© IBM Corporation 2008

Global Mirror Whitepaper

Author: Nick Clayton
European Storage Competence Centre

Date: 15/09/2008

Version: V2

Contents

Table of Figures

1. Notices and Disclaimer

Copyright © 2008 by International Business Machines Corporation.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This information may include technical inaccuracies or typographical errors.

IBM may make improvements and/or changes in the product(s) and/or programs(s) at any time without notice. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT.

IBM shall have no responsibility to update this information. IBM products are warranted according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein.

Any performance data contained herein was obtained in a controlled, isolated environment. Actual results that may be obtained in other operating environments may vary significantly. While IBM has reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

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

IBM, ESCON, FICON, System z, DS8000 and z/OS are trademarks of International Business Machines Corporation in the United States, other countries, or both. Other company, product or service names may be trademarks or service marks of others.

2. Design objectives

Global Mirror (Asynchronous PPRC) is a two-site unlimited distance data replication solution for both System z and Open Systems data. IBM's Global Mirror Solution addresses customer requirements for a long distance (typically over 300km) storage based data replication solution for both Open systems and System z data that scales and provides cross volume/cross storage subsystem data integrity/data consistency.

Before examining in detail the architecture and operation of Global Mirror, it is helpful to first review the design objectives that guided the development of this solution. The section below provides some background on how the design objectives influenced the development of Global Mirror.

Achieve an RPO of 3-5 seconds with sufficient bandwidth and resources

Many customers have implemented z/OS Global Mirror (previously known as XRC) as an asynchronous replication solution and have become accustomed to the RPO that this best of breed solution is available to provide. Global Mirror has been designed to be able to provide a similar RPO of 3-5 seconds when sufficient bandwidth and resources are available.

Do not impact production applications when insufficient bandwidth and/or resources are available

Previous asynchronous replication solutions have used a cache sidefile to store updates before transmission to the remote site. As a result, they have included pacing mechanisms to slow down production write activity and allow the mirroring to continue if the replication solution falls behind. Global Mirror has been designed not to use a sidefile and so requires no pacing mechanism.

Be scalable and provide consistency across multiple primary and secondary disk subsystems

A replication solution should not be limited to a single primary or secondary disk subsystem as this may result in an existing solution becoming non-viable if the storage and throughput requirements outgrow the capabilities of a single disk subsystem. Allowing a single consistency group to span multiple disk subsystems also allows different cost storage to exist within the same Global Mirror environment. (For example a Global Mirror session might span an ESS, DS6800 and DS8300.)

Allow for removal of duplicate writes within a consistency group before sending to remote site

Bandwidth is one of the most expensive components of an asynchronous replication solution and minimising the usage of bandwidth can provide significant cost savings for a replication solution. If multiple updates are made to the same piece of data within a consistency group then only the latest update need be sent. Depending on the access patterns of the production workload, significant savings might be seen in the bandwidth required for the solution.

Allow for less than peak bandwidth to be configured accepting a higher RPO at peak times

Many customers have significant peaks in the write activities of their workloads, which may be 2-3 times higher than the average write throughput. These peaks are often at times where maintenance or batch activities are taking place and so there may not be sufficient justification to provide the bandwidth to maintain a very small RPO at these times. However, the activities that take place are likely to be time critical and so the production workload should not be impacted if sufficient bandwidth is not available

Provide consistency between different platforms especially between System z and open systems

With the increase in applications spanning multiple servers and platforms there is a requirement to be able to provide a consistent asynchronous replication solution that can handle workloads from multiple servers and specifically for both CKD (System z) and FB (open systems) data. Global Mirror can be used on any devices that are defined on the disk subsystem including System z, System i and UNIX/Windows workloads.

3. Overall architecture

For any asynchronous replication function, we can consider three major functions that are required in order to provide consistent mirroring of data. These are:

1. Creation of a consistent point across the replicated environment
2. Transmission of required updates to the secondary location
3. Saving of consistent data to ensure a consistent image of the data is always available.

With Global Mirror, the functions are provided as shown in Figure 1.

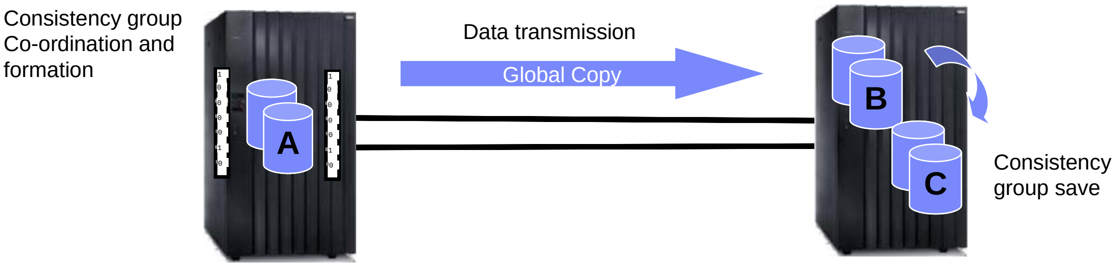


Figure 1 Global Mirror device topology

The primary disk subsystems provide functionality to co-ordinate the formation of consistency groups across all involved devices which are said to be in a Global Mirror session. Fibre channel links provide low latency connections between multiple disk subsystems ensuring that this process involves negligible impact to the production applications. The consistency group information is held in bitmaps rather than requiring the updates to be maintained in cache.

These consistency groups are sent to the secondary location using Global Copy (previously known as PPRC-XD). Using Global Copy means that duplicate updates within the consistency group are not sent and that if the data sent is still in the cache on the primary disk subsystems that only the changed blocks are sent.

Once the consistency group has been sent to the secondary location this consistent image of the primary data is saved using FlashCopy. This ensures that there is always a consistent image of the primary data at the secondary location.

3.1 FlashCopy Space Efficient (FlashCopy-SE) and Global Mirror

It is possible to use a space efficient device as the C device in a Global Mirror environment. If an additional D copy has been provided for testing (see section 5.2) then it can also be used for this device. Generally it is not recommended to use FlashCopy-SE devices where only a C device is provided or for both the C and D devices as filling the space efficient repository would result in no usable data being available on the remote site during the resynchronisation process.

There are additional overheads when using space efficient devices with FlashCopy and so it is important to understand the affect of this on the Global Mirror performance. Generally the impact will be higher for workloads with a large sequential write content and where the x-site bandwidth is large enough that the network is not the bottleneck. However for workloads with a more random write workload or with a high read to write ratio the use of FlashCopy-SE can provide significant savings.

Depending on the usage it may be more appropriate to make either the C or the D device space efficient. If testing will only be done on an occasional basis and will not last a significant time then using FlashCopy-SE for the D device will mean that in normal operation there is no impact on performance. However this will mean that it is not possible to recover on the same D device in a disaster so testing and actual recovery would be different.

3.2 Dependant write consistency

Global Mirror provides consistency in a similar fashion to PPRC and FlashCopy consistency groups by using a data freeze, which preserves the order of dependant writes. This is in contrast to z/OS Global Mirror, which provides consistency using timestamps, and other solutions, which might provide consistency using sequence numbers.

Using the data freeze concept allows the creation of consistency across multiple disk subsystems without requiring a common and extremely accurate time source like the Sysplex Timer in a System z environment.

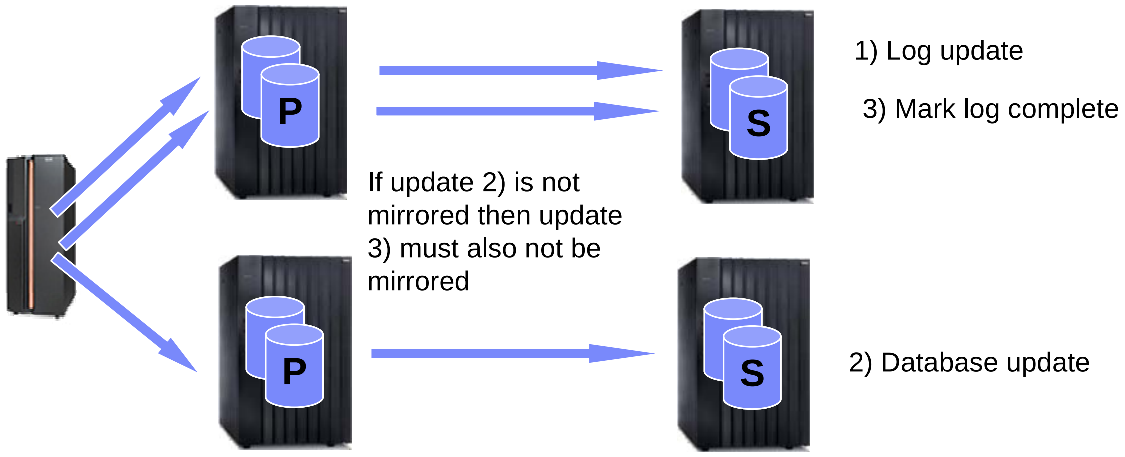

P
S
1) Log update
3) Mark log complete
If update 2) is not mirrored then update 3) must also not be mirrored
P
S
2) Database update

Figure 2 Dependant write consistency

Using the data freeze concept, consistency is obtained by temporarily inhibiting write IO to the devices in an environment and then performing the actions required to create consistency. Once all devices have performed the required actions the write IO is allowed to resume.

This might be suspending devices in a Metro Mirror environment or performing a FlashCopy when using consistent FlashCopy. With Global Mirror, this action is the creation of the bitmaps for the consistency group and we are able to create the consistent point in approximately 1-3ms; Section 4.3 discusses the impact of consistency group formation on average IO response times.

Other solutions using a data freeze are not as highly optimised as Global Mirror. Therefore, they might take longer to perform the consistency creation process. For example creating a consistent FlashCopy might take a number of seconds depending on the size of the environment.

3.3 Communication between primary disk subsystems

In order to form consistency groups across multiple disk subsystems the different disk subsystems must be able to communicate. This communication path must be resilient and high performance in order to provide minimal impact to the production applications. In order to provide this path Fibre Channel links are used, which can be direct connections or over a SAN.

