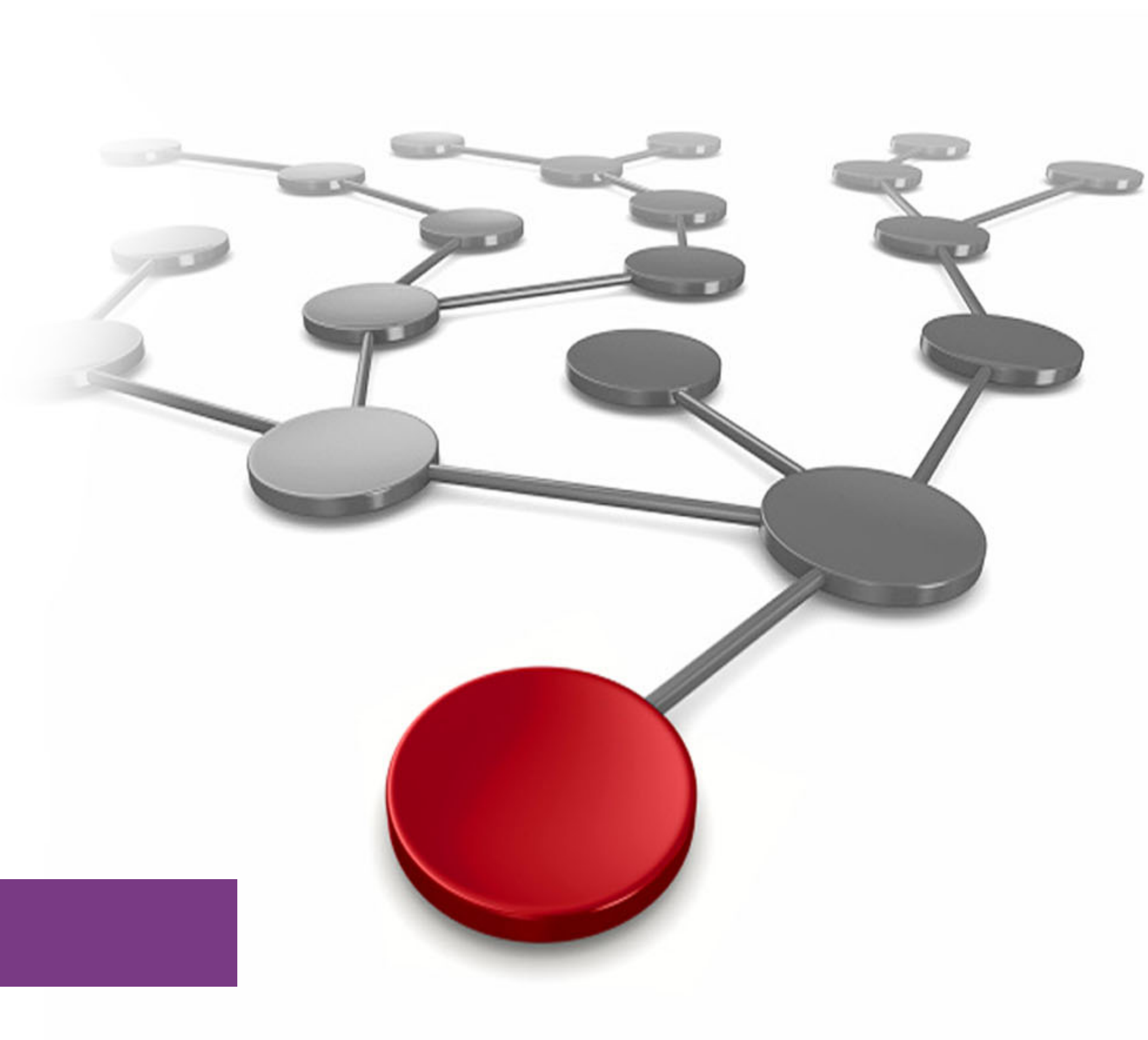
ibm.com/redbooks

IBM DS8000 and IBM Z Synergy

DS8000 Release 9.3 and z/OS 2.5

Peter Kimmel

Jörg Klemm



Storage

Redpaper

IBM Redbooks

IBM DS8000 and IBM Z Synergy (DS8000 Release 9.3 and z/OS 2.5)

July 2022

REDP-5186-07

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

Eighth Edition (July 2022)

This edition applies to z/OS Version 2, Release 5, and DS8000 Release 9.3.

