames are held until they are processed. If frames are arriving while the receiver is busy processing earlier frames and all buffer credits are used up, the transmitter will not be able to send data anymore. The data flow is then disrupted.

To calculate the correct buffer credits, consider the following items:

► The distance to the remote site and the speed of the link

  The longer the distance, the more frames can be sent to fill up the link. For example, on a 10 km link with a bandwidth of 1 Gbps, roughly one single full Fibre Channel frame fits on the link. With 2 Gbps, it is two full frames, with 4 Gbps four frames, and so on.

► The average size of the Fibre Channel frames

  A Fibre Channel frame size is a maximum of 2148 bytes long, whereby the payload can vary up to 2112 bytes. With data replication, the maximum payload is typically used. However, verify what the actual average frame size is.

► The round-trip time

  Although the distance between the two replication sites is fixed and the expected latency can be calculated, the total round-trip time should be the foundation for the buffer-buffer credit calculation. The values of total round-trip time that should be used can be measured by using the FCIP gateway.

3.4.2 Configuring the FCIP gateway

The following are considerations when configuring the FCIP gateway:

► Tunneling
► Max/min bandwidth
► Keep alive timeout value
► SCSI Fast Write or SCSI Write Acceleration

Tunneling

The FCIP gateway allows you to transmit Fibre Channel frames transparently over IP networks. This result is achieved by building one or more tunnels between the FCIP gateways. At each end of a tunnel, a Fibre Channel port is presented to the participating Fibre Channel devices.

Depending on the gateway vendor, with IP links or IP circuits it is possible to allow trunking of multiple IP links. This configuration enables optimized bandwidth utilization and management of link redundancy. For more information, see the documentation of the implemented FCIP gateway vendor.

Max/min bandwidth

In IP networks, the traffic load is controlled by an algorithm to avoid an over commitment of the network, which would lead to packet losses and performance problems. The way that the algorithm works is that when a bunch of IP data must be delivered, a *segment window of data* is defined and sent to the remote site. The system then waits for an acknowledge. If the acknowledge is received quickly, the segment window is doubled until a defined segment window threshold has been reached. If the traffic can still be handled, the segment window is increased from now on in a linear manner.

When the ceiling of the available bandwidth is reached, depending on the implemented algorithm, the segment window is either set back to one or is reduced to 50% and the algorithm starts over again.

This process sometimes leads to the saturation of the links, displaying a sort of sawtooth shape instead a flat line close to the ceiling of the bandwidth. In this case, the overall throughput of the link will not reach the capacity of the link as ordered by the provider. With FCIP routers, this effect can be reduced by supplying a threshold value for the expected maximum bandwidth and a minimum bandwidth. The FCIP routers then use their own congestion algorithm that helps to flatten the bandwidth deviation to a minimum.

For more information, see the user documentation of your FCIP routers. Monitor the bandwidth behavior and, if necessary, adopt both values.

Keep alive timeout value

As mentioned, the FCIP gateways are using a tunnel with typically two or four IP links or IP circuits. The tunnel itself is stateless and thus cannot be monitored except whether the connection is still available or not. But the underlaying IP links can do the monitoring by using keepalive messages that are returned from the remote gateway. If one link is not responding, the sender waits for a timeout value before link is brought down. The embedded routing protocol of the gateway then uses the next IP links. If this link has the same problem, the sender again waits for the response until the keep alive timeout value is due.

If the keep alive timeout values are larger than the timeout value of the PPRC links, the DS8000 sets the link that was most recent used to send data as degraded. This process occurs even if other IP links are still available. To avoid this situation, the keepalive timely value should be in total less than the PPRC timeout value for a link, which is typically 6 seconds.

SCSI Fast Write or SCSI Write Acceleration

In the standard SCSI model, each SCSI write command is done in two phases, where a command request is first sent from the initiator to that SCSI target. The target is sent back to the initiator that tells whether the target is ready or not. When ready, in a second phase the write command including the data is sent to the target.

With SCSI Fast Write or SCSI Write Acceleration, the sender gateway sends the command request and the data to the receiver in one go. The receiver then sends back only one acknowledge to the sender. The aim of this function is to reduce protocol interaction and saving transmission time.

Newer DS8000 code levels tolerate the SCSI write acceleration, but the DS8000 does not get any advantage from it. Therefore, in this case disable the SCSI write acceleration.

3.4.3 Compression

Compression is a method to reduce the amount of data before transmission. With compression, the data stream is analyzed for redundant pattern in a specific window of the data stream. When patterns are found, they are replaced by a shorter representation of this pattern.

This function requires some processing resources on both the sender and receiver gateways. Different implementations are available. In general, a hardware-based implementation allows a compression rate between 1:1.5 and 1:3, and software-based implementations can achieve higher compression rates. The hardware-based solutions are faster, and the software-based solutions consume some operation time, which is in addition to the link latency.

Remember that the compression ratio is not a fixed value. This value can vary depending on the data that is transmitted. In the diagrams for the bandwidth estimation in Figure 3-3 on page 38, you can see that a compression ratio of 2 has been assumed. In implementations for customers, this value has commonly been achieved.

> **Tip:** When sizing the links between both sites, do not assume that a compression ratio is too optimistic. The effect of compression might offer some buffer of bandwidth capacity.

3.5 DS8000 configuration considerations

The following list summarizes some considerations about the host adapter setting and configuration:

► Always isolate host connections from Remote Copy connections (MM, GM, z/GM, GC, and MGM) on a host adapter basis. Isolate CKD host connections from FB host connections on a host adapter basis.

► Always have symmetric paths by connection type (that is, use the same number of paths on all host adapters that are used by each connection type). For z/OS, all path groups should be symmetric (that is, have a uniform number of ports per HA), and spread path groups as widely as possible across all CKD HAs.

► When possible, isolate asynchronous from synchronous copy connections on a host adapter basis.

► When possible, use the same number of host adapter connections (especially for z Systems) as the number of connections that come from the hosts.

► Size the number of host adapters needed based on expected aggregate maximum bandwidth and maximum IOPS (use Disk Magic or other common sizing methods that are based on actual or expected workload).

► For optimal performance with 2 Gb and 4 Gb HAs, avoid using adjacent ports by using the information in Table 3-2. For 8 Gb HAs, the port order is less important, except that for 8-port cards, it is preferable to use ports 1 - 4 before ports 5 - 8.

4

Performance tuning

This chapter discusses enhancements and tuning aspects.

Generally speaking, Global Mirror is an autonomic solution that provides 3 - 5 second recovery point objective (RPO) with no impact to production workloads. However, some tuning might be required in special cases, mainly in very long-distance implementations. This chapter also presents Global Mirror tuning for implementations at longer distances and when available data replication bandwidth is low.

This chapter includes the following sections:

► Global Mirror/GlobalCopy data synchronization
► Managing peak activity
► Bandwidth reduction
► Global Mirror tuning

4.1 Global Mirror/GlobalCopy data synchronization

With DS8000 microcode release 7.4, the data synchronization process was redesigned to improve remote copy services in general, including Metro Mirror and Global Copy technologies. These improvements are particularly important when Global Mirror replication falls behind and must catch up, or when residual data of a recently formed consistency group is being sent to secondary disk systems. RPO is directly positively impacted by these improvements because synchronization is key to how fast Global Copy data transmission is done during both Global Mirror catch up and consistency group formation processes.

The objective is to finish the primary-to-secondary data transmission as quickly as possible. In addition to multi-target support, the current design, including new internal algorithms, has the following characteristics:

► Minimum impact to production workload by not over driving the back end of primary disk systems. Data transmission workload is balanced across PPRC ports, extent pools, device adapters, and ranks.

► Prioritizing synchronization by doing more important tasks first to prioritize consistency group formation of a Global Mirror session over any other data synchronization potentially being run by other Copy Services task.

► The number of copy agents by volume being synchronized scales with volume size. Volume extent is now the unit of work for internal copy process, meaning that multiple extents of a single volume can be simultaneously transmitted during the data synchronization process.

4.2 Managing peak activity

Global Mirror preserves primary disk systems performance at the expenses of RPO. Therefore, you can deliberately underestimate the bandwidth without causing any performance impact at the primary site. If there are write peaks during short periods of time and the RPO values are acceptable, Global Mirror can be implemented in environments with relatively low bandwidth or when network cost takes priority over lower RPO requirements.

Figure 4-1 shows a typical production workload profile with a relatively low write rate during the online period and significant peaks at various points overnight. A bandwidth of at least about 15 MBps must be provided if low RPO is required during the whole period.

However, if higher RPO is acceptable when high write activity occurs during the overnight period, then you can configure as little as 8 MBps of bandwidth. This setting reduces the network requirements by around 47 percent.

The minimum bandwidth that can be provided must allow for consistency groups formation during at least some periods of the day. It also must allow the environment to catch up after any significant delays. Sizing tools can determine the minimum bandwidth based on RPO requirements.
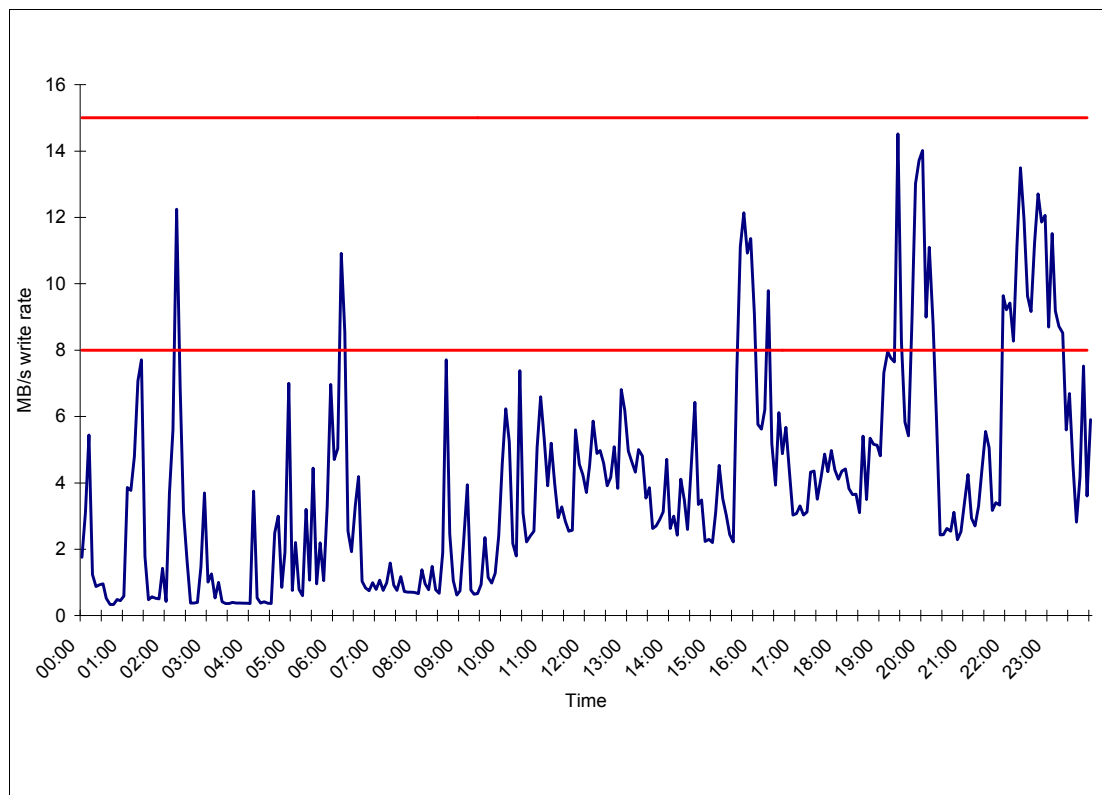

Figure 4-1   Production workload profile and Global Mirror bandwidth options

4.3 Bandwidth reduction

Because Global Mirror removes duplicate updates within a consistency group before sending it to the secondary location, less data is expected to be transmitted in comparison with data that was updated by application servers on primary disk systems. The amount of savings depends on the workload and the interval between consistency group formations.

Figure 4-2 shows the number of writes in MBps sent by Global Mirror to secondary disk systems and the related production write activity generated by application servers at the primary site in a specific period. The blue line in the graphic shows the percentage of the application servers write activity that is sent to secondary disk systems. For readability purposes, workload activity was sorted from high to low.