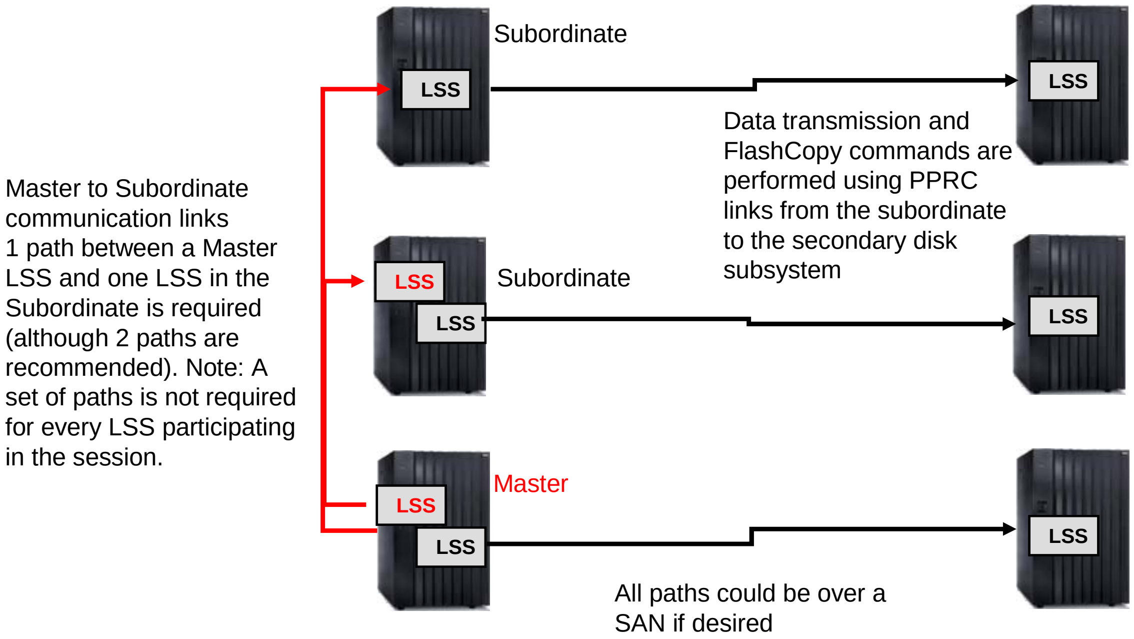

Subordinate
LSS
LSS
Data transmission and FlashCopy commands are performed using PPRC links from the subordinate to the secondary disk subsystem
Master to Subordinate communication links
1 path between a Master LSS and one LSS in the Subordinate is required (although 2 paths are recommended). Note: A set of paths is not required for every LSS participating in the session.
LSS
Subordinate
LSS
LSS
Master
LSS
LSS
LSS
All paths could be over a SAN if desired

Figure 3 Communication links between primary disk subsystems

With Global Mirror, a controlling function, known as the master, runs on one of the primary disk subsystems. This process coordinates the consistency group formation process over all the subordinate disk subsystems using the PPRC links. The master function can be started on any primary disk subsystem so long as the required connectivity is available.

3.4 Global Mirror setup process

The setup process for Global Mirror is in two stages. First, the topology of the environment is defined by creating the Global Copy and FlashCopy relationships that will be used by Global Mirror. Then the Global Mirror session definitions are created and the Global Mirror session is started on the master disk subsystem.

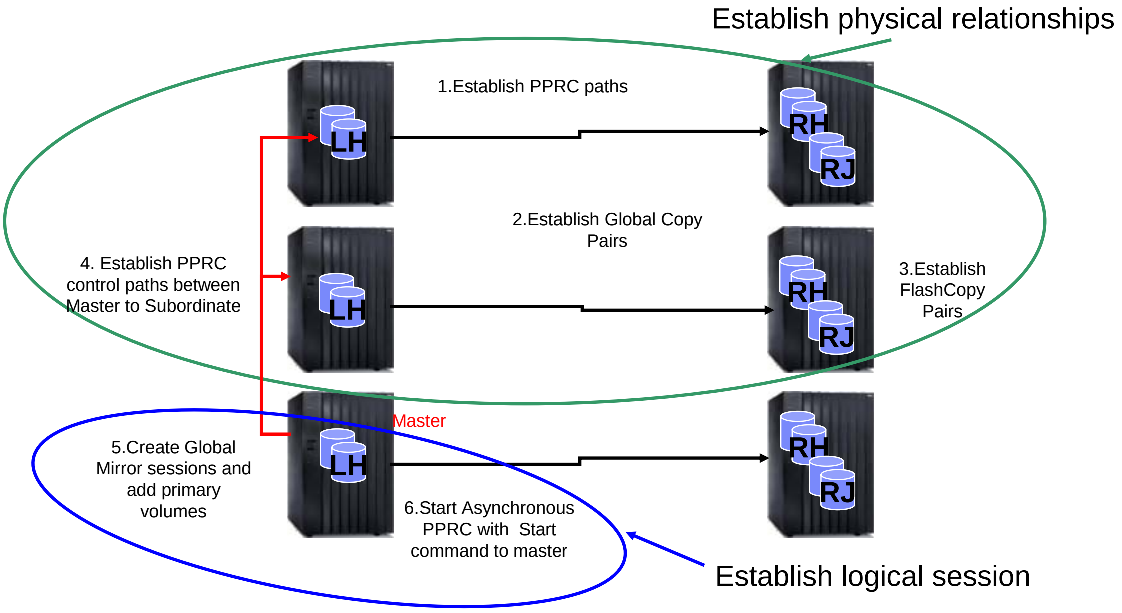

Establish physical relationships
1.Establish PPRC paths
LH
RH
RJ
2.Establish Global Copy Pairs
4. Establish PPRC control paths between Master to Subordinate
LH
RH
RJ
3.Establish FlashCopy Pairs
Master
5.Create Global Mirror sessions and add primary volumes
LH
RH
RJ
6.Start Asynchronous PPRC with Start command to master
Establish logical session

Figure 4 Global Mirror setup process

3.5 Adding volumes to a Global Mirror session

The process for adding volumes to a Global Mirror session is similar to the process for setting up the environment. When additional volumes are added to the session they will initially be placed in a Join Pending state until they have performed their initial copy. Once this has completed they will then join the session when the next consistency group is formed.

Adding a new disk subsystem 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 might take a few seconds and is done in order to define the new disk subsystem and its control paths to the master process. This has minimal impact to the environment and only results in a small increase to the RPO for a brief period.

3.6 Current restrictions

The Global Mirror architecture allows for the creation of multiple Global Mirror sessions on a primary disk subsystem with different devices associated with each session. Currently only a single active Global Mirror session is allowed per primary disk subsystem so all devices being replicated on a single disk subsystem must be in the same consistency group.

The maximum supported number of disk subsystems participating in the Global Mirror environment is eight including both primary and secondary disk subsystems. It is possible to submit an RPQ if a larger number of disk subsystems are required and there are currently customers running in larger environments. The architecture allows for sixteen primary disk subsystems and does not have any explicit restriction on the number of secondary disk subsystems.

3.7 Differences between Global Mirror and z/OS Global Mirror

z/OS Global Mirror (previously called XRC) is IBM's System z based asynchronous replication technology. It uses a combination of microcode on primary disk subsystems and a z/OS software component that performs the data movement and manages the consistency of data. This section will discuss the differences between Global Mirror and z/OS Global Mirror.

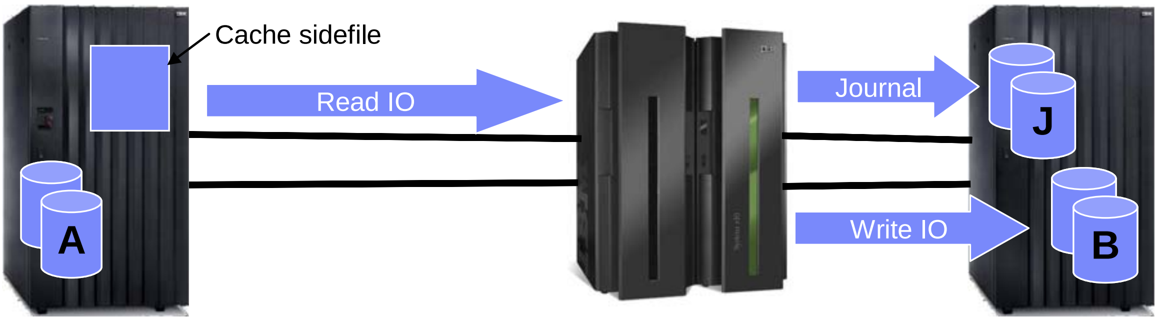


Figure 5 z/OS Global Mirror architecture

System z only scope for z/OS Global Mirror

z/OS Global Mirror is only able to manage CKD devices and can only provide consistency outside of a single storage control session where the production operating system is timestamping the application writes. Global Mirror is able to handle both CKD and FB devices as it uses the data freeze concept to provide consistency and so does not require a common time source for all servers.

Use of cache by z/OS Global Mirror

z/OS Global Mirror like many other asynchronous replication solutions temporarily holds the production workload updates in the cache on the primary disk subsystem. If the environment is configured with sufficient bandwidth and resources this is not an issue, but as cache is a finite resource, we have two options when the cache fills up. We can either suspend the mirroring or we can pace the production systems in order to prevent further increase in cache usage. Global Mirror does not use a cache sidefile and so there is no requirement to pace production or suspend the mirror when behind.

Predictability of RPO vs. impact to production

As z/OS Global Mirror is often configured to pace the production writes it is possible to know both an average RPO (referred to as delay by z/OS Global Mirror) under non-stress conditions and what a reasonable maximum RPO might be. However, it is possible that the mirror will either suspend or production workloads will be impacted if the capability of the replication environment is exceeded due to unexpected peaks in the workload or a under configured environment.

With Global Mirror, as we do not pace the production writes, it is possible that the RPO will increase significantly if the production workload exceeds the resources available for Global Mirror. However, we will not have to perform a suspension and subsequent resynchronisation of the mirror and we will not inject any delays into the production IO activity.

Requirement for additional copy during resynchronisation

When z/OS Global Mirror suspends the mirror due to a network outage or other cause in order to preserve a consistent image in the secondary site we need to take a point in time copy of the secondary devices before resynchronisation. With Global Mirror, as the architecture already includes two copies of the data on the secondary site, the C devices will contain a useable image of the production systems even during resynchronisation.

Removal of duplicate updates before transmission

With z/OS Global Mirror, the formation of consistency groups is performed by a z/OS software component (the SDM) and so when the data is read from the primary disk subsystem we have to read every write. With Global Mirror, as we have the consistency group formation on the primary disk subsystem, we are able to remove duplicate updates within a consistency group before they are sent to the secondary disk subsystem.

Note: z/OS Global Mirror will also remove duplicate updates before writing to the secondary disk subsystems. However this does not result in bandwidth savings as for a disaster recovery solution the SDM is located in the secondary location.