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®
CICS®
Db2®
DS8000®
Easy Tier®
FICON®
FlashCopy®
GDPS®
HyperSwap®
IBM®
IBM Cloud®
IBM Security™
IBM Z®
IBM z Systems®
IBM z13®
IBM z14®
Parallel Sysplex®
POWER®
POWER9™
RACF®
Redbooks®
Redbooks (logo) ®
System z®
z Systems®
z/Architecture®
z/OS®
z/VM®
z/VSE®
z13®
z15™
zEnterprise®

The following terms are trademarks of other companies:

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Microsoft, 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

IBM Z® has a close and unique relationship to its storage. Over the years, improvements to the IBM zSystems® processors and storage software, the disk storage systems, and their communication architecture consistently reinforced this synergy.

This IBM® Redpaper™ publication summarizes and highlights the various aspects, advanced functions, and technologies that are often pioneered by IBM and make IBM Z and IBM DS8000® products an ideal combination.

This paper is intended for users who have some familiarity with IBM Z and the IBM DS8000 series and want a condensed but comprehensive overview of the synergy items up to the IBM z16 server with IBM z/OS® V2.5 and the IBM DS8900 Release 9.3 firmware.

Authors

This paper was produced by a team of specialists from around the world:

Peter Kimmel is an IT Specialist, IBM Redbooks Project Leader, and Advanced Technical Skills team lead of the Enterprise Storage Solutions team at the IBM ESCC in Frankfurt, Germany. He joined IBM Storage in 1999, and since then has worked with all DS8000 generations, with a focus on architecture and performance. Peter co-authored several DS8000 IBM publications. He holds a Diploma (MSc) degree in physics from the University of Kaiserslautern.

Jörg Klemm is a Senior Mainframe Consultant working for IBM Platinum Business Partner SVA System Vertrieb Alexander GmbH in Germany. He has over 20 years of experience in IBM working directly with customers, primarily focused on enterprise storage and specialized in Business Continuity solutions. His areas of expertise include Copy Services and GDPS. Jörg has been delivering GDPS solutions for over 15 years.

Thanks to the authors of the previous editions of this paper:

Bertrand Dufrasne
Alexander Warmuth
Ewerson Palacio
Marlon Cerqueira
Andre Coelho
Marcelo Morae
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 IBM Redbooks 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, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on LinkedIn:

  https://www.linkedin.com/groups/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

Introduction

This chapter provides background information about the topics that are covered in this paper and the reasons that make the IBM DS8000 storage system and IBM Z servers an ideal combination.

Although most of the synergy items that are described here apply when the DS8900F storage system is combined with a comprehensive z/OS software level, some of the items also apply to other operating systems that deploy on IBM Z, including Linux on IBM Z, IBM z/TPF, IBM z/VM®, and IBM z/VSE®.

This chapter includes the following topics:

1.1 IBM DS8000 storage system and IBM Z synergy

This section describes the historically strong ties between the IBM DS8000 storage system and IBM Z.

1.1.1 IBM Z server heritage

IBM owns the architecture of the IBM Z hardware and fundamental software, such as IBM z/OS, z/VM, Linux on IBM System z®, z/TPF, or z/VSE with their components and subsystems, such as Input/Output Supervisor (IOS), Channel Subsystem, and access methods to data on connected disk storage systems.

Since the advent of mainframe servers in 1964 with the IBM System/360 server family, this server architecture was continuously enhanced and developed to provide the most reliable application server on the market while still maintaining operational efficiency. IBM Z today can reach any level of scalability that is needed to run the most demanding workloads.