This example shows no bandwidth constraint, and the RPO must be around a few seconds. Even in this case, you can see that a considerable percentage of write activity from application servers does not need to be transmitted to secondary disk systems.
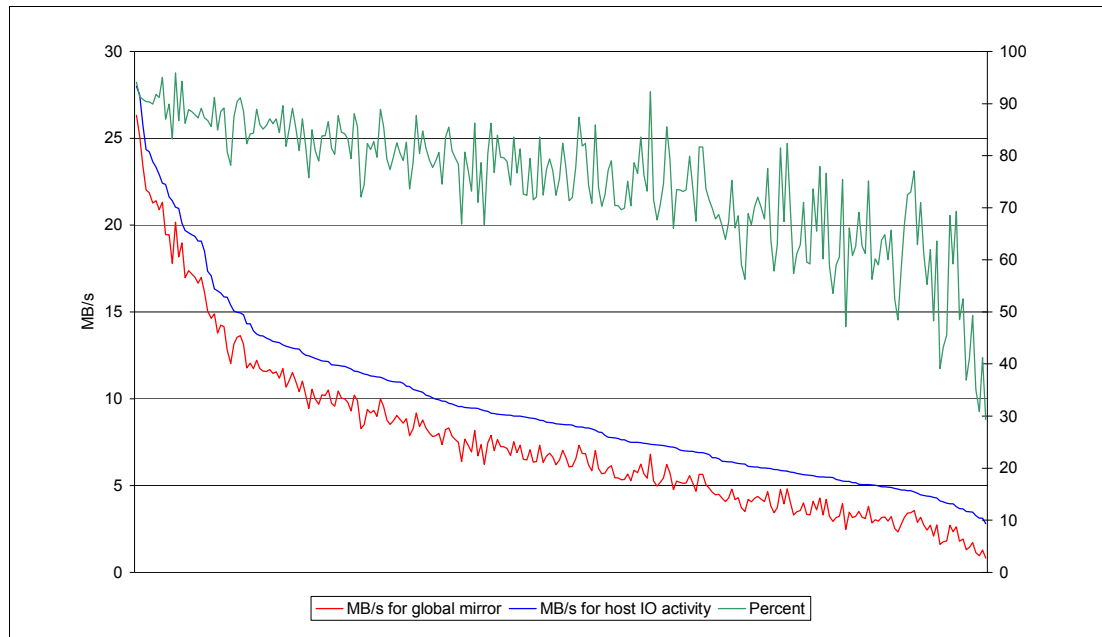

Figure 4-2   Comparison between Global Mirror data transmission and Write activity in primary disk subsystems.

Note in Figure 4-2 that the higher the write throughput in MBps, the higher the percentage of primary site write activity that must be sent to secondary disk systems. High write throughput usually denotes preponderant sequential write activity that usually does not change within a short period. As a result, most of the primary site write activity must be sent to the secondary site.

The opposite behavior appears on the right portion of the graphic, which shows lower percentages of write activity at the primary site being transmitted to the secondary site. This figure assumes that the predominant workload pattern is more randomized with small write updates mostly occurring within the same logical track.

4.4 Global Mirror tuning

Global Mirror meets RPO requirements without requiring any tuning in almost all implementations. Although the goal is to achieve the lowest RPO as possible, RPO values depend on distance and available bandwidth, assuming other required resources are correctly sized. Some tuning might be required in very long-distance implementations or in environments with low bandwidth.

This section describes Global Mirror tuning that might be required to achieve the lowest RPO as possible in an environment.

4.4.1 Global Mirror externalized parameters

As described in Chapter 1, "Global Mirror overview and architecture" on page 1, Global Mirror externalizes three parameters. These parameters can be changed to adjust the consistency group formation mechanism based on environment characteristics like distance between primaries and secondaries disk systems, available bandwidth, link characteristics, and workload:

► **Maximum coordination time**: To form a consistency group, Global Mirror serializes all primary volumes of all primary disk systems that are participating in a Global Mirror session. The design point for coordination time is 2 - 3 milliseconds. The default value is more than enough to coordinate the serialization among thousands of volumes across up to 17 primary disk subsystems with one master and 16 subordinates. This is the Global Mirror architectural limit. Therefore, leave the maximum coordination time as default.

► **Maximum drain time**: This is the maximum time that is allowed to drain the residual data of a consistency group to save that consistency group at the remote site. When drain time is not enough to allow the forming consistency group to be sent to secondary disk systems, that consistency group formation fails. After this failure, the RPO tends to be high for some time.

Increasing drain time value can mitigate, and in some cases eliminate, unsuccessful consistency groups formation attempts. In scenarios where temporary workload spikes are observed or when replication links are healthy but presenting high latency caused by either link technology, link quality, or long distances, the maximum drain time parameter value can be increased. A preliminary evaluation should be done, but increasing the maximum drain time value should not cause any negative impact, and can help to achieve lower RPO.

► **Consistency Group interval time**: This parameter determines how long to wait for the next consistency group to start. In most implementations where the lowest RPO is required, this parameter should be left as default. However, when required RPO is measured in hours, this parameter can be adjusted to allow the next consistency group formation to happen according to the desirable RPO.

For example, an intercontinental distance implementation requires an RPO of around 8 hours. Adjusting the consistency group interval time to achieve an 8-hour RPO lets Global Copy send data to the remote site in the most efficient manner and without any Global Mirror consistency group formation attempt during nearly 8 hours. As an initial setup, the consistency group interval time should be set to a value that is the required RPO minus the maximum drain time. For example, if the required RPO is 8 hours or 480 minutes and the maximum drain time is set to 10 minutes, the consistency group interval time should initially be set to 28,200 (480 minutes - 10 minutes = 470 minutes (28,200 seconds). Smaller adjustments can then be made as required.

4.4.2 Pokeables or internal switches

DS8000 pokeables, also known as internal switches, are internal controls that are externalized by DS8000 user interfaces for informational purposes only. Therefore, they cannot be changed by any user interface.

IBM technical support must be engaged to evaluate and change any internal switch as required. Some of these internal switches determine levels of parallelism when data is transmitted by Global Mirror or Global Copy. Default settings do not need to be tuned except for longer distances implementations or low-bandwidth scenarios where RPO is greater than expected, and one of the three situations is encountered:

► The total write throughput to be transmitted is higher than 1,600 MBps.

► Latency is higher than 40 ms and the total write throughput to be transmitted is higher than 1,000 MBps.

► The available bandwidth is lower than 100 Mbps.

4.4.3 Extreme distance tuning

At very long distances, mainly when high-bandwidth networks are involved, the default settings for Global Mirror and Global Copy need to be changed to allow for optimal performance. The set of internal tuning switches is referred to as *Global Mirror Extreme Distance RPQ*, and can be changed only by IBM technical support.

To achieve optimum parallelism in data transmission at extreme long distances, the following switches can have their default values changed, based on the environment and workload characteristics:

► TCB Usage: Controls the allowed number of updates being sent in parallel by each DS8000 main controller (server).

► Group size: Controls the number of updates that can be sent in parallel for each copy services process.

► DA Limit: Limits the number of tasks that can be running for each DS8000 device adapter.

► HA Port Limit: Limits the number of tasks that are running for each DS8000 host adapter port.

► Control Command Limit: Controls the number of FlashCopy commands that are issued in parallel for each DS8000 main controller (server).

► Port BW Max Limit: Controls the maximum bandwidth for each host adapter port.

These internal switches control other aspects of Global Mirror/Global Copy behavior and might have their defaults changed as required, generally in longer distance implementations:

► Volume Synchronization Delay: Sets the minimum time between successive scans of the out-of-synch tracks for a volume. When primary disk systems are running microcode release 7.3 or lower, which does not have Global Copy collisions enhancement, this value can be increased for longer distances to reduce potential impact in the productive environment due collisions.

► GM Polling Target: Controls when Global Mirror decides to form a consistency group when falling behind. This value indicates how optimistic Global Mirror is when deciding to form a consistency group when falling behind.

4.4.4 Path and port configuration for optimal data transmission.

If ports that are assigned for PPRC paths are shared between the two DS8000 main controllers (Server 0 and Server 1), then Server 0 and 1 interfere with each other during data transmission. One of the two servers will decrease its data transmission in a PPRC shared port if it detects that the other server is sending data to the remote site. This adjustment is made so that each server does not overdrive that PPRC port.

According to the DS8000 architecture, volumes are logically connected to logical subsystems (LSSs), and LSSs are owned by only one of the two DS8000 internal servers. Volumes in even LSSs are managed by Server 0, and volumes in odd LSSs are managed by Server 1. Because transmission paths are created on a per-LSS basis, it is better to split logical paths for even and odd LSSs over different ports. Splitting paths that way ensures that both Server 0 and Server 1 have redundancy across DS8000 I/O enclosures, host adapters, SAN components, and external links.

Figure 4-3 illustrates four replication ports of two host adapters, supposedly installed in different I/O enclosures. Logical paths dedicated for even LSSs are managed by DS8000 Server 0, and odd LSSs are managed by DS8000 Server 1 over different ports while keeping redundancy across host adapters and two storage area networks. This configuration assumes that each SAN is connected to dedicated and different external links.

As shown in Figure 4-3, the correct cabling and logical configuration would be to have each port of each host adapter being connected to a different SAN. Path A and Path D would be dedicated for LSSs owned by one server, in this case Server 0. The remaining Paths B and C would be created for LSSs owned by the other server, Server 1.
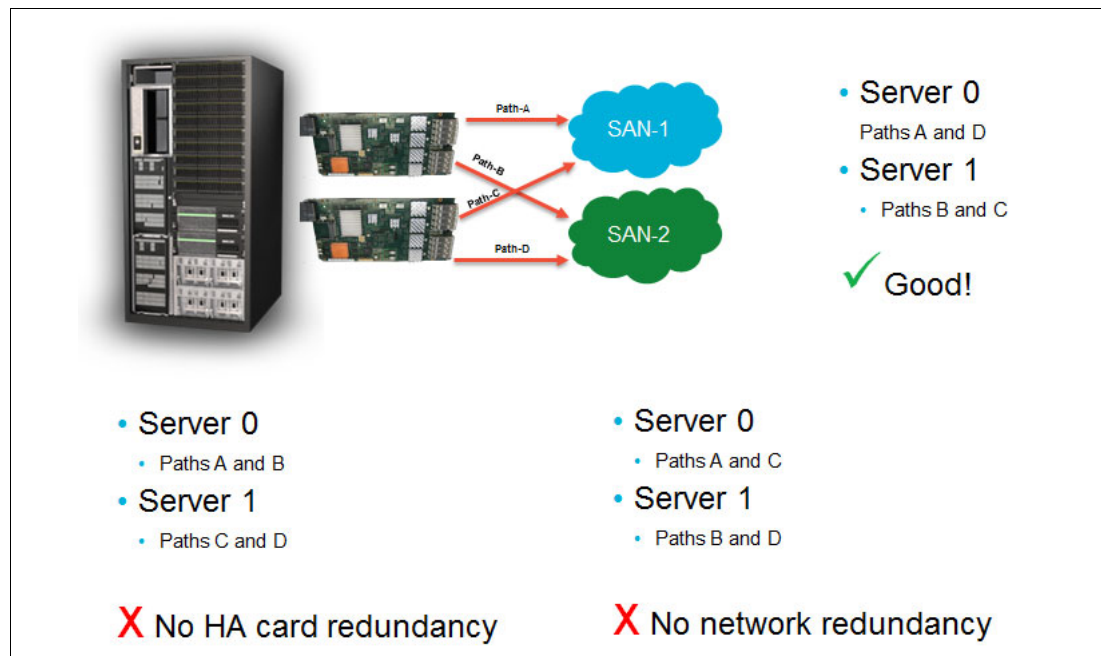

Figure 4-3   Splitting logical paths over different ports while maintaining redundancy for maximum data transmission throughput.

4.4.5 Environments with very low bandwidth

Implementations with very low bandwidth tend to present excessive retransmissions and retries due to the Peer-to-Peer Remote Copy (PPRC) internal response time threshold being reached. Any data transfer with latency over the threshold limit is assumed by DS8000 as a sign of congestion in the network or a saturation at the secondary disk system.

With very low bandwidth, external links are quickly saturated, causing transmissions with high response time. If the internal PPRC target response time threshold is reached, then a potential network congestion is detected that makes the data transfer dramatically decrease to nearly zero. This change eliminates the network congestion. Because the network congestion is not detected anymore, data transmission throughput increases until the threshold limit is reached again. As a result, the transmission throughput pattern is repeated up-and-down cycles like the sawtooth graphic that is shown in Figure 4-4.

By reducing the value of the PPRC response time target internal switch, this effect is attenuated, as can be observed in the same graphic in Figure 4-4. With this internal switch tuning, the available bandwidth is more efficiently utilized as retries and retransmission rates are reduced.
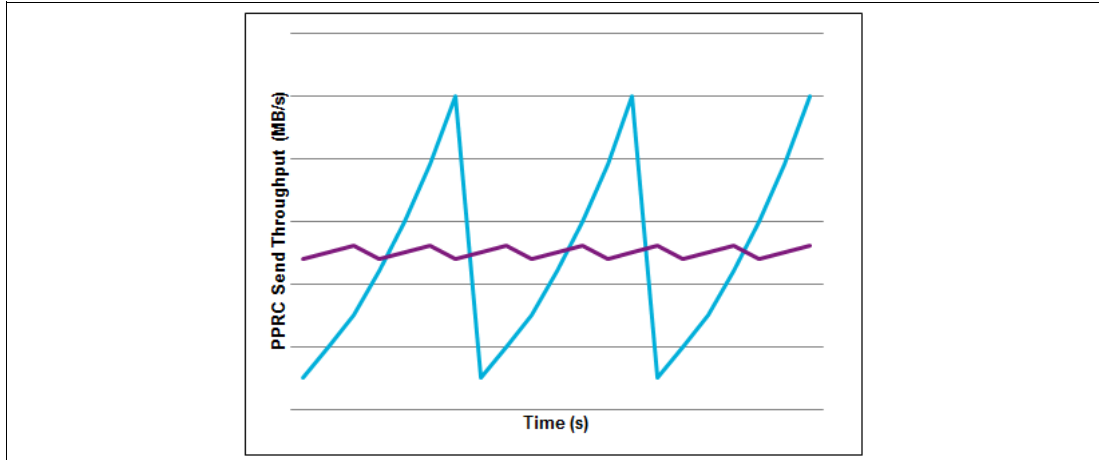
Figure 4-4   Transfer pattern in very low-bandwidth implementations

Global Mirror recovery

This chapter presents a high-level overview of the recovery process for a Global Mirror environment.