Mixed vendor implementations

z/OS Global Mirror is also supported on disk subsystems from other vendors who have licensed and implemented the interfaces from IBM and it is possible to run with a heterogeneous environment with multiple vendors disks. Currently Global Mirror is only available on IBM disk subsystems and so a heterogeneous environment is not possible. It is possible that other vendors will choose to license the Global Mirror architecture as they have done with Metro Mirror and z/OS Global Mirror. Target z/OS Global Mirror volumes can also be from any vendor, even if the target subsystem does not support z/OS Global Mirror, thus enabling investment protection. However, in order to provide for failback, customers will likely prefer that the remote systems also support z/OS Global Mirror.

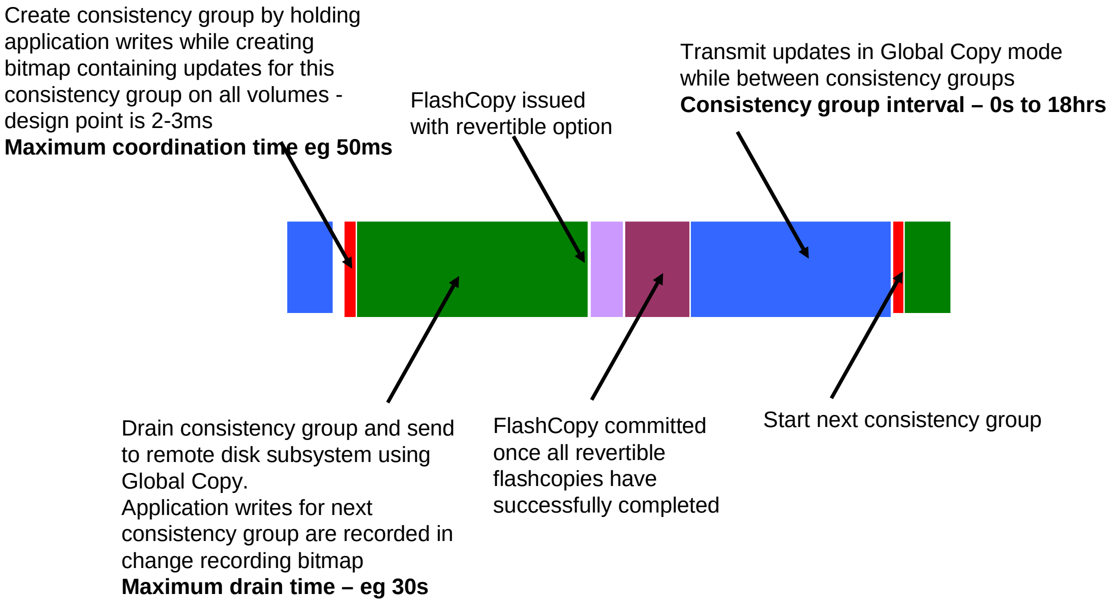
4. Consistency group creation process

4.1 Process overview

As we have discussed earlier the process for forming consistency groups can be broken down conceptually into three steps:

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

Figure 6 shows a more detailed breakdown of the consistency group formation process. The tuneable options available are shown in **bold** and are discussed in detail in Section 4.2 Tuneable values



Create consistency group by holding application writes while creating bitmap containing updates for this consistency group on all volumes - design point is 2-3ms
Maximum coordination time eg 50ms

FlashCopy issued with revertible option

Transmit updates in Global Copy mode while between consistency groups
Consistency group interval – 0s to 18hrs

Drain consistency group and send to remote disk subsystem using Global Copy.
Application writes for next consistency group are recorded in change recording bitmap
Maximum drain time – eg 30s

FlashCopy committed once all revertible flashcopies have successfully completed

Start next consistency group

Figure 6 Consistency group creation process

If the maximum drain time is exceeded then Global Mirror will determine how long it will take before another consistency group could be expected to be formed and will transition to Global Copy mode until it is possible to form consistency groups again. The production performance will be protected but the RPO will be allowed to increase.

In this way, Global Mirror maximised the efficiency of the data transmission process by allowing duplicate updates to be sent and allowing all primary disk subsystems to send data independently. Global Mirror will check on a regular basis whether it is possible to start forming consistency groups and will do so as soon as it calculates that this is possible.

4.2 Tuneable values

Global Mirror has three tuneable values to modify its behaviour. In most environments, the default values should be used.

Maximum Coordination Time

The maximum coordination time is the maximum time that Global Mirror will allow for the determination of a consistent set of data before failing this consistency group. Having this cut-off ensures that even if there is some error recovery event or communications problem the production applications will not experience significant impact from consistency group formation.

The default for the maximum co-ordination time is 50ms, which is a very small value compared to other IO timeout values such as MIH (30seconds) or SCSI IO timeouts. Hence, even in error situations where we might trigger this timeout, Global Mirror will protect production performance rather than impacting production in an attempt to form consistency groups in a time where there might be error recovery or other problems occurring.

Maximum Drain Time

The maximum drain time is the maximum amount of time that Global Mirror will spend draining a consistency group before failing the consistency group. If the maximum drain time is exceeded then Global Mirror will transition to Global Copy mode for a period of time in order to catch up in the most efficient manner. While in Global Copy mode the overheads will be lower than continually trying and failing to to create consistency groups

The previous consistency group will still be available on the C devices so the effect of this will simply be that the RPO increases for a short period. The primary disk subsystem will evaluate when it would be able to be possible to continue to form consistency groups and will restart consistency group formation at this time.

The default for the maximum drain time is 30 seconds, which allows a reasonable time to send a consistency group while ensuring that if there is some non-fatal network or communications issue that we do not wait too long before evaluating the situation and potentially dropping into Global Copy mode until the situation is resolved. In this way, we again protect the production performance rather than attempting (and possibly failing) to form consistency groups at a time when this might not be appropriate.

If we are unable to form consistency groups for 30 minutes, by default Global Mirror will form a consistency group without regard to the maximum drain time. It is possible to change this time if this behaviour is not desirable in a particular environment.

Consistency Group Interval

The consistency group interval is the amount of time Global Mirror will spend in Global Copy mode between the formation of each consistency group. The effect of increasing this value will be to increase RPO and can increase efficiency of bandwidth utilisation by increasing the number of duplicate updates that occur between consistency groups and do not need to be sent from the primary to the secondary disk subsystems

However, as it also increases the time between successive FlashCopies increasing this value is not necessary and may be counter productive in high bandwidth environments as frequent consistency group formation will reduce the overheads of Copy on Write processing (see section 9.4).

The default for the Consistency Group Interval is 0 seconds so Global Mirror will continuously form consistency groups as fast as the environment will allow. In most situations, we would recommend leaving this parameter at the default and allowing Global Mirror to form consistency groups as fast as possible given the workload as it will automatically move to Global Copy mode for a period of time if the drain time is exceeded.

4.3 Response time impact from consistency group creation process

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

This process must therefore be fast enough that an impact is extremely small. With Global Mirror, the process of forming a consistency group is designed to take 1-3ms. If we form consistency groups every 3-5 seconds then the percentage of production writes impacted and the degree of impact is therefore very small.

The example below shows the type of impact that might be seen from consistency group formation in a Global Mirror environment.

We assume that we are going 24000 IOs per second with a 3:1 R/W ratio. We perform 6000 write IO/sec and each write I/O takes .5 ms and it take 3ms to create a consistent set of data.

Approximately (.0035*6000) = 21 write I/Os are affected by the creation of consistency.

If each of these 21 I/Os experiences a 3 ms delay, and this happens every 3 seconds, then we have an average RT delay of (21*.003)/18000 = .0035 ms.

0.0035ms average impact to a 0.5ms write is a 0.7% increase in response time and normal performance reporting tools will not detect this level of impact.

4.4 Collisions

Global Mirror does not use a cache sidefile in order to avoid the issues with cache filling up that are seen with other asynchronous replication solutions. However, this does have the implication that if updates are made to tracks in a previous consistency group that have not yet been sent then this previous image needs to be protected. We call this situation a collision.

In order to do this we will delay the completion of the write until the previous track image has been sent to the secondary disk subsystem. This is done immediately and with the highest priority.

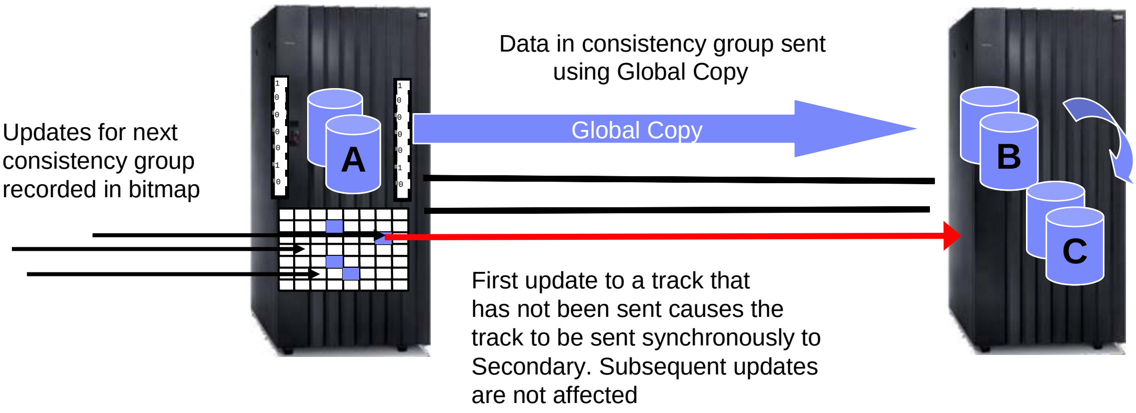

Data in consistency group sent using Global Copy
Global Copy
Updates for next consistency group recorded in bitmap
A
B
C
First update to a track that has not been sent causes the track to be sent synchronously to Secondary. Subsequent updates are not affected

Figure 7 Collisions in Global Mirror

For most intensive write workloads such as log volumes updates are performed in a sequential fashion and the same piece of data is not updated on a regular basis. However even for workloads such as an IMS WADS dataset, the collision for a particular track will only occur once for each consistency group and so the synchronous overhead is only seen once per consistency group.

Considerable analysis was performed on the potential impact of collisions for the IMS WADS dataset and it was determined that a very small percentage of the writes would experience collisions and so, similar to the impact of consistency group formation, this effect would be very small.

5. Recovery process

This section aims to give a high-level overview of the recovery process for a Global Mirror environment. More details are available in the various manuals referred to in the References section.