1.1.2 Disk storage system heritage

With disk storage systems, IBM also has a rich history and extensive experience. IBM created the first randomly accessed disk storage system in 1956 and pioneered many technological breakthrough achievements. IBM invented, developed, and built more advanced disk storage systems, which culminated today with the most advanced disk storage server: IBM DS8000 storage systems, and its flagship, the DS8900F.

The DS8900F features three system types:

- DS8910F Flexibility Class models 993 and 994
- DS8950F Agility Class models 996 and E96
- DS8980F Analytic Class models 998 and E96 (recently announced)

1.1.3 Connecting layer

The piece that brings these products together and helps them interact optimally is a connection technology that IBM also invented and continuously enhances and improves.

The technology’s first appearance was in 1990 when IBM presented a newly created serial I/O protocol between IBM Z and storage systems called *ESCON*. It was implemented to run through optical fiber technology to replace the copper-based bus and tag channel interface technology. It also overcame the performance and distance limitations that were imposed by these parallel channels.

In 1998, this serial I/O interface technology was enhanced to become the IBM Fibre Channel connection (IBM FICON®) technology by using full-duplex Fibre Channel and multiplexing capabilities. FICON is the IBM interpretation of Fibre Channel technology that IBM enhanced for IBM Z optimal connection performance, reliability, and security.

In 2017, IBM announced IBM zHyperLink technology for the IBM z14 server and the DS8880 storage system. IBM zHyperLink works with a FICON point to point or storage area network (SAN) infrastructure to provide low-latency connectivity to FICON storage systems. With the arrival of the IBM z15 and the DS8900F all-flash storage system family in late 2019, IBM zHyperLink was further improved. With IBM z16, new features, such as data in-flight encryption and 32 GFC connections, are now available.

1.1.4 Putting the pieces together: Synergy

Because IBM owns all of these building blocks and their respective architectures, integrating these components to get the best overall efficiency from the complete solution is easier. This *synergy* is based on the fact that each component understands the potential of other components and can best use it. Integrating the DS8000 storage system, IBM Z, and communication components result in a combination that offers more than just the sum of its individual components.

1.1.5 Mainframe and storage end-to-end configuration model

A schematic view of the key layers when combined from the end-to-end and integrated IBM Z and DS8000 configuration model is shown in Figure .

The key layers include the following major components:

- IBM Z and z/OS operating system, including z/VM, Linux on zSystems, z/TPF, and z/VSE.
- IBM DS8000 storage system, particularly the DS8900F storage system with its rich functions, such as IBM zHyperLink.
- Optional SAN fabric devices that support FICON, provided mainly by Brocade and Cisco.

  This layer includes end to end intelligent and advanced error detection and error recovery. A directly connected configuration also is available.