This chapter includes the following sections:

► Taking an additional copy for Disaster Recovery testing
► General recovery principle
► Autonomic behavior

5.1 Taking an additional copy for Disaster Recovery testing

One of the common requirements for disk replication design, as part of an overall Disaster Recovery (DR) solution, is the ability to practice DR tests at the remote site while production data is still being replicated to the DR site. With Global Mirror, this operation is achieved by taking an additional FlashCopy of the latest consistent data at the remote DR site. Moreover, this approach gives the opportunity of taking regular additional copies, perhaps once or twice a day, for other purposes.

The diagram in Figure 5-1 shows the entire process.
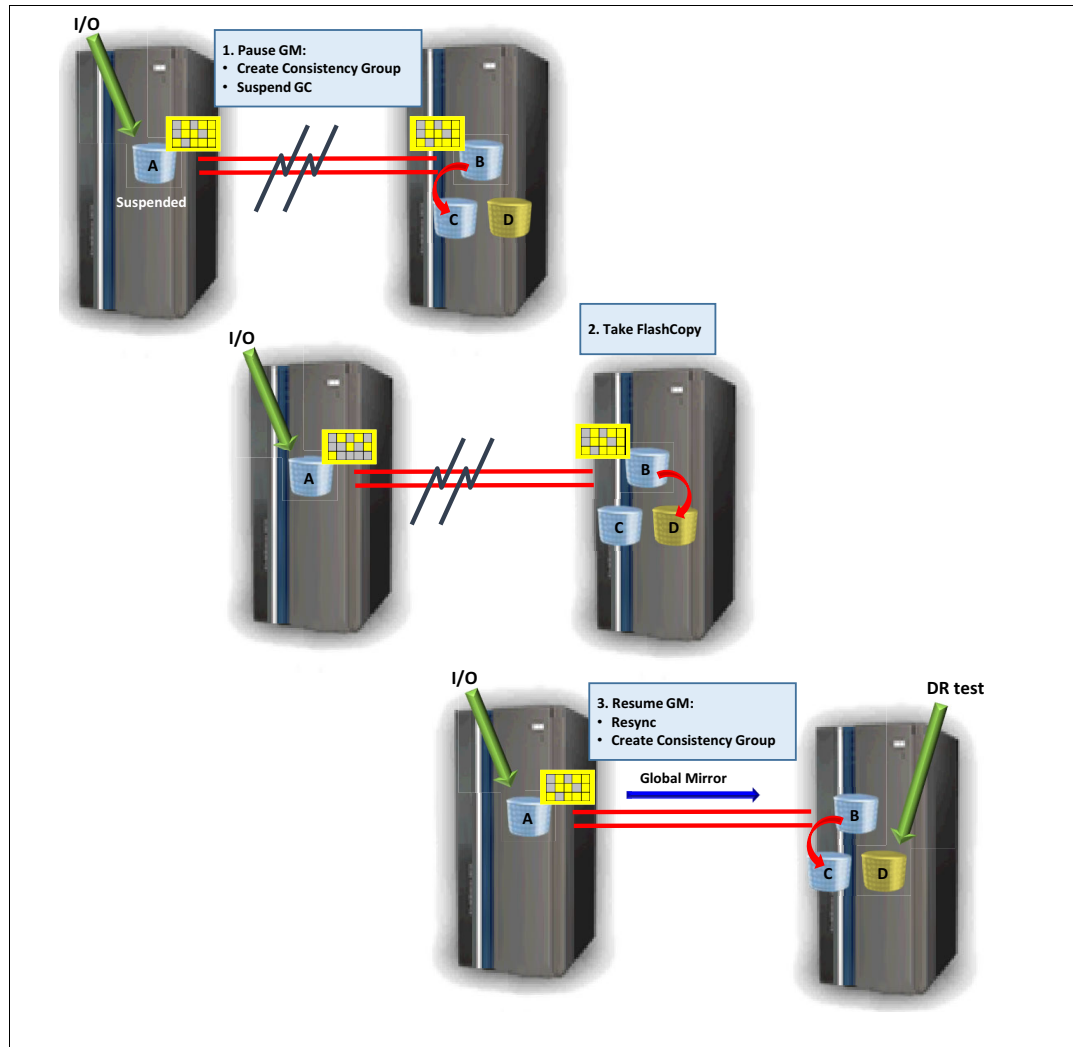
Figure 5-1   Process to take an additional copy for DR testing

To take a consistent copy of data at the DR site, the GM session must be paused.

The *GM pause With Consistency* function provides a simple method for obtaining a consistent copy of data at DR site. As soon as the GM pause command is issued, the new consistency group is created and Global Copy pairs are immediately suspended to stop any further updates coming from the primary (production) to the secondary volumes.

After the consistency group is created, the session state changes to *paused*. The primary Global Copy volumes change to *suspended* state, which means all further updates to these volumes are recorded into the bitmap.

Now that the GM session is paused with a consistent copy of data, the FlashCopy for DR test can be taken.

As soon as the FlashCopy relationship is established, you can resume the GM session. Resume starts resynchronizing Global Copy pairs and, according to GM design, creates a new consistency group.

The following commands and parameters should be used with respective interfaces when pausing GM with consistency:

▶ DS CLI: With `pausemgir`, use the following parameter: `-withsecondary`
▶ TSO: `RSESSION` command with `ACTION CGPAUSE`
▶ GDPS supports GM pause with consistency (APAR PM65428)
▶ Copy Services Manager support GM pause with consistency (starting with Copy Services Manager, formerly known as Tivoli Productivity Center for Replication, Release 5.1.1.1)

5.2 General recovery principle

The Global Mirror general recovery scenario can be subdivided into two general operations:

1. First, the underlaying Global Copy must be failed over from the remote site to the local site.

2. The FlashCopy relationship must be reversed from the FlashCopy target volumes to the FlashCopy source volumes.

By reference to Figure 5-2, the term *H1* is used for the local or primary volumes, *H2* for the remote or secondary volumes, and *J2* for the journal or FlashCopy target volumes.
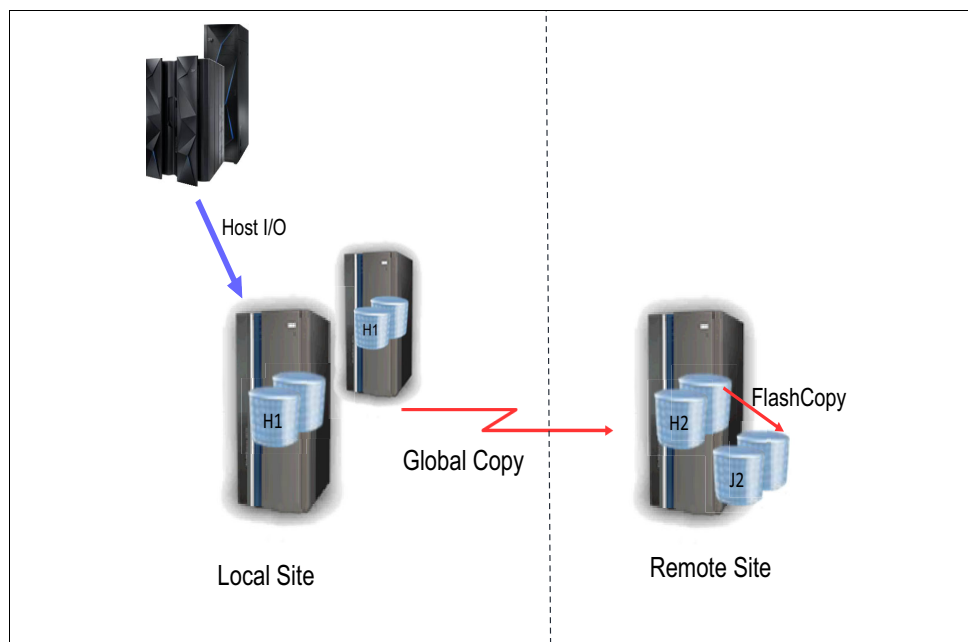

Figure 5-2   Global Mirror layout

5.2.1 Planned recovery scenario

Possible reasons for planned recovery scenarios might be to test the Global Mirror consistency or when the safety of the production is jeopardized because of physical maintenance operations at the primary site. In those cases, the production is failed over to the remote site.

The first step in a planned recovery scenario is to stop the Global Mirror immediately before performing the recovery. The goal is to preserve data consistency at the remote site.

Next, the underlaying Global Copy is turned around from H2 to H1 by using a `failoverpprc` DSCLI command or the `CRECOVER` TSO command. The H2 volumes are now ready for host access, but they still contain inconsistent data. To provide consistent data to the H2 volumes, the FlashCopy must be reversed from the Jx volumes back into the H2 volumes. The host can now access the data from H2.

5.2.2 Unplanned recovery scenario

Unplanned recovery scenarios must be performed when disaster strikes at the primary site or a serious outage of the data center infrastructure impacts the production so that the production cannot be continued at the primary site.

FlashCopy recovery stages

In an unplanned recovery scenario, check the status of the last saved consistency group. The status can be obtained by querying the FlashCopy relations and inspecting the status of the revertible bit and the sequence number for each FlashCopy relation. Figure 5-3 shows the different phases with the various possible status descriptions. If sequence numbers are all equal and all revertible bits are not set, the FlashCopy operation was not impacted and no further actions are required. However, in all other cases, the FlashCopy relation requires manual intervention as follows:

1. When the sequence numbers are different and the revertible bits are equal or different, the last FlashCopy operation was not completed for all FlashCopy pairs. In this case, the FlashCopy target volumes Jx does not provide a consistent state of data. Therefore, the previous consistency group must be restored by using a `revertflashcopy` command.

2. When the sequence numbers are all equal and there is a mix of revertible and non-revertible volume pairs, that data is already on the FlashCopy target volumes Jx. In this case, the FlashCopy operation must be committed to finalize the most recent FlashCopy operation manually.
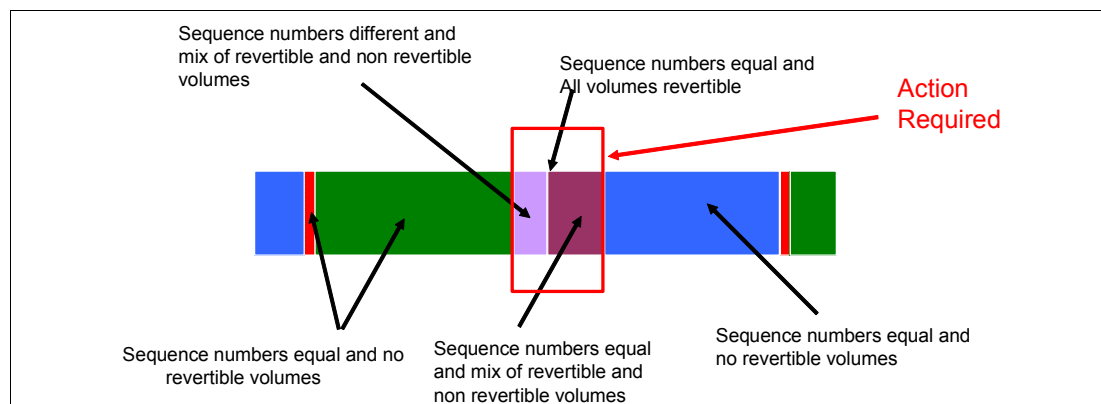

Figure 5-3   Parts of consistency group formation process where recovery action is required

The second stage is to recover the environment and enable production systems to be restarted on the H2 volumes. Then prepare for a potential return to the primary site. This recovery is performed according to the following process:

1. Fail over the H2 volumes. This action places the H2 volumes in a primary suspended state. It allows for a resynchronization of the Global Copy relationship to be performed to return to the primary site, assuming the primary disk system has survived.

2. Fast Reverse Restore the FlashCopy relationship with the Jx volumes. This action restores the latest consistency group to the H2 volumes and starts a background copy for those tracks that have been modified since the latest consistency group.

3. FlashCopy from the H2 volumes to the Jx volumes to save an image of the last consistency group. This step is optional. It preserves an image of the production devices at the recovery point in case this might be required.

4. Restart production systems.

5.3 Autonomic behavior

Global Mirror is designed to handle certain conditions such as a loss of connectivity automatically, without requiring user intervention.

5.3.1 PPRC paths

If PPRC paths are removed unexpectedly for some reason, then the disk system automatically reestablishes these paths when the error situation is resolved.