5.1 Recovery of Global Mirror environment

There are two stages to the recovery of a Global Mirror environment.

The first stage is to check the exact status of Global Mirror at the time of the outage. Depending on where Global Mirror was in the consistency group formation process at the time of the failure event there may be actions required to ensure that the "C" copy is consistent.

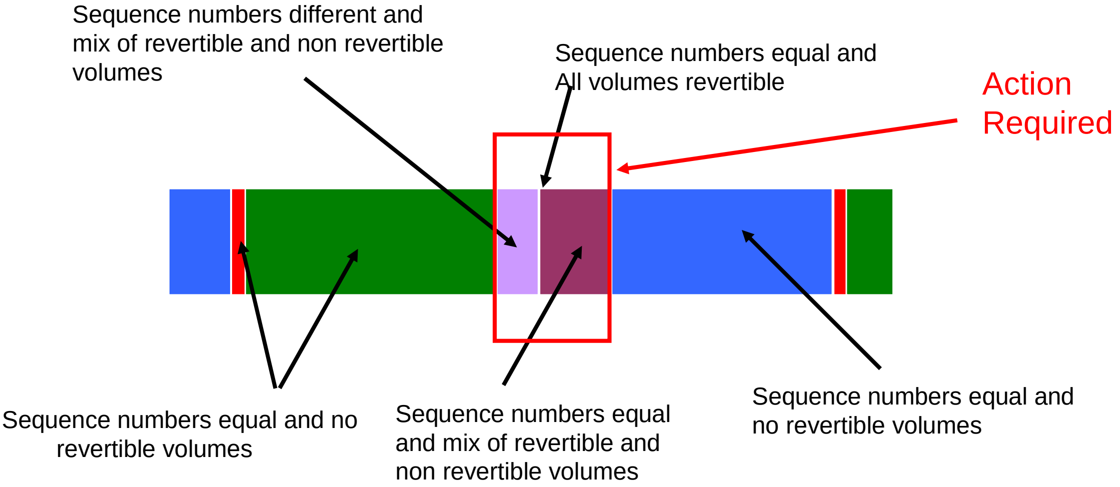

Sequence numbers different and mix of revertible and non revertible volumes
Sequence numbers equal and All volumes revertible
Action Required
Sequence numbers equal and no revertible volumes
Sequence numbers equal and mix of revertible and non revertible volumes
Sequence numbers equal and no revertible volumes

Figure 8 Parts of consistency group formation process where recovery action is required

If the Global Mirror environment was part way through the FlashCopy process when the failure occurred then this is similar to an in-flight transaction in a database environment. If the commit process has not yet started then we must revert the FlashCopy to back out the consistency group and if the commit process has started, we must complete this process to commit the consistency group.

1) Commit/revert FlashCopy relationships if required.

The second stage is to recover the environment and enable production systems to be restarted on the B devices and prepare for a potential return to the primary site. This is performed with the following process:

2) Failover the B devices. This will place the B devices in a primary suspended state and will allow for a resynchronisation of the Global Copy relationship to be performed in order to return to the primary site assuming the primary disk subsystem has survived.
3) Fast Reverse Restore the FlashCopy relationship with the C devices. This will restore the latest consistency group to the B devices and will start a background copy for those tracks that have been modified since the latest consistency group.
4) FlashCopy from the B devices to the C devices to save an image of the last consistency group. This step is optional but preserves an image of the production devices at the recovery point in case this might be required.
5) Restart Production systems.

Figure 9 shows the Global Mirror recovery process with a green colour indicating the location of the restartable production image.

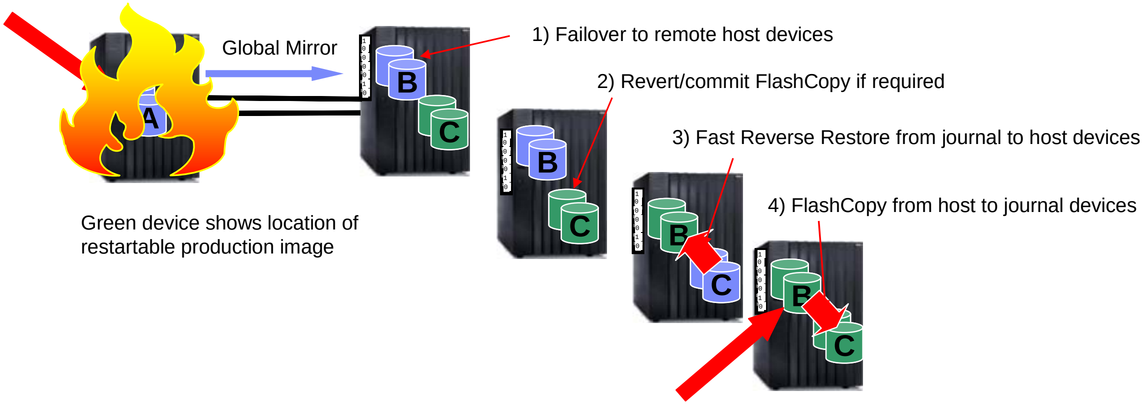

Global Mirror
1) Failover to remote host devices
2) Revert/commit FlashCopy if required
3) Fast Reverse Restore from journal to host devices
4) FlashCopy from host to journal devices
Green device shows location of restartable production image
A
B
C

Figure 9 Global Mirror recovery process

5.2 Taking an additional copy for DR testing

If there is a requirement to perform disaster recovery testing while maintaining the currency of the mirror in a Global Mirror environment or for taking regular additional copies perhaps once or twice a day for other purposes then the following process is performed.

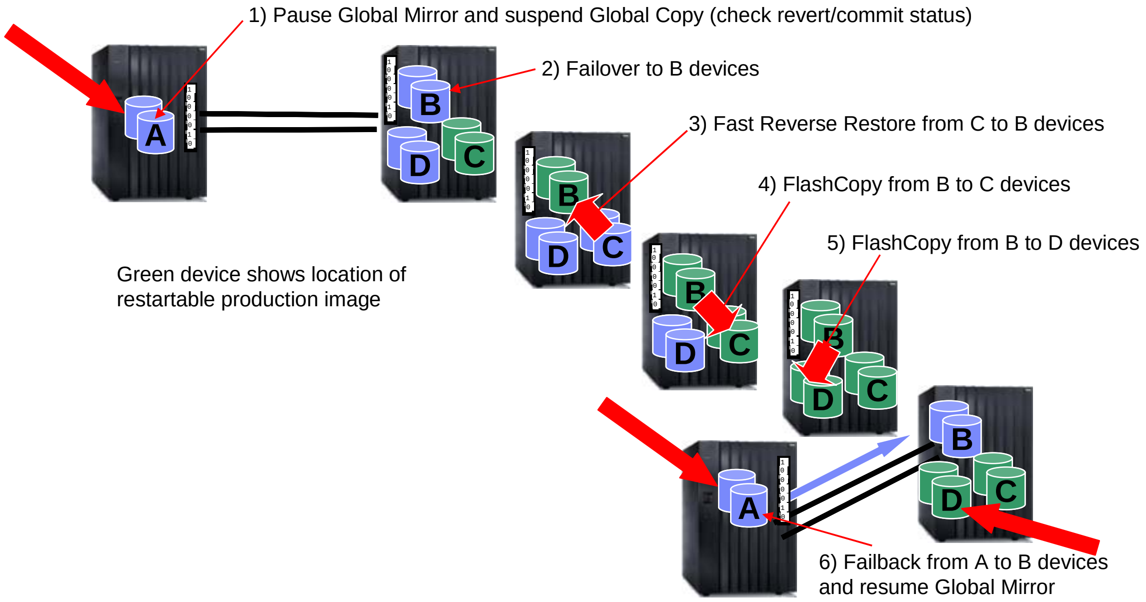

1) Pause Global Mirror and suspend Global Copy (check revert/commit status)
2) Failover to B devices
3) Fast Reverse Restore from C to B devices
4) FlashCopy from B to C devices
5) FlashCopy from B to D devices
6) Failback from A to B devices and resume Global Mirror
Green device shows location of restartable production image
A
B
C
D

Figure 10 Process to take an additional copy for testing

This is similar to the procedure for recovering the Global Mirror environment with the addition of steps to pause and resume the mirroring plus the creation of the additional test copy.

6. Autonomic behaviour

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

PPRC paths

If PPRC paths are removed unexpectedly for some reason then the disk subsystem will automatically re-establish these paths when the error situation is resolved. In a z/OS environment, the following messages will be seen on the console indicating that these events have occurred. Similar SNMP alerts are produced for open systems environments.

```
IEA498I 4602,PRSA53,PPRC-PATH ALL PPRC PATHS REMOVED UNEXPECTEDLY
            SSID=4600 (PRI)=0175-38711,CCA=02

IEA498I 4100,IMNA76,PPRC-PATH ONE OR MORE PPRC PATHS RESTORED
            SSID=4100 (PRI)=0175-38711,CCA=00
```

PPRC pairs

If Global Copy pairs suspend for some reason other than a user command then the disk subsystem will automatically attempt to restart the mirroring for these pairs. As the consistent set of disks is the FlashCopy secondary devices, this will not compromise the integrity at the secondary site. This is different from Metro Mirror where the resynchronisation is always via command as the Metro Mirror secondaries are the consistent devices.

```
*IEA491E 4003,PRSA60,PPRC SUSPENDED,COMMUNICATION_TO_SECONDARY_FAILURE,
      (PRI)=0175-38711,CCA=03(SEC)=0100-63271,CCA=03
IEA494I 4003,PRSA60,PPRC PAIR SUSPENDED,SSID=4000,CCA=03

IEA494I 4601,IMNA02,PPRC PAIR PENDING,SSID=4600,CCA=01
```

It is possible to disable this behaviour if desired.

Global Mirror session

Once the Global Copy pairs are restarted and have resynchronised then Global Mirror will resume the formation of consistency groups unless in doing so it might result in inconsistent data on the secondary disk subsystem.

One example of such a condition is where a communications failure occurs half way through a FlashCopy event. In this case we have to perform a revert/commit action as described in section 4 before restarting Global Mirror. In this case the Global Mirror session will have entered what is called a “Fatal” state.

7. Management tools

There are a number of options for management tools when using Global Mirror. System z provides inband management capabilities while in an Open Systems environment out of band management using TCP/IP is used.