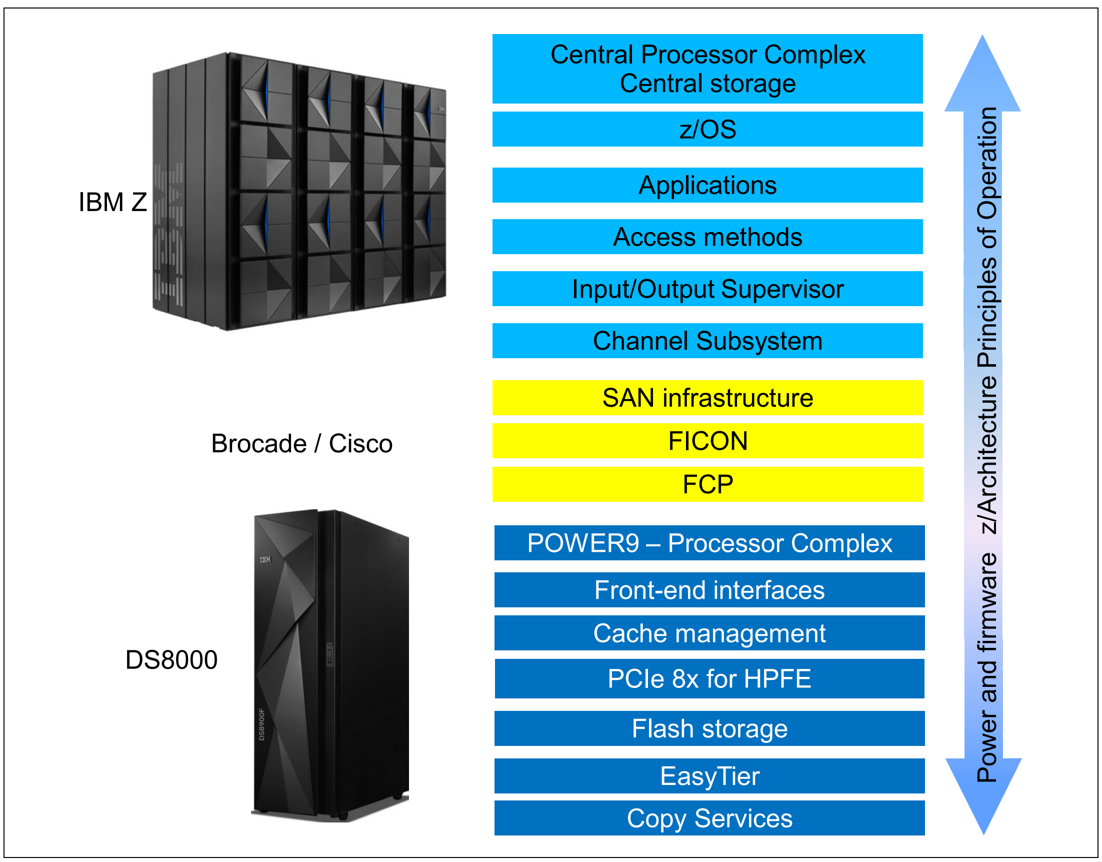


Figure 1-1 Synergy potential within the mainframe configuration model

Figure 1-1 on page 3 shows the close and integrated cooperation between the IBM Z server and the DS8000 storage system. It also indicates the main ingredients of each building block within the mainframe configuration model. In particular, the DS8000 storage system includes the following building blocks:

- Comprehensive IBM Power Systems servers are at the heart of the DS8000 storage system.
- Central storage or server memory that provides firmware-assisted caching services.
- Host adapters to connect through the front end to application servers.
- Device adapters that connect through PCIe x8 for attachment to High-Performance Flash Enclosures (HPFEs) on the back-end storage.
- Rich IBM Copy Services and IBM Easy Tier® functions.

Figure 1-1 on page 3 also shows the foundation upon which the synergy items are built that are described in this paper.

1.2 Synergy items

This section highlights the most important and popular synergy items of IBM Z server and DS8000 storage system.

1.2.1 Disaster recovery and high availability items

For more information about the following IBM Copy Services functions that the DS8000 storage system provides, see Chapter 2, “Disaster recovery, high availability, and cyber resiliency” on page 7:

- IBM FlashCopy® with all its options

  IBM FlashCopy is a popular function and the foundation for solutions, such as the IBM Db2® BACKUP SYSTEM utility. BACKUP SYSTEM is closely interconnected in z/OS with Data Facility Storage Management Subsystem (DFSMS) software constructs, such as Copy Pools and Copy Pool Backup Storage Groups (CPBSGs).
- Global Copy

  This function is used to asynchronously replicate volumes without ensuring consistent data at the target site.
- Global Mirror (GM)

  This function is used to asynchronously replicate volumes while ensuring consistent data at the target site by combining Global Copy and FlashCopy.
- Metro Mirror (MM)

  This popular function synchronously replicates data from a source volume to a target volume.
- Safeguarded Copy

  Safeguarded Copy is a solution for Logical Corruption Protection (LCP). It must be managed with IBM Copy Services Manager (CSM) or IBM Geographically Dispersed Parallel Sysplex® (IBM GDPS®). The IBM Z Cyber Vault solution is based on these elements.