5.3.2 PPRC pairs

If Global Copy pairs suspend for some reason other than a user command, then the disk system automatically attempts to restart the mirroring for these pairs. As the consistent set of disks is the FlashCopy secondary devices, this process does not compromise the integrity at the secondary site. This feature is different from Metro Mirror where the resynchronization is always by using command because the Metro Mirror secondaries are the consistent devices.

Tip: You can disable this behavior, if wanted.

5.3.3 Global Mirror session

After the Global Copy pairs are restarted and have resynchronized, Global Mirror resumes the formation of consistency groups unless doing so might result in inconsistent data on the secondary disk system.

One example of such a condition is where a communications failure occurs halfway through a FlashCopy event. In this case, you must perform a revert/commit action. The Global Mirror session will have entered what is called a "Fatal" state.

Topologies and solution scenarios

Different Global Mirror topologies can be considered when planning for Global Mirror implementation.This chapter describes a number of scenarios that can be used for setting up a Global Mirror configuration based on disaster recovery requirements.

The examples in this chapter use the following symbols:

► H: Volumes that are attached to hosts

► J: Journal volumes (FlashCopy)

► Numbers next to the volumes are site indicators. The following is an example for a two-site configuration:

– 1: Primary site (original production site)

– 2: Secondary site (original DR site)

This chapter includes the following sections:

► Asymmetrical configuration
► Symmetrical configuration
► Multiple Target PPRC with Global Mirror
► Metro Global Mirror
► Four-site configuration
► Data Migration example scenario

6.1 Asymmetrical configuration

With an asymmetrical configuration, Global Mirror can only be used from the primary site to the disaster recovery (DR) site. This type of configuration would be typical for a disaster recovery configuration where the production systems would run in the secondary location only during an unplanned outage of the primary location. As shown in Figure 6-1, each primary H1 volume is in a replication relationship with its associated H2 host volume and a corresponding J2 journal volume for creating consistency groups.
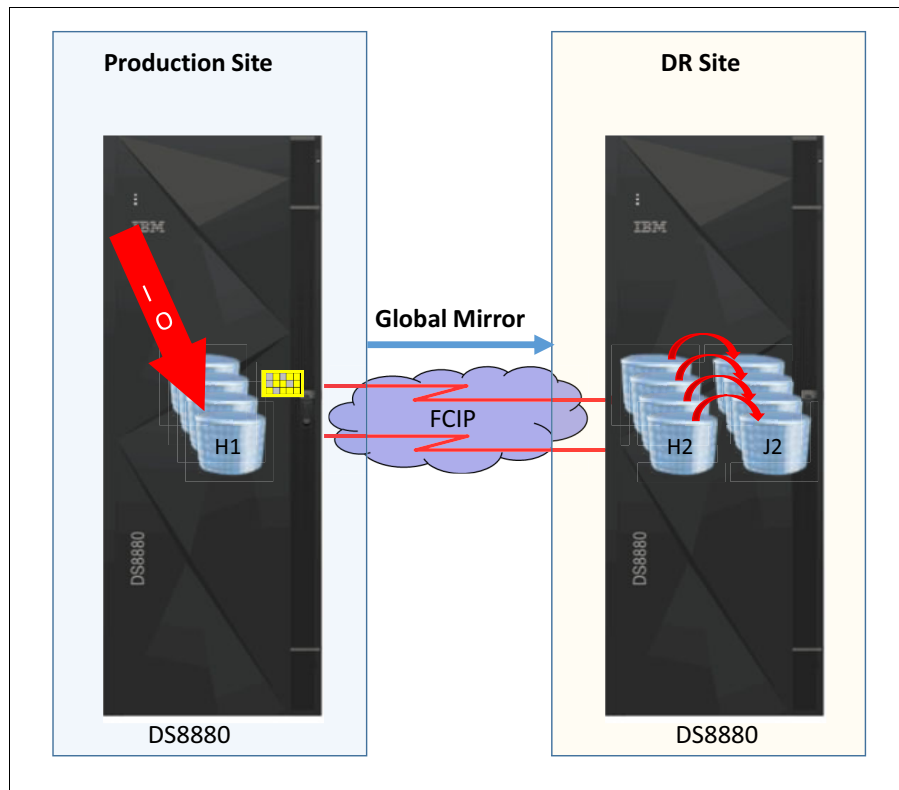
Figure 6-1   Asymmetrical Global Mirror configuration

Because Global Mirror uses two copies of data in the secondary (DR) location, there are twice as many physical drives in this location as in the production location if the same size drives are used.

> **Guideline:** The preferred practice is to use the same size and RPM for journal volumes. In some situations, it might be cost effective to use space efficient FlashCopy to reduce the total capacity that is required for Global Mirror configuration at the secondary DR site.

6.1.1 Return to primary site with asymmetrical Global Mirror configuration

After production workloads are moved to the recovery site, Global Copy must be used to return back to the primary site (see Figure 6-2). No disaster recovery capability is provided in the reverse direction. Therefore, it is unlikely that this type of configuration production would run for extended periods of time at the secondary location, unless forced to by unavailability of the primary site.
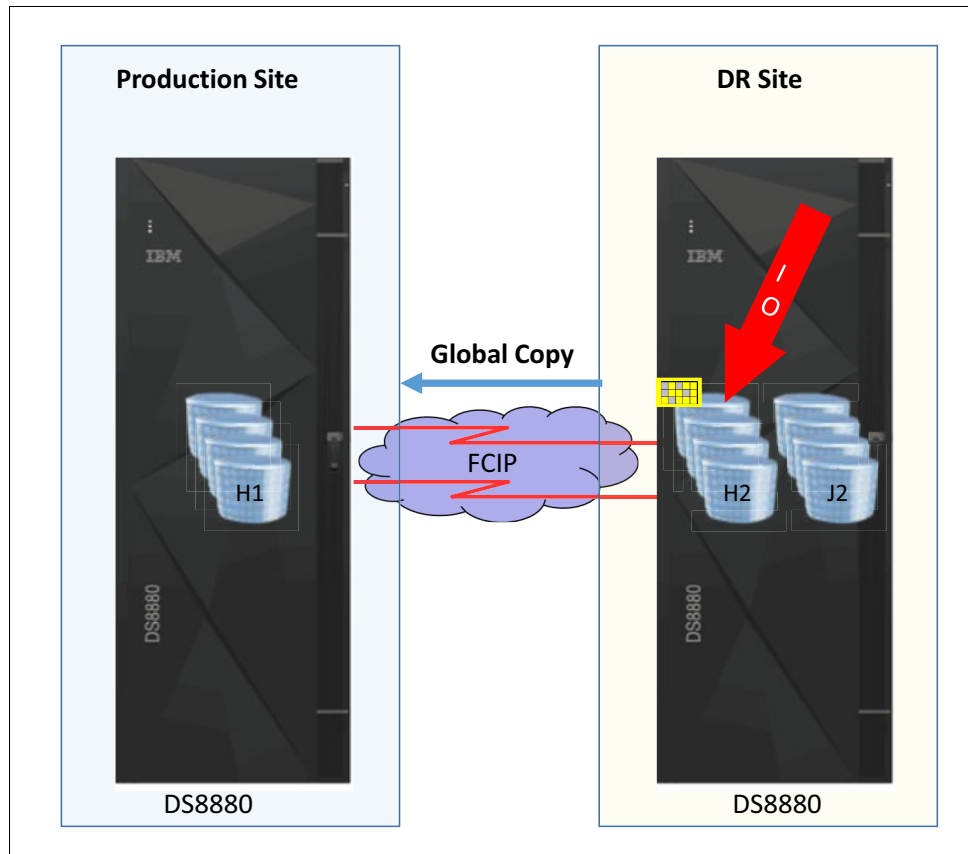

Figure 6-2   Return to primary site with asymmetrical Global Mirror configuration

Monitor the Global Copy status by using Copy Services Manager (CSM) or any command interface to determine how many tracks are out of sync between H2 and H1 volumes. After these out-of-sync numbers are static, you can stop I/O at the secondary DR site (shut down all systems) and wait until the out of sync is zero. You can then fail back to the H1 volumes at the primary site, start your systems off H1 volumes, and resume the Global Mirror session again from H1 to H2 volumes.

6.2 Symmetrical configuration

With a symmetrical configuration, additional disk capacity is also required for journal FlashCopy volumes at the primary site. Therefore, both primary and secondary disk systems have identical configurations, as shown in Figure 6-3 on page 60. The additional capacity at the primary site can also be used for regular FlashCopy operations. For example, the capacity can be used for backing up the data to disk and then dump to tape without extended outages for the production systems.
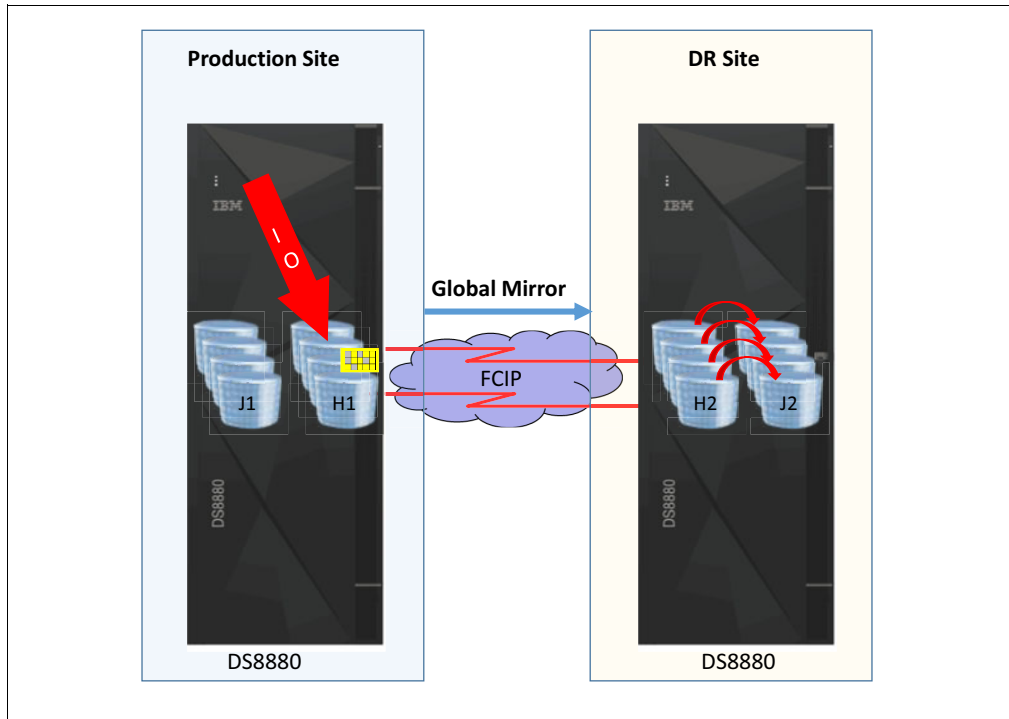
Figure 6-3   Symmetrical Global Mirror configuration

With FlashCopy capacity at both sites, it is possible to provide a disaster recovery solution using Global Mirror in both directions between the two sites. This type of configuration would typically be used where production workloads might run for extended periods of time in either location. In Figure 6-4, the Global Mirror session direction is from H2 to H1 volumes and the consistency groups are formed on journal J1 volume set.
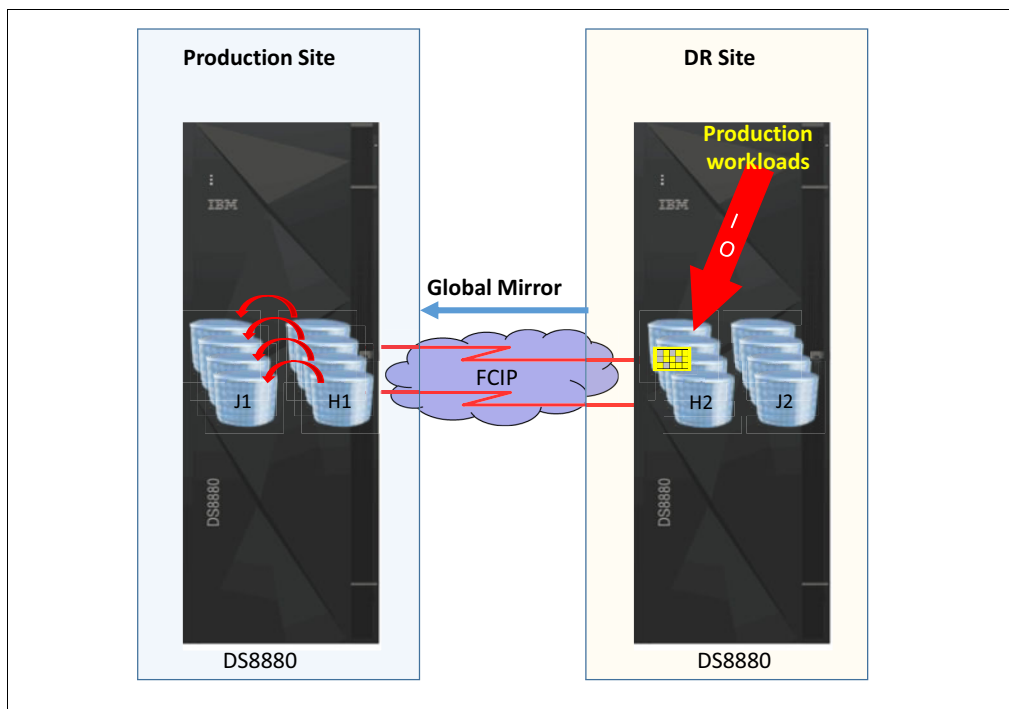

Figure 6-4   Running production workload in secondary location with symmetrical configuration

6.3 Multiple Target PPRC with Global Mirror

Different multi-site DS8000 copy services topology options are available when it comes to three or four site configurations. Multiple Target Peer-to-Peer Remote Copy (MT PPRC) is one of them. It allows you to have a single primary volume in a continuous copy services relationship with two target volumes. MT PPRC enhances capability and flexibility for disaster recovery and migration solutions by using synchronous, asynchronous, or a combination of both synchronous and asynchronous replications.