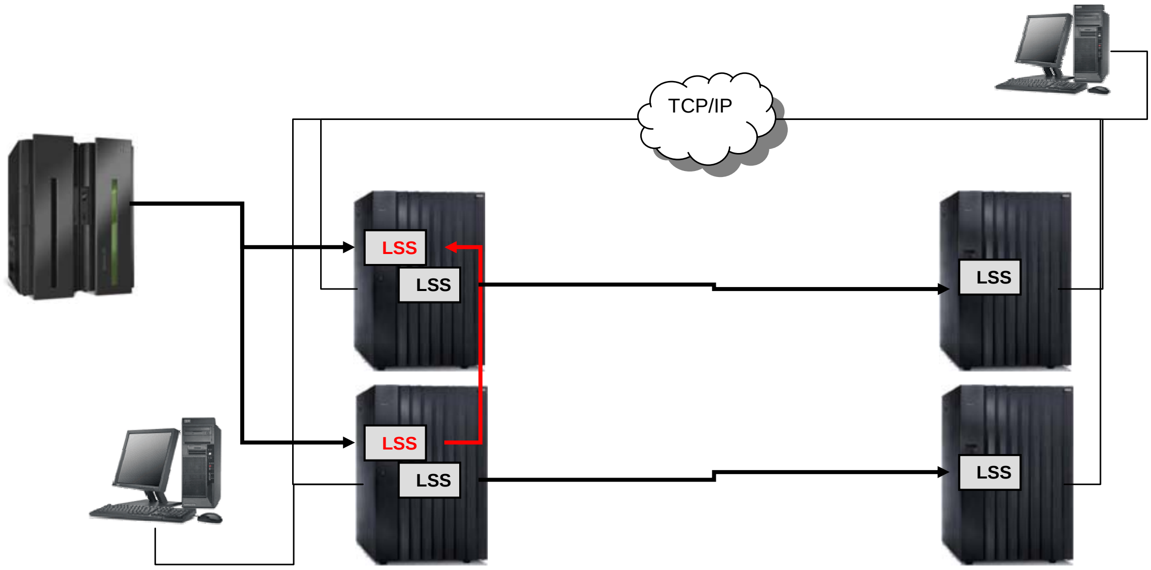


Figure 11 Inband and out of band management of Global Mirror

Command interfaces for Global Mirror are available in both System z and Open Systems environments to allow a customer to develop their own automation solution as well as providing problem diagnosis capabilities if a solution offering is used to manage the environment.

IBM's disaster recovery and replication management solution offerings also have been extended to provide support for Global Mirror. In most cases if these solutions fit the requirements for the environment being used these would be the recommended option as this provides a supported management solution without requiring code to be written specifically for each environment.

7.1 Command interfaces

The DS Command Line Interface contains support for the management of a Global Mirror environment. With its support for scripting and use of ranges and lists of devices within a single command, this is a powerful environment to use if a bespoke solution is required.

TSO and the ANTRQST API contain support for Global Mirror for both CKD and FB devices and can be used to manage either a purely System z environment or, if all disk subsystems have System z devices, can manage a heterogeneous environment with both CKD and FB devices.

ICKDSF contains support for Global Mirror with CKD devices and can be used to manage Global Mirror for z/OS, VM and VSE environments. There is a collection of REXX utilities available for management of Global Mirror using ICKDSF. These utilities allow for the creation of ICKDSF jobs from a configuration file, which will automate the operations required to manage Global Mirror.

Both ICKDSF and TSO/ANTRQST require connectivity to the secondary disk subsystem in order to perform the actions required for recovery or testing in a Global Mirror environment. This could either be from a system running in the secondary location or for testing, could be by having connectivity from the production systems. Connectivity for the production systems could potentially be provided over ISLs in the same SAN that is used for the PPRC links.

The DS CLI requires IP connectivity to the HMC or copy services server for the disk subsystems it is managing both in the primary site and to the recovery site for the actions required for recovery or testing.

7.2 Global Mirror Management and Disaster Recovery Solutions

Global Mirror has a number of supported management solutions to provide additional facilities beyond the raw command interfaces. These solutions provide different capabilities depending on the exact requirements for a particular situation.

7.2.1 TPC for Replication (TPC-R)

TPC for Replication is a remote copy management solution for disaster recovery that supports a variety of different environments including Metro Mirror, FlashCopy, Global Mirror and Metro Global Mirror. TPC-R runs as a WebSphere application on System x, System p or System z servers to provide a management environment for replication including Global Mirror. It uses a TCP/IP interface to the disk subsystem in order to manage the Global Mirror environment.

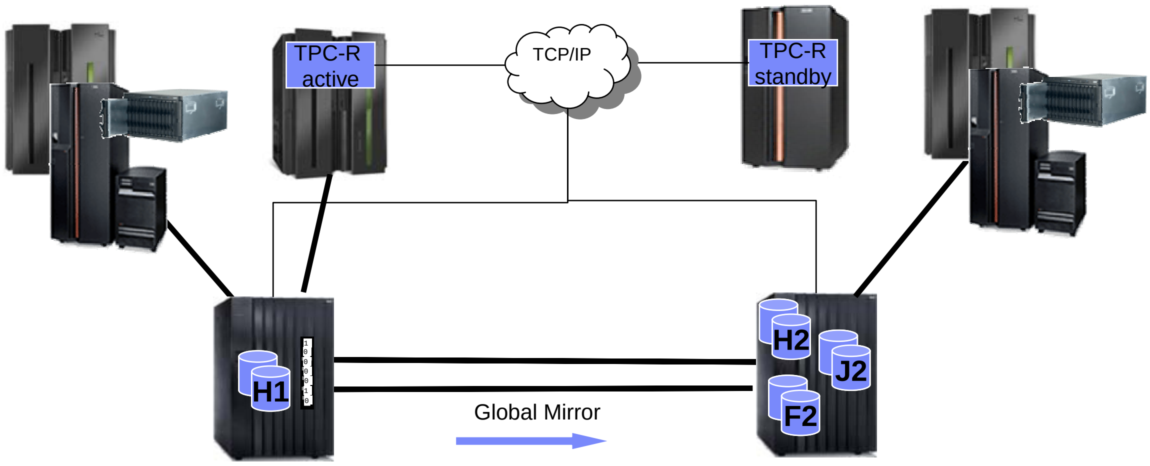


Figure 12 TPC for Repli cation environnent

TPC-R 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.

7.2.2 GDPS Global Mirror (GDPS/GM)

GDPS provides a range of solutions for disaster recovery and high availability in a System z centric environment. GDPS/GM provides support for Global Mirror within a GDPS environment. GDPS builds on facilities provided by System Automation and Netview and utilises inband connectivity to manage the Global Mirror relationships.

GDPS is delivered as part of a services engagement which includes both the software and services to assist in 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 subsystems and is where the day-to-day management of Global Mirror is performed. The R-sys function runs in the secondary site with access to the secondary disk subsystems and is where the recovery of the production systems is managed.

As well as managing the operational aspects of Global Mirror GDPS/GM also provides facilities to restart System z production systems in the recovery site. By providing scripting facilities it provides a complete solution for the restart of a System z environment in a disaster situation without requiring expert manual intervention to manage the recovery process.

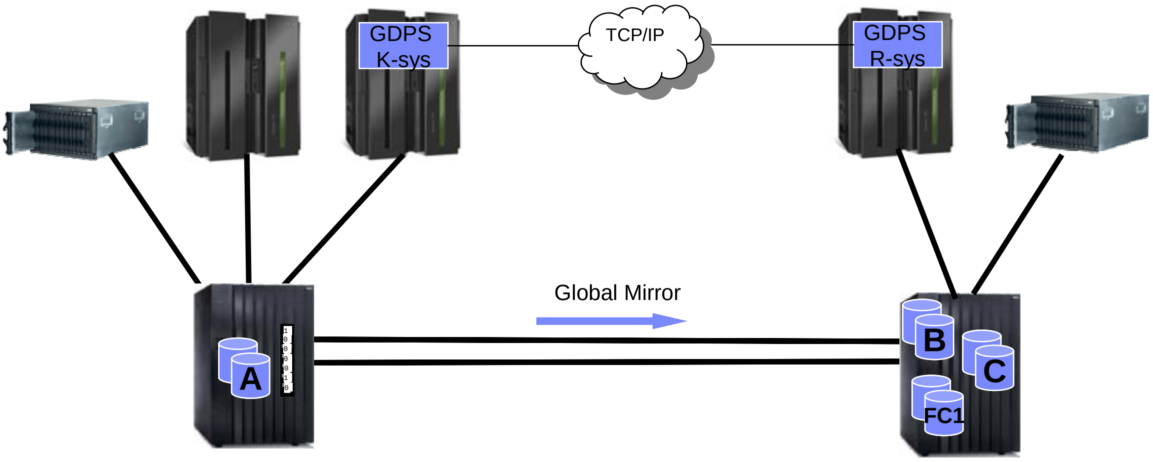


Figure 13 GDPS/GM environment

GDPS provides the capability to use an additional set of devices on the remote site for testing purposes.

GDPS supports both System z and Open Systems devices in a Global Mirror environment. However, GDPS requires that the disk subsystems be shared between the System z and open systems environments, as it requires CKD device addresses in order to issue the commands to manage the environment.

7.2.3 Power HA for i (HASM)

System i PowerHA for i (HASM) is the IBM high availability disk based clustering solution for the IBM i5/OS operating system and is available with V6.1. PowerHA for i supports both Global Mirror and Metro Mirror for data replication when using independent ASPs for application data.

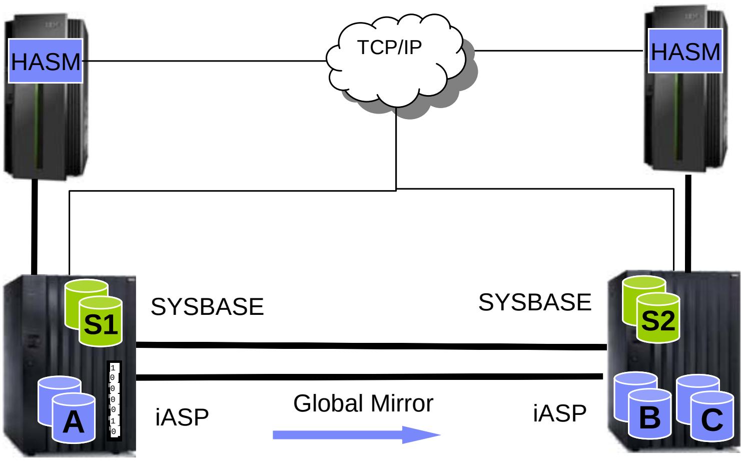


Figure 14 Power HA for i environment

For earlier versions of i5/OS the Copy Services Toolkit provides support for Global Mirror. This also supports the use of Global Mirror to replicate a full system environment if not using independent ASPs.