- IBM z/OS Global Mirror (IBM zGM)

  This 2-site disaster recovery (DR) solution is formerly known as *Extended Remote Copy* (XRC). IBM zGM still requires proper firmware support in the primary or source storage system but relies mainly on its software-based component, which is the System Data Mover (SDM) within z/OS.

- Three-site solutions

  These solutions include cascaded Metro/Global Mirror (MGM) and Multiple Target Peer-to-Peer Remote Copy (MT-PPRC), which provide two copies that are based on a single source volume. These copies can be synchronously replicated copies or asynchronously replicated copies, or a mix of synchronous and asynchronous replications. Another three-site multi-target volume replication relationship consists of an MM relationship and an IBM zGM replication relationship from the same primary or source volume.

- Four-site solutions

  These solutions are also known as *symmetrical solutions* and provide a high availability (HA) environment within a metropolitan region with an active DR configuration to another region. This configuration takes advantage of cascaded and MT-PPRC functions. In this case, both regions have a symmetrical configuration. Therefore, it has HA and DR (HA/DR) protection, regardless of where the system is running.

> **Note:** All copy services functions can be managed through CSM or GDPS.

1.2.2 Data protection and backup

This support is integrated into system-managed storage within z/OS and can be used by any other utility or application.

A user of this software-based support is the BACKUP SYSTEM Db2 utility. Its availability started with z/OS V1.8 and was significantly enhanced in subsequent versions; for example, to include FlashCopy Consistency Groups support.

For more information about the software-based interface to FlashCopy, see Chapter 3, “Data protection and backup” on page 33.

1.2.3 More information

For more information about managing and configuring synergy items, see Chapter 4, “Management and configuration” on page 45.

For more information about the DS8000 storage system and z/OS performance synergy items, see Chapter 5, “IBM Z and DS8000 performance” on page 71.

2

Disaster recovery, high availability, and cyber resiliency

This chapter describes the various data and volume replication techniques that are implemented in DS8000 Copy Services, which provide the foundations for disaster recovery (DR) operations and enable high data availability (HA) and cyber resiliency.

DS8000 Copy Services are complemented by management frameworks and functions that are built directly into the IBM Z software and firmware. This functional complementarity allows a further enhanced automation approach that can handle almost any potential incident, even in an unattended environment that is composed of IBM Z servers and DS8000 storage systems.

This chapter includes the following topics:

2.1 DS8000 Copy Services functions

The DS8000 storage systems provide broad and rich copy services functions. They can be used for 2-, 3-, or 4-site solutions. These scenarios are described in this chapter.

Note: For more information, see *IBM DS8000 Copy Services: Updated for IBM DS8000 Release 9.1*, SG24-8367.

2.1.1 Metro Mirror 2-site synchronous volume replication

Metro Mirror (MM) is the synchronous volume replication approach that uses DS8000 firmware. The technology that is used is known as Peer-to-Peer Remote Copy (PPRC).

Figure 2-1 shows the basic sequence of operations. The goal is to ensure the safe and consistent data arrival at the receiving site of MM with the least cycles possible between both sites. MM achieves this goal quickly.