This section provides only a high-level description for Global Mirror option in combination with Metro Mirror. More details about MT PPRC can be found in *IBM DS8870 Multiple Target Peer-to-Peer Remote Copy*, REDP-5151.

6.3.1 Overview of a Metro Mirror and Global Mirror topology

The existing customers with two data centers within metropolitan distance (usually up to 100 km) might consider enhancing their DR topology with a remote third site. This site is usually located out of the region and beyond the supported distance for synchronous replication.

With MT PPRC configuration, data is synchronously mirrored to one secondary site and is asynchronously mirrored to a separate remote disaster recovery site (see Figure 6-5). When the primary DS8000 with H1 volumes detects that an MT PPRC configuration exists, it creates Multiple Target Incremental Resynchronization (MTIR) pairs between H2 and H3. The MTIR pairs serve two main purposes:

▶ They enable an active relationship to be quickly established between the two secondary volumes by converting the existing pair rather than establishing a new pair.

▶ They provide a change recording mechanism to track which data is potentially different between the two secondary volumes. This mechanism is what allows the resynchronization to be an incremental copy rather than a full copy.
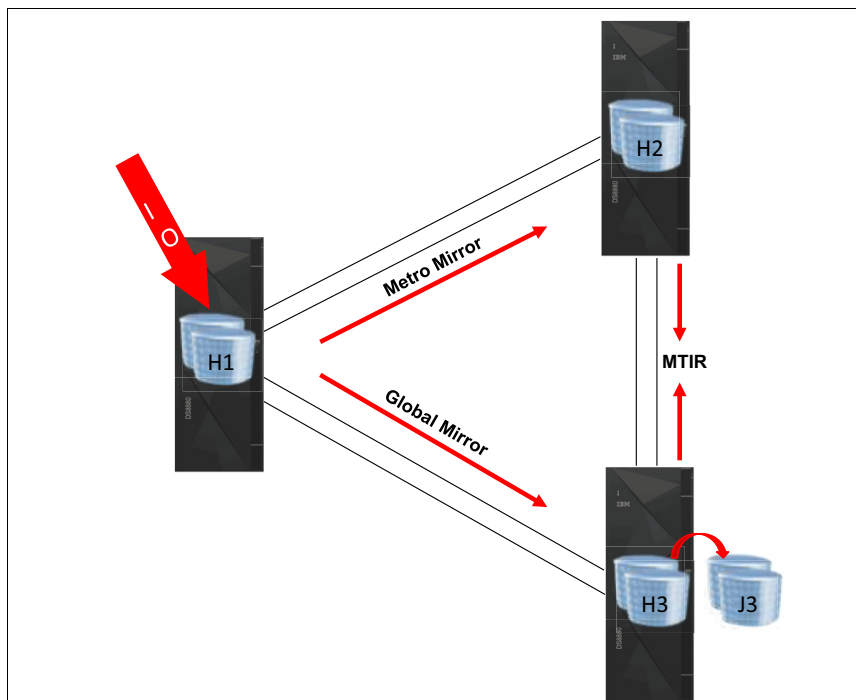

Figure 6-5   MT PPRC with Metro Mirror

During any planned or unplanned outage for H2 site, Global Mirror from H1 to H3 site still provides disaster recovery protection.

When H3 site encounters an unplanned or planned outage, the Global Copy relationship between H1 and H3 is suspended and Global Mirror stops forming consistency groups. The Metro Mirror replication continues to run and provides protection during H1 site failure.

During an H1 site outage, the production workload can fail over to the H2 site and continue replication between H2 and H3 site, thus maintaining DR capability.

Note: Use Copy Services Manager or Geographically Dispersed Parallel Sysplex (GDPS) when managing 3-site or 4-site replication topologies.

6.4 Metro Global Mirror

Metro Global Mirror provides a 3-site or 3-copy solution using both synchronous and asynchronous replication. This topology can provide a local synchronous copy of data either to another site within synchronous distance or within the same campus or data center.

Additionally, Global Mirror is used to continually mirror data from the Metro Mirror secondary devices, providing an out-of-region copy. As shown in Figure 6-6, H2 volumes are defined as secondary Metro Mirror volumes. These volumes are in a cascaded Global Mirror relationship with primary Global Mirror volumes that asynchronously replicate data to H3 volumes at the remote site. In this example, the Global Mirror session between H2 and H3 site volumes is asymmetrical. However, based on client's requirements, both symmetrical and asymmetrical Global Mirror configurations are supported.
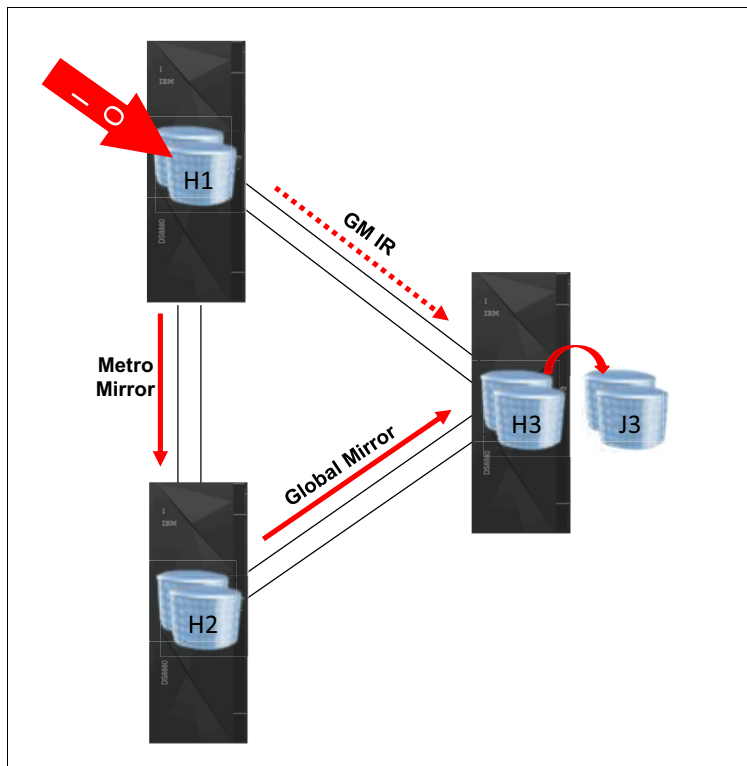

Figure 6-6   Metro Global Mirror configuration

During a H1 site outage, production workload is failed over to the intermediate H2 site and the Global Mirror session between H2 and H3 continues to provide disaster recovery protection.

Although the links between H1 and H3 are optional, the preferred practice is to always configure them. The advantage of having these links is beneficial when the intermediate H2 site fails. The MGM Incremental Resync function offers the capability to establish the Global Mirror relationship between the local H1 and remote H3 sites without needing to replicate all the data again. The MGM topology with Incremental Resync provides a more flexible and efficient disaster recovery protection.

When H3 site is not available, the H1 to H2 Metro Mirror configuration is not affected.

6.5 Four-site configuration

Customers with three or four data centers within and out of the region might consider a combination of Metro Global Mirror with Multiple Target PPRC (MT PPRC) as shown in Figure 6-7. Metro Global Mirror configuration between H1, H2, and H3 systems can be extended by adding a Metro Mirror session between H1 and H3 with MT PPRC. This MT PPRC configuration provides high availability and disaster recovery protection within metropolitan distances, while MGM extends data protection during metropolitan and regional disasters.
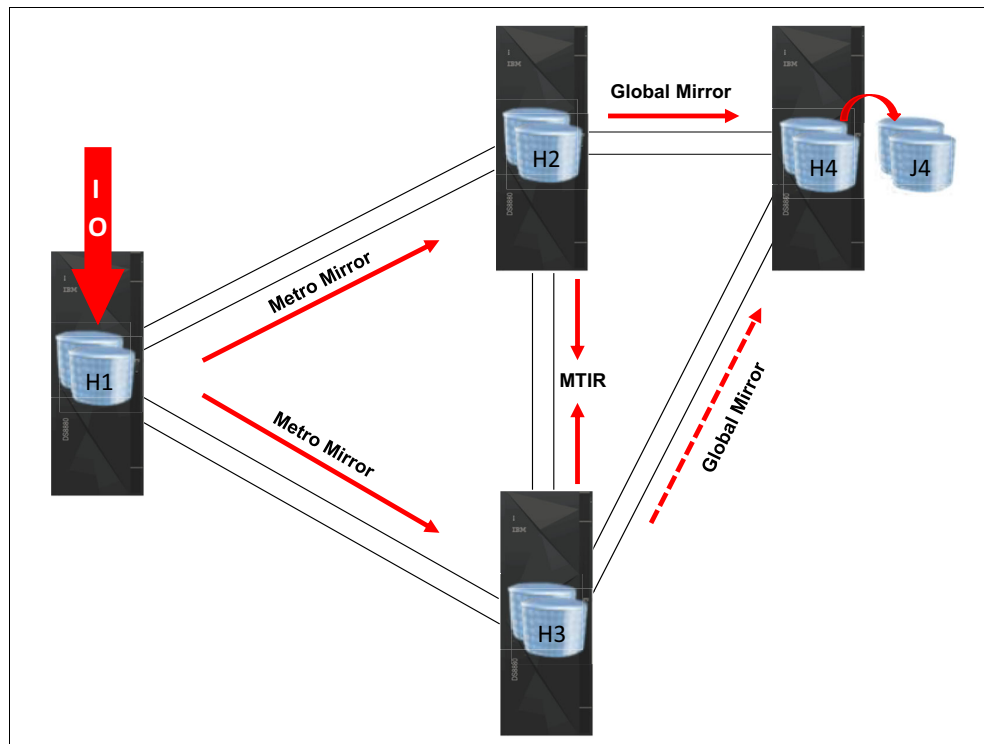

Figure 6-7   Four site configuration with MGM and MT PPRC

During a failure at site H1, the production applications can be moved to run at H3, and the Incremental Resynchronization capabilities of Multiple Target PPRC (MTIR) can be used to establish an active Metro Mirror relationship (H3 to H2). This situation results in an MGM configuration where there is Metro Mirror H3 to H2 and Global Mirror H2 to H4.

Even after a failure of the primary production site, there is still the full protection of an MGM environment where Metro Mirror H3:H2 provides a high availability capability and the Global Mirror H2 to H4 provides for long-distance disaster recovery.

In case the H2 site is not operational, there is an option to cascade H3 to H4 by forming a new MGM configuration.

6.6 Data Migration example scenario

This section describes the migration topology and process that is involved when replacing the old DS8870 systems in a Global Mirror session with the newer DS8880 disk models, as illustrated in Figure 6-8.

For the examples in this section, the following terms are used:

► H1 is the current primary site where the production applications are running.
► H1' is the new Global Mirror primary site that is replacing the current H1.
► H2 is the current Global Mirror secondary site to which H1 is mirroring data.
► H2' is the new Metro Mirror secondary site that is replacing the current H2.
► J2 is the current Global Mirror journal (FlashCopy) secondary site.
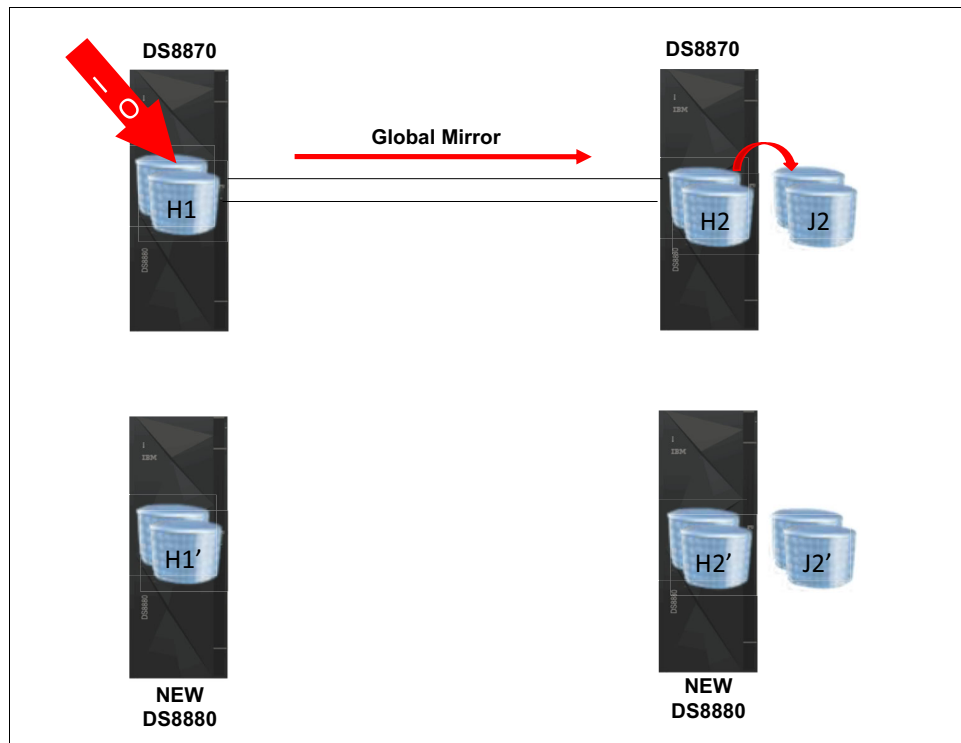► J2' is the new Global Mirror journal (FlashCopy) secondary site.