8. Solution scenarios

This section discusses a number of scenarios and topologies for implementation of Global Mirror.

8.1 Asymmetrical configuration

With an asymmetrical configuration, Global Mirror can only be used from the primary site to the recovery site. This type of configuration would be typical for a disaster recovery configuration where the production systems would run in the secondary location only if there was an unplanned outage of the primary location.

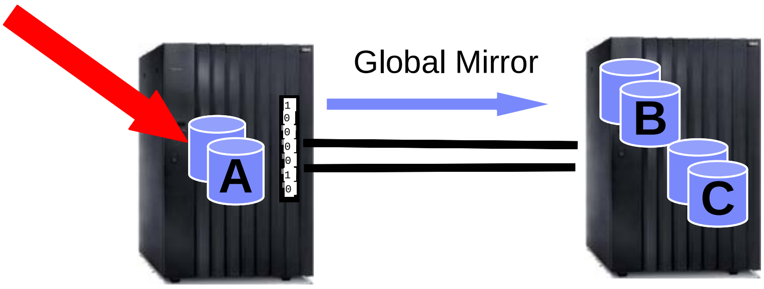


Figure 15 Asymmetrical Global Mirror environment

Once production workloads are moved to the recovery site then Global Copy must be used to return to the primary site. As no disaster recovery capability would be provided in the reverse direction it is unlikely that in this type of configuration we would choose to run for extended periods of time in the secondary location unless forced to by unavailability of the primary site.

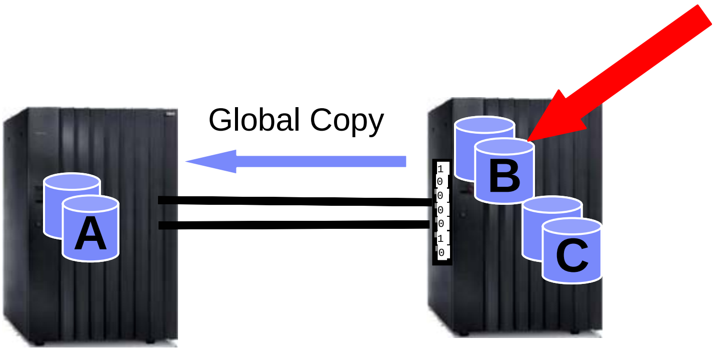


Figure 16 Return to primary site with asymmetrical Global Mirror environment

As Global Mirror uses two copies of data in the secondary location there would be twice as many physical drives in this location as in the production location if the same size drives were used.

In some situations, it may be cost effective to use larger drives in the secondary location. Spreading the production data over all these drives should provide equivalent performance in a disaster situation while reducing the overall cost of the solution.

8.2 Symmetrical configuration

With a symmetrical configuration, we must also supply additional disk capacity for FlashCopy in the primary site. This could also be used for regular FlashCopy, for example for backing up the data without extended outages for the production systems.

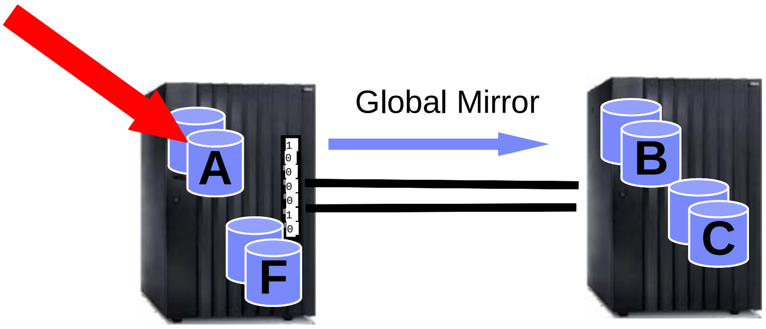


Figure 17 Symmetrical Global Mirror configuration

As we have FlashCopy capacity in 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 the production workloads might run for extended periods of time in either location.

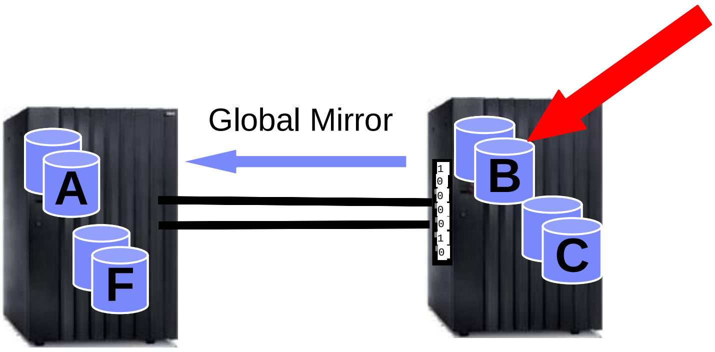


Figure 18 Running in secondary location with symmetrical configuration

8.3 Testing with Global Mirror

Unlike other solutions with only a single copy of data in the secondary location it is possible to perform testing in a Global Mirror environment while maintaining a DR position in the secondary location. However we cannot continue to keep this copy up-to-date, as both copies are required for Global Mirror to run.

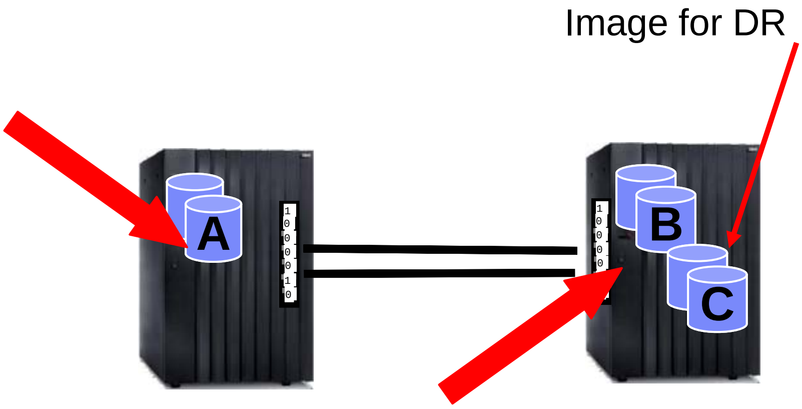


Figure 19 Testing without an additional copy of data

In order to perform testing while the Global Mirror environment is running we need to provide an additional copy of data in the recovery location. Global Mirror is briefly suspended in order to take this copy and then can be restarted while the testing takes place on this extra copy.

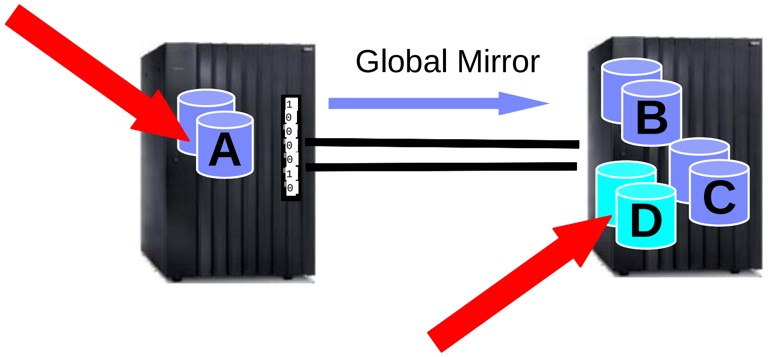


Figure 20 Testing with Global Mirror

8.4 Metro Global Mirror

Metro Global Mirror provides a 3-site or 3-copy solution using both synchronous and asynchronous replication.

This can provide a local synchronous copy of data either to another site within synchronous distance or within the same campus or datacentre. Additionally Global Mirror is used to continually mirror data from the Metro Mirror secondary devices providing an out-of-region copy.

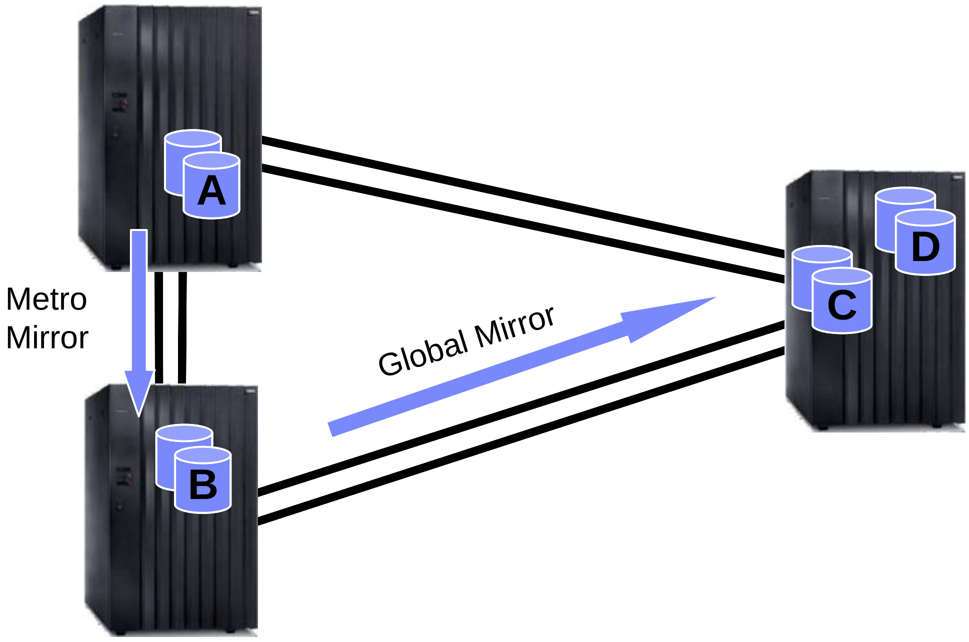


Figure 21 Cascading configuration with Global Mirror

8.5 Controlled Point in Time copy with Global Mirror

It is possible to setup Global Mirror to form a consistency group in a controlled fashion to take a regular but infrequent Point in Time copy of the production workload in a remote location. This might be used to provide an RPO of 6/12/24 hours or to provide a consistent copy of data at particular points relative to business and application activity.

When running in this fashion the normal mode of operation would be running with Global Mirror paused. Global Copy will be sending data from the primary to the secondary disk subsystems ensuring that the difference between primary and secondary is kept to the minimum.

When an update of the consistent copy of data on the secondary disk subsystem is required Global Mirror is resumed and is left running until a single consistency group has formed.Global Mirror is configured with a large consistency group interval to ensure we only form a single consistency group and will then return to running in Global Copy mode.