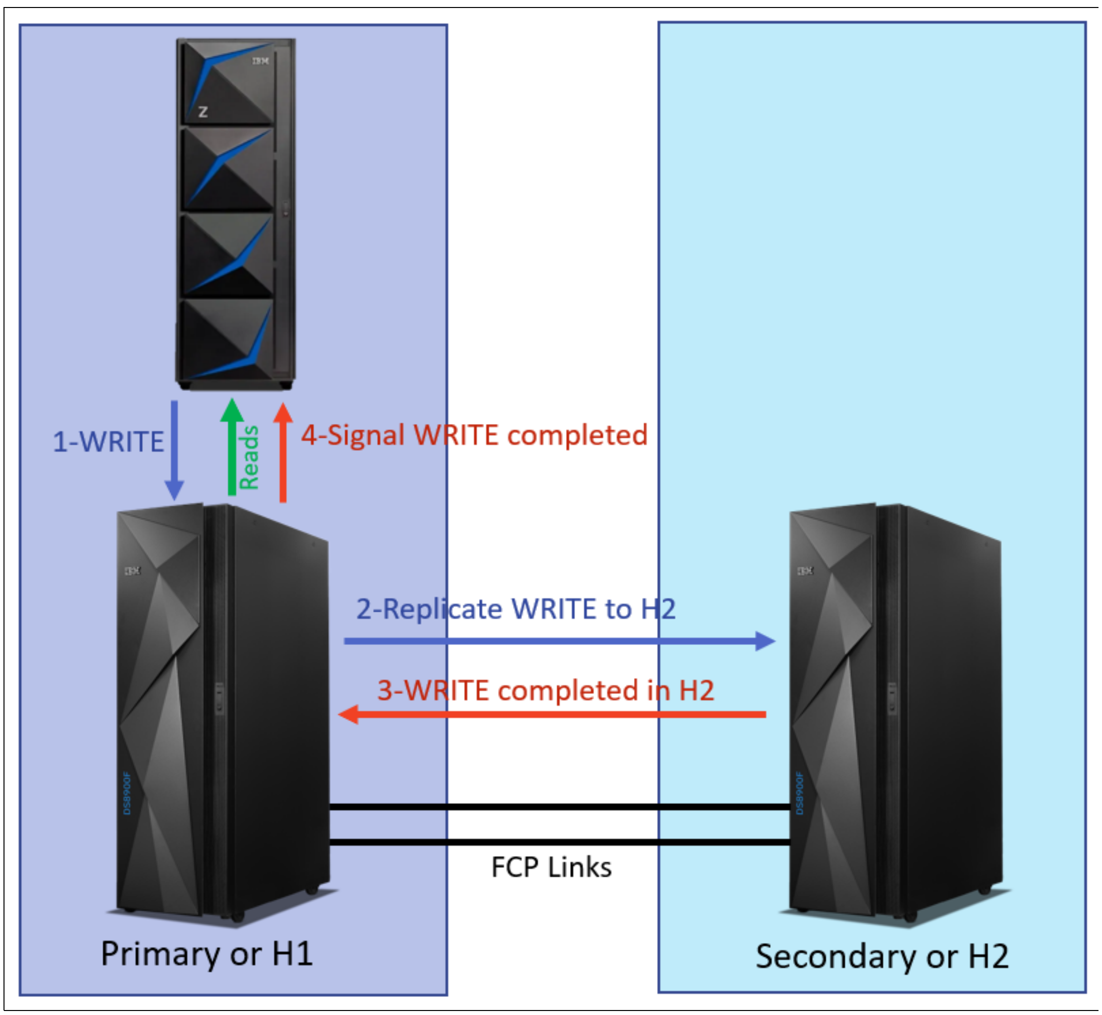


Figure 2-1 Metro Mirror basic operation

A balanced approach between pre-deposited MM writes at the receiving site, and occasional feedback from the receiving control unit (CU) back to the sending CU allows the use of the Fibre Channel links and MM paths between both sites as efficiently as possible.

Before Release 9.2, in an MM full-duplex scenario, the auxiliary storage was not directly accessed for reads, even though it was physically closer to the host. In Figure 2-1 on page 8, if IBM HyperSwap® was used, the host stops reading from H1, and all read I/O are done from H2 after it is made the primary storage.

The same situation occurs when a full duplex with an active-active configuration is used. Figure 2-2 shows the I/O process before a HyperSwap in a traditional environment. The host at site 2 reads data from H1 after it becomes the primary storage.