Figure 6-8   Replacing DS8870 systems in Global Mirror session with DS8880 systems

The proposed migration example to migrate data from a primary or secondary DS8870 storage system in Global Mirror session is MT PPRC. The use of MT PPRC allows for migration procedures with either few or no periods of time when the system is not protected by mirroring.

> **Note:** All DS8870 storage systems must have Multiple Target PPRC support and the correct license features.

For more information about DS8870 migration, see *DS8870 Data Migration Techniques*, SG24-8257.

6.6.1 Replacement of Global Mirror secondary DS8870 system

The general method of this migration is to use the MT PPRC capability to start Global Copy from the existing H1 primary site to the new H2' secondary site (Figure 6-9). Global Copy is asynchronous data replication, which does not use journal volumes to create consistency groups. Therefore, it is mainly used for data migration. However, the new DS8880 at the secondary site requires provisioned space for journal volumes that are used in the final migration step.
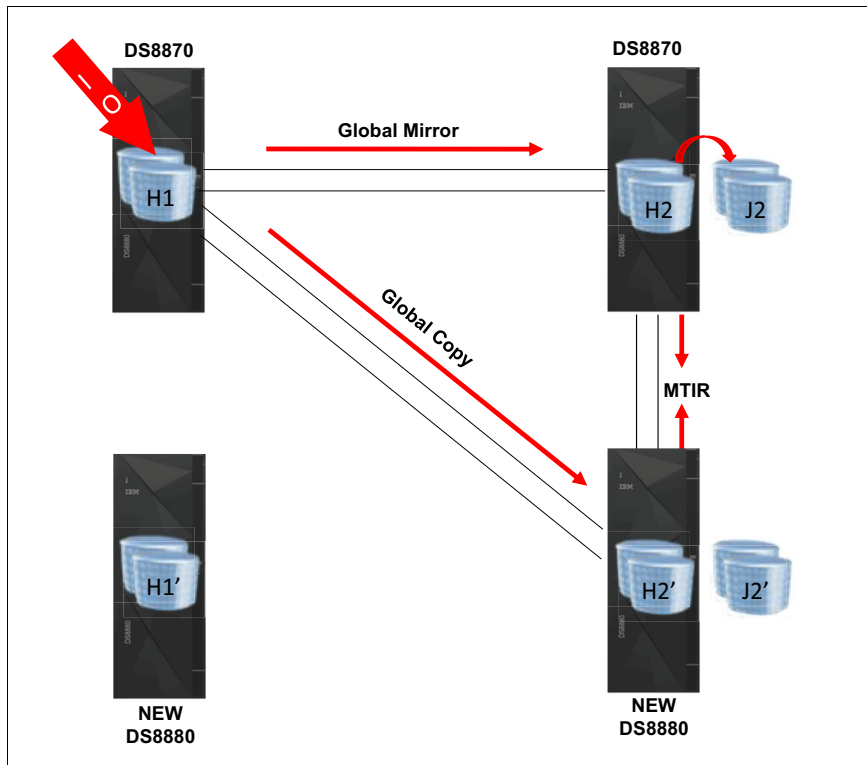

Figure 6-9   Replacing Global Mirror secondary DS8870 system

After all of the volume pairs H1 to H2' have passed the first round of copy (they never reach full sync/duplex state, but the out of sync data amount is minimal), the migration can start as follows (see Figure 6-10):

1. Remove H1 to H2 Global Copy pairs from the H1 to H2 Global Mirror session.
2. Delete Global Mirror session H1 → H2.
3. Add the Global Copy H1 → H2' pairs into a new H1 → H2' Global Mirror session.
4. Start the new H1 → H2' Global Mirror session.

The RPO increases during the conversion of H1 to H2' Global Copy into Global Mirror. The new consistency group will be created after the new Global Mirror session between H1 and H2' is started.
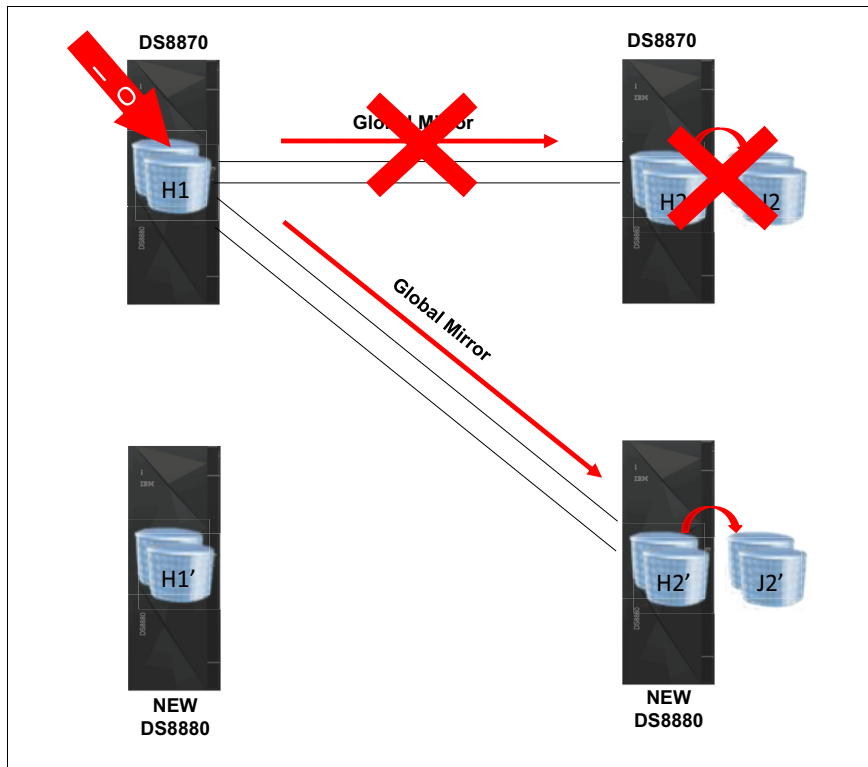

Figure 6-10   Start new Global Mirror session between H1 and H2' and remove H2

6.6.2 Replacement of Global Mirror primary DS8870 system

Similarly to replacing secondary DS8870 in a Global Mirror session, MT PPRC can be used to replace the primary Global Mirror DS8870 system. Because the new target DS8880 H1' volumes are in the local, primary site, the H1' volumes can be defined in a synchronous Metro Mirror session between H1 and H1'. As you can see in Figure 6-11 on page 67, the secondary DS8880 is already replaced (as described in "Replacement of Global Mirror secondary DS8870 system" on page 65) and the Global Mirror session is active between old H1 and new H2' volumes.

With z/OS and AIX, you can start the H1 to H1' Metro Mirror session with IBM HyperSwap® enabled. HyperSwap can transparently swap the production workload from H1 to H1' volumes, thus avoiding the applications and systems outage.
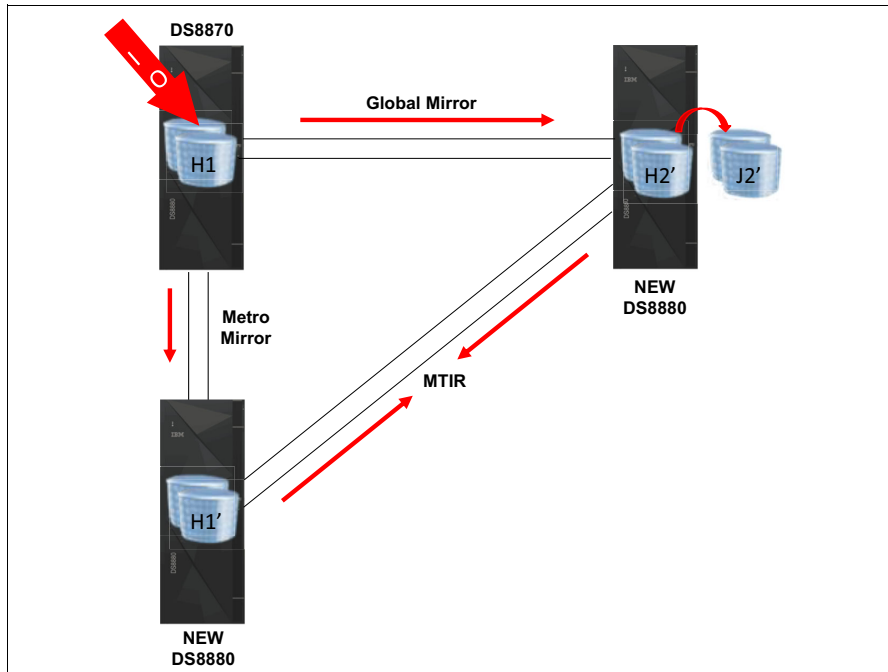
Figure 6-11   Replacing Global Mirror primary DS8870 system

After the H1 and H1' volumes are fully synchronized and in Duplex state, you can invoke HyperSwap if it is enabled, or alternatively perform a failover to the H1' volumes, by shutting down the workload to H1 volumes and starting it from H1' volumes. Thanks to the Multiple Target Incremental Resynchronization capability between the new H1' and H2' volumes, the new Global Mirror relationship is quickly started and the new consistency group created.

At this stage, you can terminate all relationships on H1 volumes and remove the old DS8870. The migration to H1' is now complete, as shown in Figure 6-12.
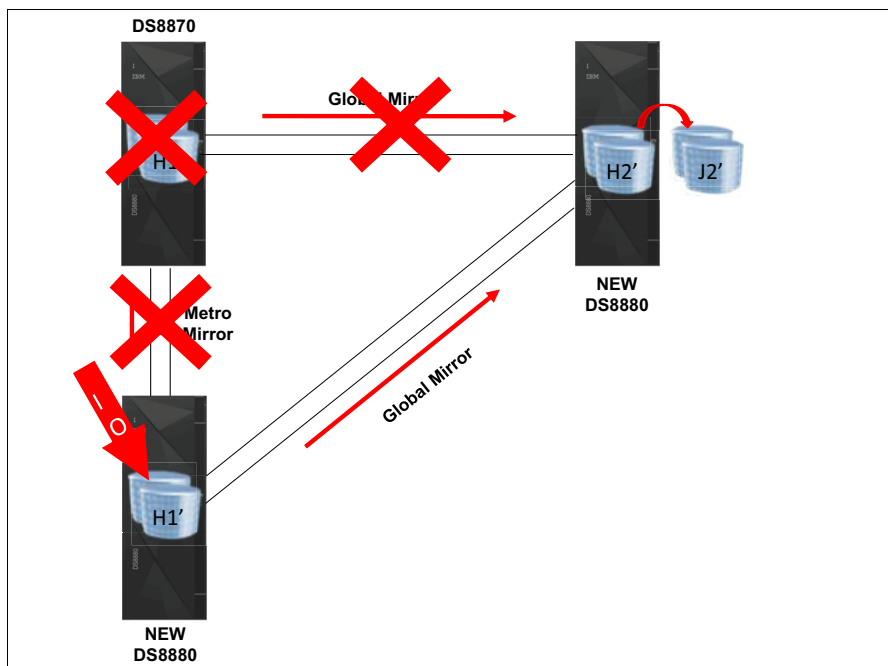

Figure 6-12   Start new Global Mirror session between H1' and H2' and remove H1

7

Global mirror management and disaster recovery solutions

Different interface options are available to configure, manage, and control Global Mirror environment. Each interface provides a set of commands that allows you to develop your own management, control, and automation solution as well as providing problem diagnosis capabilities for the Global Mirror environment.

However, the IBM disaster recovery and replication management solution offerings also have been extended to provide support for Global Mirror. These solutions provide different capabilities depending on the exact requirements for a particular situation. In most cases, if these solutions fit the requirements for the environment being used, these would be the preferred option as it provides a supported management solution without requiring code to be written specifically for each environment.