As soon as the consistency group is formed Global Mirror should be paused to prevent the formation of further consistency groups/

9. Performance considerations

This section contains information on the performance aspects of Global Mirror.

9.1 Managing peak activity

With asynchronous replication solutions that use a cache sidefile on the primary disk subsystem and/or the secondary disk subsystem, only a finite amount of data can be held in the cache. If the mirror falls more than a certain amount of time behind, the replication solution must either pace the production applications or suspend the mirror.

As Global Mirror does not use a cache sidefile, it is possible to deliberately under configure the bandwidth provided in order to reduce the total of the solution. If there are significant peaks then this cost saving could be considerable as the network costs are often a very large portion of the ongoing costs.

Figure 22 shows a typical profile for a production workload with a relatively low write rate during the online day and significant peaks at various points overnight. A bandwidth of around 15MB/s might be provided if a low RPO was required at all points during a 24 hour period.

However if an increased RPO was acceptable at times if high write activity during the overnight period then we might potentially configure as little as 8MB/s of bandwidth which would reduce the network requirements by around 47%.

The minimum bandwidth that can be provided must allow for the formation of consistency groups during at least some periods of the day and must allow for the environment to catch up after any significant periods of delay. This minimum bandwidth will be at least the average write bandwidth after taking account any savings due to duplicate write activity.

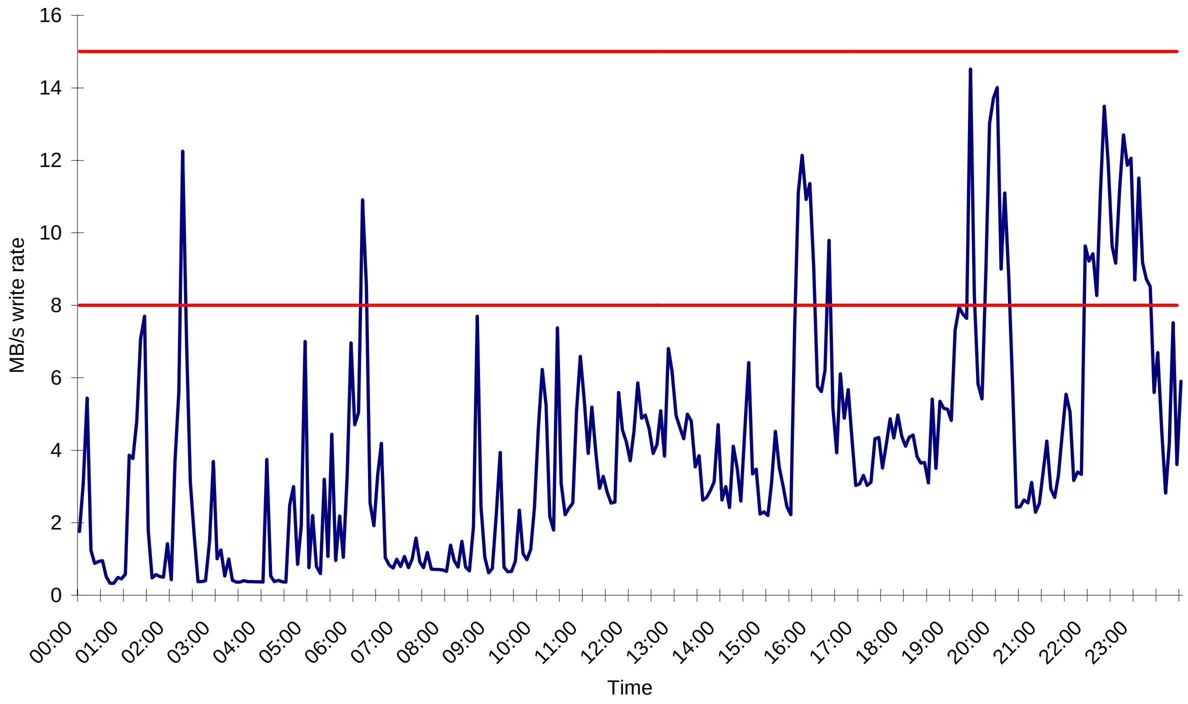


Figure 22 Production workload profile and Global Mirror bandwidth options

With Global Mirror, the production workloads will continue without significant impact during the peak period and the Global Mirror solution will transition automatically to Global Copy mode when a consistency group is not transmitted to the secondary site within the maximum drain time.

When the peak activity has passed and Global Copy is once more able to drain the changed data in a timely manner the disk subsystems will transition automatically back to Global Mirror and resume the formation of consistency groups. At all points, a consistent restartable image of the production data will be available in the recovery location.

9.2 Bandwidth reduction

As Global Mirror removes duplicate updates within a consistency group before sending them to the secondary location less data will tend to be transmitted than is written to the primary devices. The amount of savings here will be workload dependant as well as depending on the interval between consistency group formation.

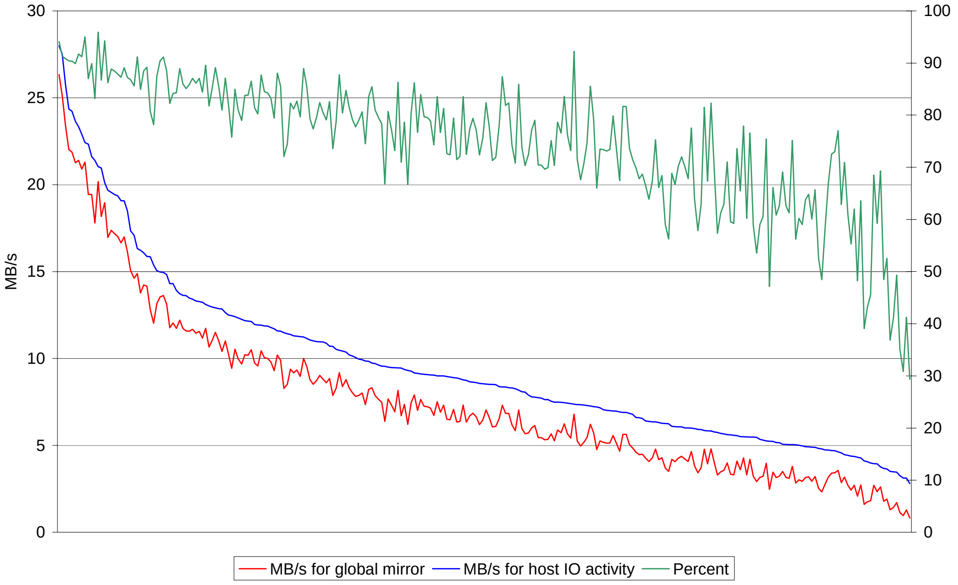


Figure 23 Percentage of production writes sent by Global Mirror

The graph in figure 23 shows the MB/s sent by Global Mirror compared to the production write activity for a particular customer environment over a period of a day. It also shows the percentage of the production writes that were sent by Global Mirror. The data is sorted by host activity from high to low.

In this case, the environment is not bandwidth constrained and the RPO is generally of the order of a few seconds. Even in this case we can see that a reasonable percentage of the production writes do not need to be transmitted to the secondary location.

Another factor we see is that for this workload when the activity is higher the percentage savings are actually lower as for this workload we have a higher proportion of sequential activity at these times. Sequential updates will not generally be duplicated and so the data that is written to the disk subsystem will have to be sent by Global Mirror.

9.3 Primary disk subsystem performance

Global Mirror will protect production application performance at the expense of the currency of the consistent copy at the remote site. Given that a cache sidefile is not used to hold updates there is no requirement to suspend the mirror (requiring manual intervention to resolve) or impact production if the resources available are not sufficient.

In normal operation, the data sent by Global Mirror is contained in the read cache of the primary disk subsystem as it is sent before being prioritised out of cache by the cache management algorithms. If Global Mirror falls behind in it's transmission of data to the secondary disk subsystem then the data that requires to be sent may have to be read from the RAID ranks on the primary disk subsystem.

This activity is performed at a lower priority than production IO and so production performance will be protected at the expense of the speed of data transmission to the secondary disk subsystem.

As well as improving the performance of the production workloads, having a balanced production workload across the available primary disk subsystem resources will improve the capability of Global Mirror to catch up when it falls behind and reduce the increase of the RPO.

9.4 Performance at distance

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

Distance can impact 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 but as distances grow extremely large there is some impact that will be experienced.

Figure 24 shows the impact of extremely long distances on the RPO of a Global Mirror solution in a test environment. With improvements delivered in the R2.4 of the DS8000 microcode, there is some small increase in the RPO at extremely long distances but the effect was not significant.

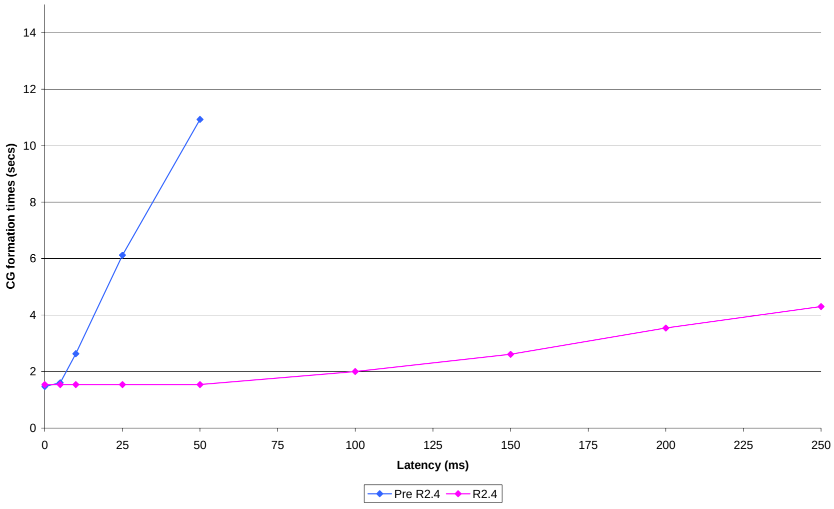


Figure 24 RPO of Global Mirror at very long distances

In order to send large amounts of data at long distances a significant degree of parallelism is required to ensure that the bandwidth can be filled. If there is not enough parallelism then the throughput will be reduced as more and more time is spend waiting for acknowledgements indicating that data has been received at the remote location. However at shorter distances the same degree of parallelism may be counter-productive.

Global Mirror provides an extreme distance RPQ for environments where the distance/latency between the local and remote sites is over 5000km.