This chapter provides more information about the following IBM Global Mirror management solutions:

► Copy Services Manager (CSM), which was previously known as Tivoli Storage Productivity Center for Replication (TPC-R)

► Geographically Dispersed Parallel Sysplex (GDPS): IBM Service offering for mainframe only

► Power HA for IBM i systems

► VMware SRM

This chapter includes the following sections:

► Copy Services Manager
► GDPS Global Mirror (GDPS/GM)
► IBM System i PowerHA for i
► VMware SRM

7.1 Copy Services Manager

Copy Services Manager (CSM) is a remote copy management solution for disaster recovery. It supports various different environments, including Metro Mirror, FlashCopy, Global Mirror, Metro Global Mirror, and Multi Target Metro Mirror. CSM runs as an IBM WebSphere® application on different open systems and mainframe (z/OS) platforms. It uses a TCP/IP interface to the disk system to manage the Global Mirror environment.

Figure 7-1 shows two CSM servers configured in High Availability mode. The server at the production site is the active CSM server that controls and manages Global Mirror sessions, while the standby CSM server at the DR site is ready to take over control in case of any planned or unplanned outage of the active CSM server (or even complete site outage). All updates performed on the active CSM server are immediately replicated over TCP/IP to the standby CSM server (for instance, adding sessions or volumes to GM sessions).
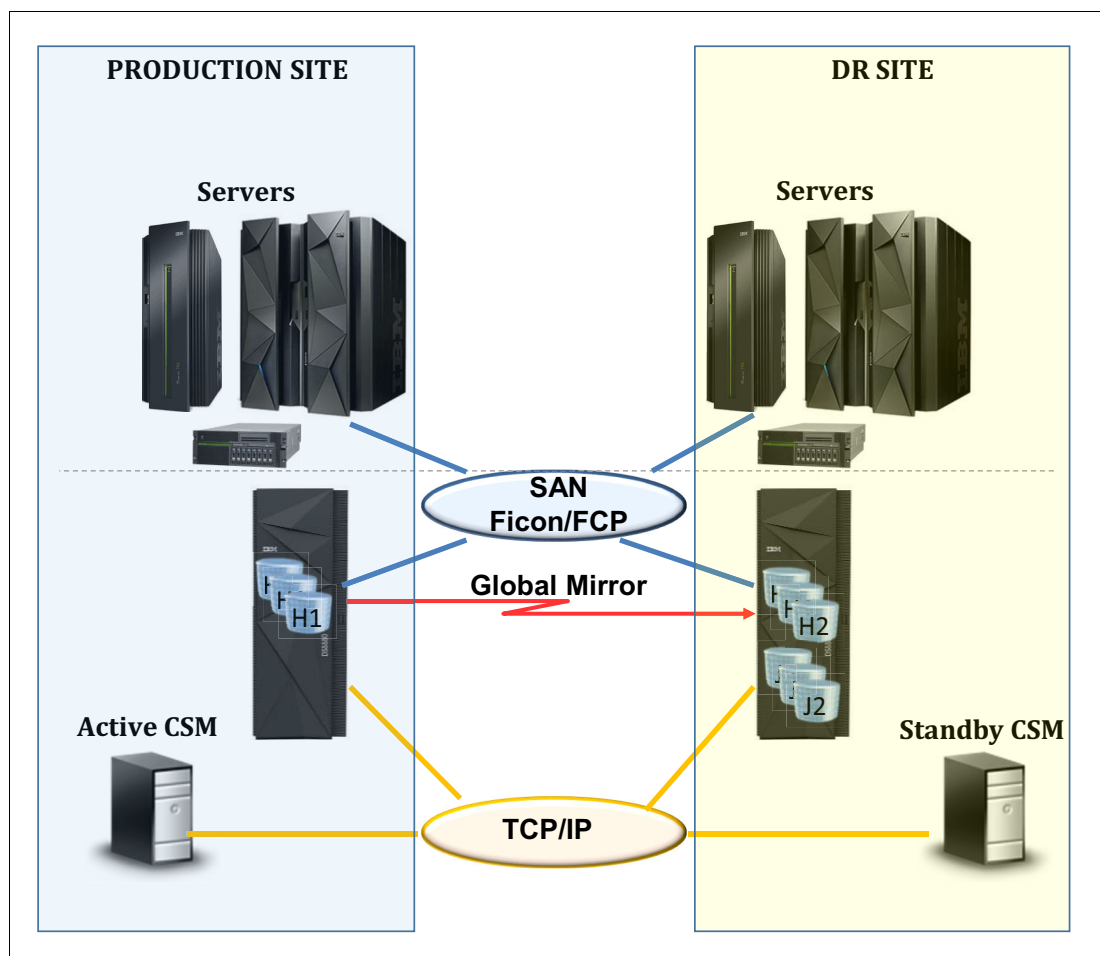

Figure 7-1   CSM environment

Global Mirror configurations can become complex, with thousands of relationships, practice and test copies, and numerous recovery scenarios. Managing all of these aspects through the storage system's GUI, DS CLI, TSO, or ICKDSF requires a large amount of effort and is prone to errors. With CSM, you can set up and manage complex and large scenarios with a few mouse clicks. For instance, setting up a GM session, practicing DR, and failover and failback scenarios during the planned or unplanned outage can be performed quickly and

under full CSM control. CSM is aware of different sites and it is simple to follow the wizard for initial configuration.

CSM supports a number of different configurations for Global Mirror including the creation of an additional testing copy for disaster recovery testing and the ability to return back to the production site after it has become available again. Without CSM, scripting all these sequences by using commands can be challenging for ongoing management and support.

The CSM GUI interface is easy to use. Figure 7-2 shows an example of the **Create Sessions** wizard. By selecting the **Global Mirror Failover/Failback with Practice** session and predefined sites, you can see a diagram of your Global Mirror configuration.

H volumes are always the volumes that are attached to your hosts. J volumes are journal volumes that are used by GM to create consistency groups and I volumes are intermediate volumes that are actually Global Copy secondary volumes. A number next to each volume letter is a site indicator of where the production application runs.
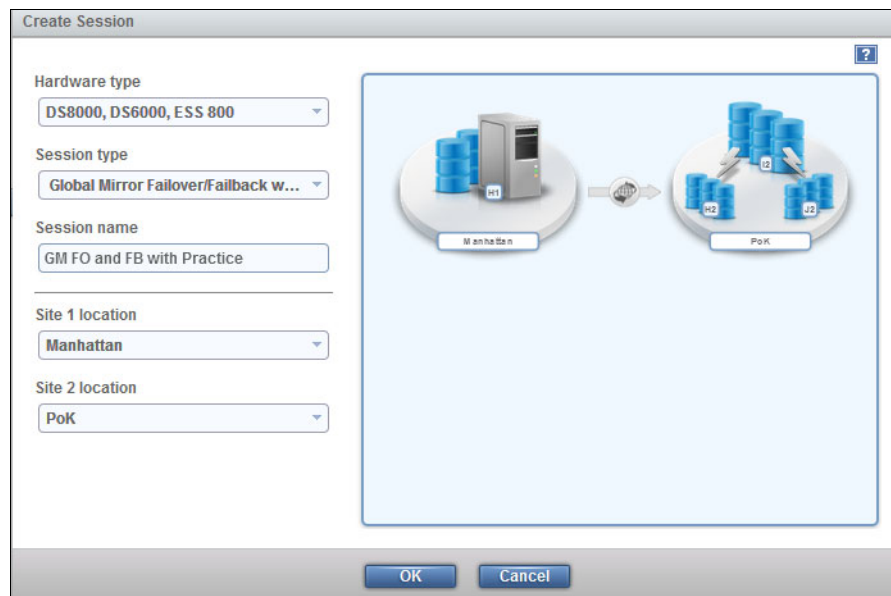

Figure 7-2   CSM Create Session wizard

CSM management console records all Global Mirror activities and message alerts, which can be broadcasted (SNMP traps, email) to the appropriate administrators.

For more information, such as installation and user guides, see the following sources of information:

► The latest IBM Knowledge Center information for IBM Copy Services Manager, found at this website.

   This website is the main entry point to all Copy Services Manager related documentation.

► Several IBM publications are available for Tivoli Storage Productivity Center for Replication (the former name of Copy Services Manager). Even though they refer to previous releases of the product, most of the information is still valid. Here are two useful publications:

   – *IBM TotalStorage Productivity Center for Replication Using DS8000*, SG24-7596.

   – *IBM Tivoli Storage Productivity Center for Replication for System z*, SG24-7563.

7.2 GDPS Global Mirror (GDPS/GM)

Geographically Dispersed Parallel Sysplex (GDPS) provides a range of solutions for disaster recovery and high availability in a z Systems centric environment. GDPS/GM provides support for Global Mirror within a GDPS environment. GDPS builds on facilities provided by System Automation and IBM NetView® and uses inband connectivity to manage the Global Mirror relationships.

GDPS is delivered as part of a services engagement that includes both the software and services to help during the planning and implementation of the solution.

GDPS/GM runs two different services to manage Global Mirror, both of which run on z/OS systems. The K-Sys function runs in the primary site with access to the primary disk systems and is where the day-to-day management of Global Mirror is performed. The R-Sys function runs in the secondary (DR) site with access to the secondary disk subsystems and is where the recovery of the production systems is managed (see Figure 7-3).

GDPS K-Sys is responsible for sending configuration information to the R-Sys. The K-Sys and R-Sys communicate information to each other using a Tivoli NetView for z/OS-to-Tivoli NetView for z/OS network communication mechanism over the wide area network (WAN).

GDPS provides the capability to use an additional set of devices (F2 volumes in Figure 7-3) on the remote site for testing purposes.
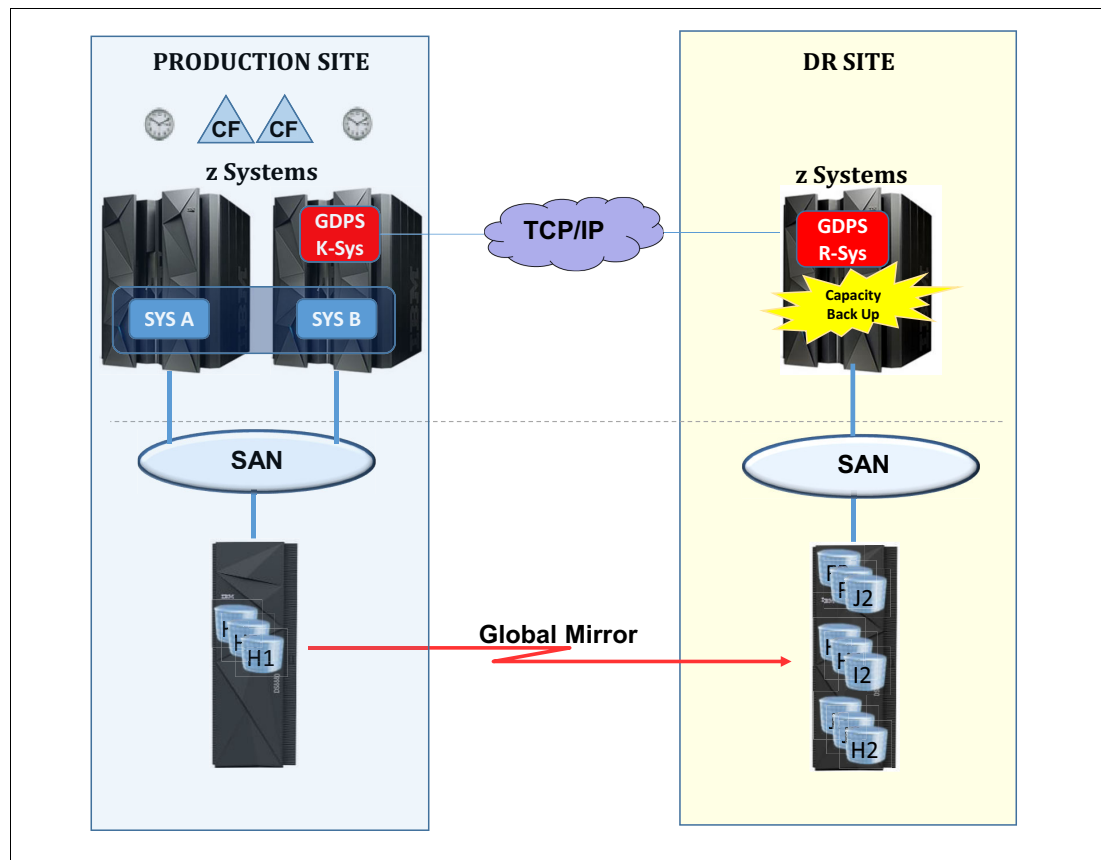

Figure 7-3   GDPS/GM environment

In addition to managing the operational aspects of Global Mirror, GDPS/GM also provides facilities to restart z Systems production systems in the recovery site. With scripting facilities, it provides a complete solution for the restart of a z Systems environment during a disaster situation. This feature does not require expert manual intervention to manage the recovery process.

GDPS supports Global Mirror management for both z Systems and Open Systems devices either in the same session as z Systems CKD disks or in a separate one. However, GDPS requires that the disk systems be shared between the z Systems and open systems environments, as it requires CKD device addresses to issue the commands to manage the Global Mirror environment.

As an alternative configuration, GDPS/GM also provides the Distributed Cluster Management (DCM) capability for managing global clusters by using Veritas Cluster Server (VCS) through the Global Cluster Option (GCO).

When the DCM capability is used, GDPS/GM does not manage remote copy or consistency for the distributed system disks (this is managed by VCS). Therefore, it is not possible to have a common consistency point between the z Systems CKD data and the distributed data. However, for environments where a common consistency point is not a requirement, DCM with VCS does provide various key availability and recovery capabilities that might be of interest.