The affect of distance is felt more for smaller updates as the bandwidth is less likely to be the bottleneck than for larger updates. Figure 25 shows the results of some extreme distance testing for 4KB random writes both with the default Global Mirror settings and with the changes provided by the extreme distance RPQ.

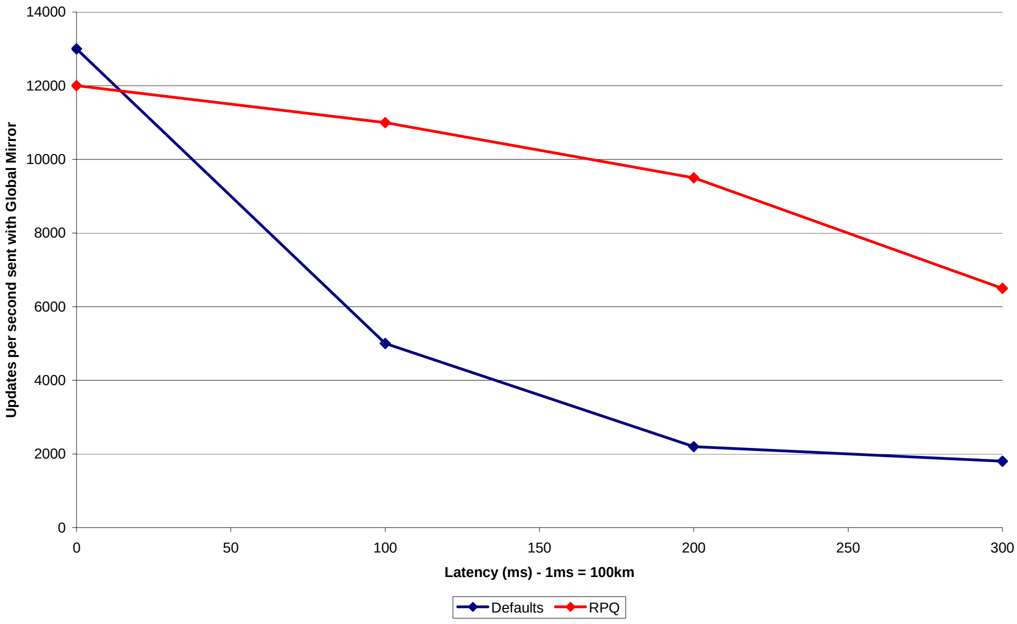


Figure 25 Throughput of Global Mirror at very long distances

Note : Global Mirror will send multiple updates to the same track in a single operation so for write streams such as a database log the number of writes from the server will be significantly higher than the number of individual updates that must be sent by Global Mirror. On open systems 16 4KB writes from a database log could be sent in a single operation.

9.5 Secondary disk subsystem performance

FlashCopy is used by Global Mirror to save consistency groups. The NOCOPY option is used here so the only background copy activity is caused by copy on write processing for the updates between each consistency group.

Figure 26 shows the process for the copy on write activity on a Global Mirror secondary. A PPRC write will occur to a track on the B device. So long as there is space in NVS this write will be acknowledged immediately. Later the previous track on the B device must be copied to the C device before allowing the update to be destaged.

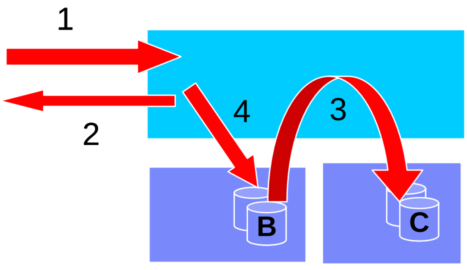


Figure 26 Copy on write processing

In normal operation when forming consistency groups every 3-5 seconds then the copy on write processing does not occur as we do not need to perform the destage before the next consistency group occurs. However when Global Mirror falls behind or is catching up there may be additional activity on the secondary RAID ranks.

It is possible that in some environments, an unbalanced workload may result in bottlenecks being seen on the RAID ranks of the secondary disk subsystems. Features such as Arrays Across Loops and the increased RAID rank bandwidth of switched fibre channel disks the chance of this occurring is much reduced. However there are a number of optimisations that can be made to improve the performance of the Copy on Write processing by following some simple rules for placement of devices on the secondary disk subsystem.

9.5.1 Placement of devices on secondary disk subsystem

In order to spread the load of any hotspots on the secondary RAID ranks it is recommended to split the FlashCopy source and target volumes on separate RAID ranks and device adapters. In this way if we have particular busy volumes the B volume RAID rank will be writing and reading and the C volume RAID rank will be writing rather than a single RAID rank performing all the activity.

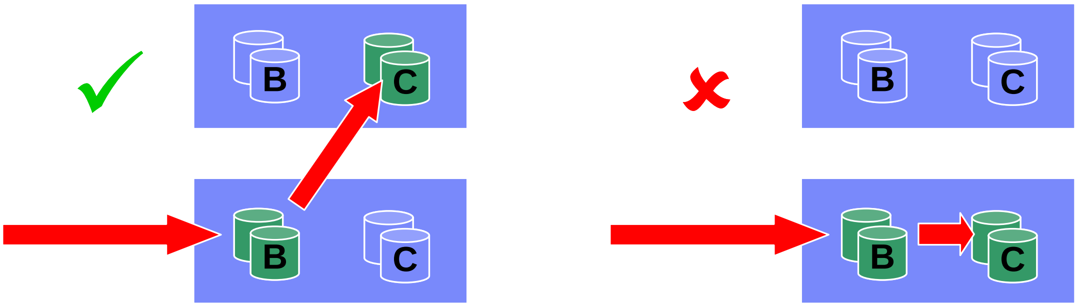


Figure 27 Spreading B and C volumes over different ranks

If we have an additional D copy for testing purposes then we might either choose to place the D volumes on the same ranks as the B and C volumes. If we intended to perform a lot of heavy activity such as backup processing or stress testing then we might dedicate a rank to the D volumes in order to protect the Global Mirror environment from this workload.

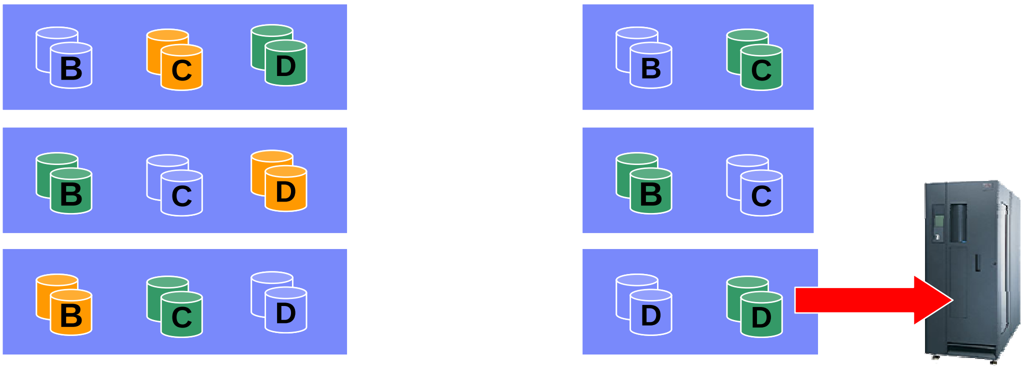


Figure 28 Options for placement of D volumes

A customer might also choose to implement smaller faster drives for the B volumes but larger drives for the C and D volumes in order to take advantage of the cost efficiencies in doing so. This does result in the ranks containing the B volumes having a higher workload that the larger ranks and so might not be recommended for optimal performance. Using a single drive size and spreading all the devices over all drives is generally a better approach.

A better configuration might be to spread the B and C volumes over both sets of drives for optimal performance of both Global Mirror and the production workloads. The D volumes would then utilise the additional space on the larger drives.

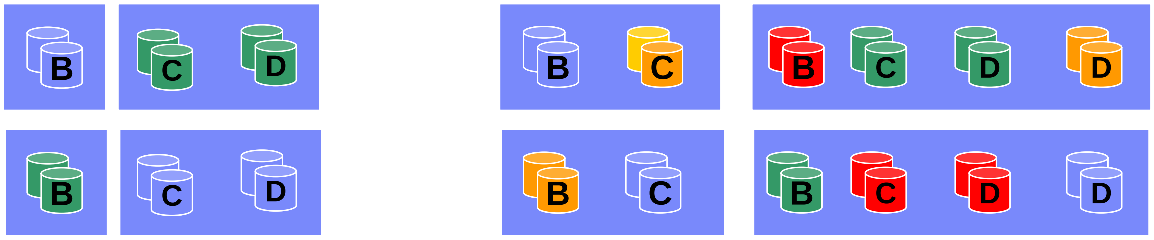


Figure 29 Volume layout with two drive sizes on secondary disk subsystem

There have also been further optimisations to FlashCopy that enhance the Copy on Write processing when the FlashCopy source and target are managed by the same storage server processor complex (cluster) on the disk subsystem. This means that the recommendation is to have a particular FlashCopy source and target both on either odd or even LSS.

10. References

This section contains references to other publications and websites containing further information on Global Mirror.

10.1 Redbooks, Manuals and Whitepapers

SG24-6787	IBM TotalStorage DS8000 Series: Copy Services with IBM eServer System z
SG24-6788	IBM TotalStorage DS8000 Series: Copy Services in Open Environments
SG24-6782	IBM TotalStorage DS6000 Series: Copy Services with IBM eServer System z
SG24-6783	IBM TotalStorage DS6000 Series: Copy Services in Open Environments
SG24-7596	IBM TotalStorage Productivity Center for Replication Using DS8000
SG24-7120	System i and IBM TotalStorage: A Guide to Implementing External Disk on eServer i5
GG24-6374	GDPS Concepts and Facilities
SC35-0428	DFSMS Advanced Copy Services
GC26-7681	IBM TotalStorage DS6000 Command-Line Interface User's Guide
SC26-7625	IBM TotalStorage DS Command-Line Interface User's Guide
SC32-0103	IBM TotalStorage Productivity Center for Replication Users Guide

10.2 Websites

IBM Storage Business Continuity website

http://www-03.ibm.com/systems/storage/solutions/business_continuity/index.html

GDPS website

http://www-03.ibm.com/servers/eserver/System z/resiliency/gdps.html

TPC for Replication website

http://www-03.ibm.com/systems/storage/software/center/replication/index.html

IBM PowerHA website

http://www-03.ibm.com/systems/power/software/availability/index.html

System i Copy services toolkit website

http://www-03.ibm.com/systems/i/support/itc/copyservicessystemi.html