GDPS/GM also provides the DCM capability for managing distributed clusters under IBM Tivoli System Automation Application Manager (SA AppMan) control. DCM provides advisory and coordinated functions between GDPS and SA AppMan-managed clusters. Data for the SA AppMan-managed clusters can be replicated using Global Mirror under GDPS control.

Thus, z/OS and distributed cluster data can be controlled from one point. Distributed data and z/OS data can be managed in the same consistency group (Global Mirror session) if cross-platform data consistency is required. Equally, z/OS and distributed data can be in different sessions and the environments can be recovered independently under GDPS control.

7.3 IBM System i PowerHA for i

IBM System i® PowerHA® for i (formerly known as HASM) is the IBM high availability disk-based clustering solution for the IBM i operating system. It is available starting with V6.1. PowerHA for i supports both Global Mirror and Metro Mirror for data replication when using independent auxiliary storage pools (iASP) for application data.
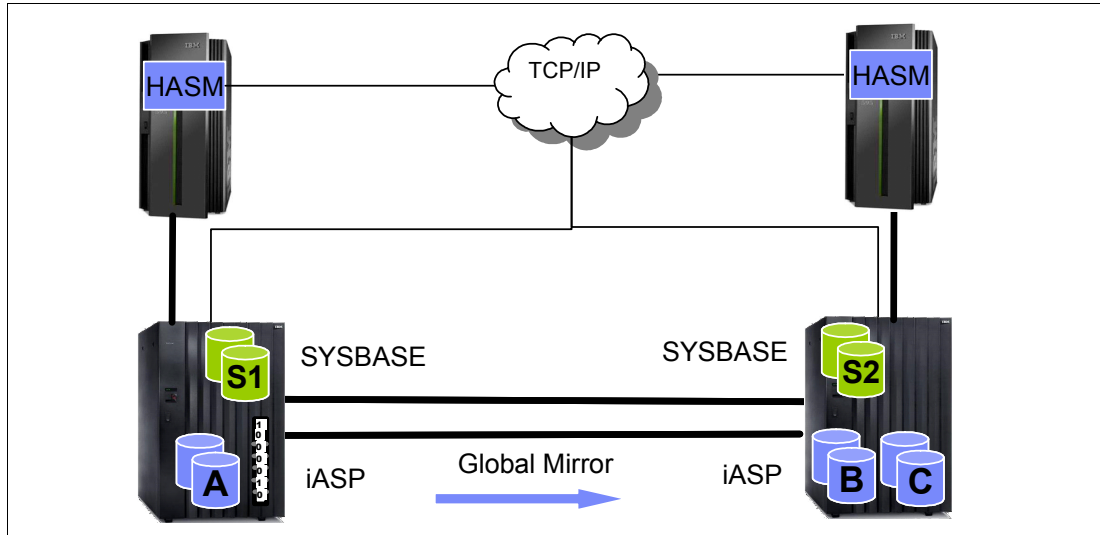
Figure 7-4 shows the PowerHA for i environment.

Figure 7-4   PowerHA for i environment

For earlier versions of i5/OS, the Copy Services Tool Kit provides support for Global Mirror. This feature also supports the use of Global Mirror to replicate a full system environment if not using iASPs. For more information, see *Implementing PowerHA for IBM i*, SG24-7405.

7.4 VMware SRM

VMWare Site Recovery Manager (SRM) VMWare software product that provides a DR solution for VMWare environments. The software has its own mechanisms to maintain the environment secure and fast recovered in cases of disaster. The high-level architecture is shown in Figure 7-5.
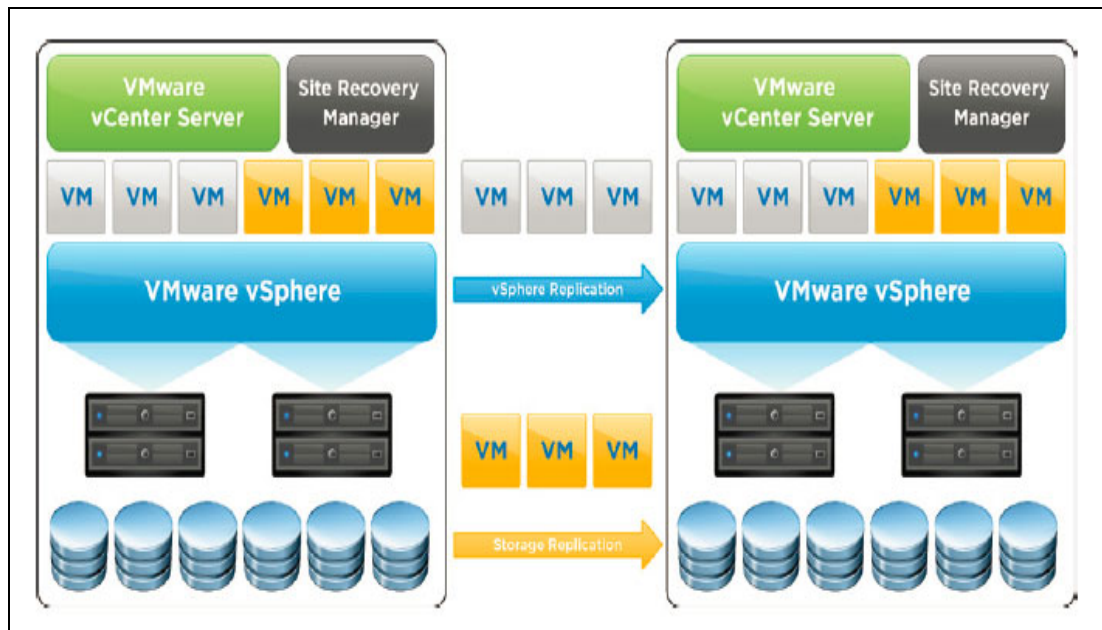

Figure 7-5   VMWare SRM Architecture

For more information, see Site Recovery Manager (SRM) Package at the VMWare website.

VMWare SRM is supported by Global Mirror technology when used together with an add-on called IBM DS8000 Storage Replication Adapter (SRA). This add-on is responsible for handling inputs from SRM and converting them into commands understandable by the DS8000. It must be deployed on both SRM servers (local and remote), as shown in Figure 7-6 and Figure 7-7 on page 76.

> **Note:** Go to Fix Central to obtain latest available version and IBM System Storage Interoperation Center (SSIC) for a compatibility software list.
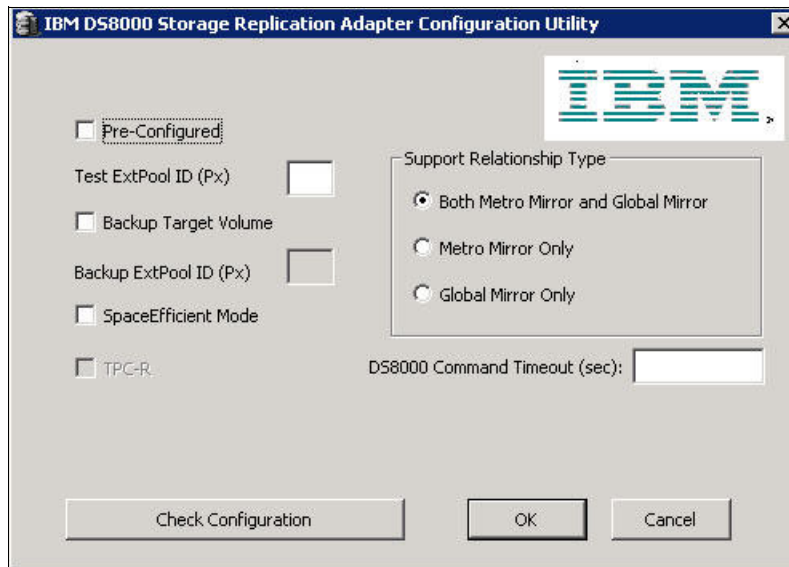

Figure 7-6   IBM DS8000 SRA for VMWare SRM Configuration Utility

A basic explanation of the options provided by the add-on is given in Table 7-1.

Table 7-1   IBM DS8000 SRA options

Option	Description
Pre-configured	The environment can be a preconfigured environment or not. A preconfigured environment is one where you create the backup target volumes and map the volumes to the VMware ESX servers at the recovery site in advance. A non-preconfigured environment is one where the SRA creates and maps snapshot volumes during tests for failover and backup target volume operations.
Test ExtPool ID(Px)	This is an available extent pool on the DS8000 where it creates FlashCopies of production volumes in case of a Test issued by SRM. For example, when creating a copy of the system for testing purposes.
Backup Target Volume/Backup ExtPool ID(Px)	This is an available extent pool where SRM makes a FlashCopy for backup purposes during failover. For example, when changing from A to B site.
Supported Relationship Type	Supports Metro Global Mirror, Metro Mirror, and Global Mirror.

| Option	Description
SpaceEfficient Mode	If specified, the system uses TSE repositories for each of the extent pools previously described. Note that with DS8880, TSE volumes are not supported, so be certain to review the compatibility before any implementation or testing.
DS8000 Command Time-out	The DS8000 Command Time-out specifies the maximum time, in seconds, to wait for a required state of volume, command results, FlashCopy, or remote copy. The wait is after certain DS8000 commands run, or to wait for command results.

Figure 7-7 shows an example of SRM.
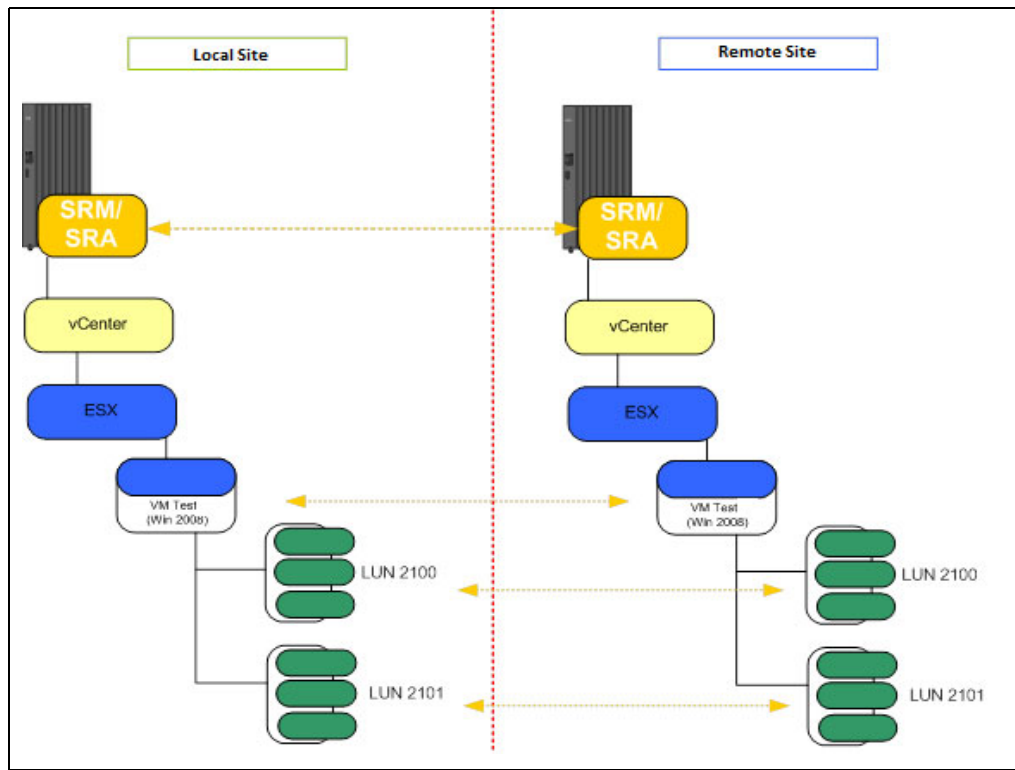
Figure 7-7   Example of SRM with SRA deployment topology

As previously explained, the add-on must be deployed at both locations to be able to successfully send commands when needed. For more information, see *IBM DS8000 Storage Replication Adapter User Guide*, SC27-4232-04.

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

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

► *DS8000 Copy Services*, SG24-8367
► *IBM DS8880 Architecture and Implementation (Release 8.2.1)*, SG24-8323
► *DS8870 Data Migration Techniques*, SG24-8257
► *IBM DS8870 Multiple Target Peer-to-Peer Remote Copy*, REDP-5151
► *IBM System Storage DS8000: Remote Pair FlashCopy (Preserve Mirror)*, REDP-4504
► *IBM TotalStorage Productivity Center for Replication Using DS8000*, SG24-7596
► *IBM Tivoli Storage Productivity Center for Replication for System z*, SG24-7563
► *Implementing PowerHA for IBM i*, SG24-7405.

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

► IBM DS8880 Knowledge Center:

https://www.ibm.com/support/knowledgecenter/ST5GLJ/ds8000_kcwelcome.html

► DFSMS Advanced Copy Services

https://www.ibm.com/support/knowledgecenter/SSLTBW_2.1.0/com.ibm.zos.v2r1.antg0
00/toc.htm

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services