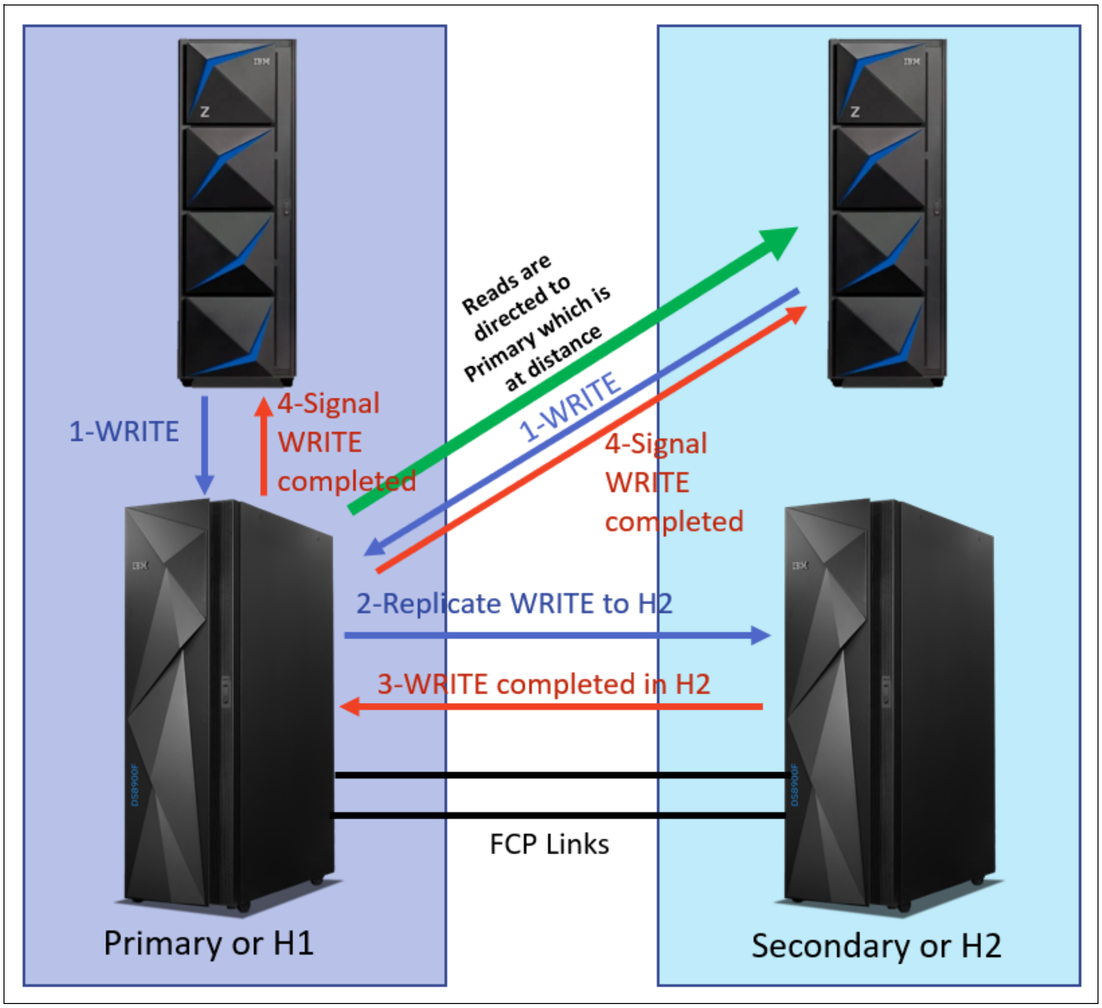


Figure 2-2 Metro Mirror without Consistent Read from Secondary

When HyperSwap is done from primary to the auxiliary storage on site 2, the host server at site 1 sends all read I/Os to the primary storage, which is now the H2 on site 2. Now, all the reads are at a distance from the site 1 host perspective, as shown in Figure 2-3.

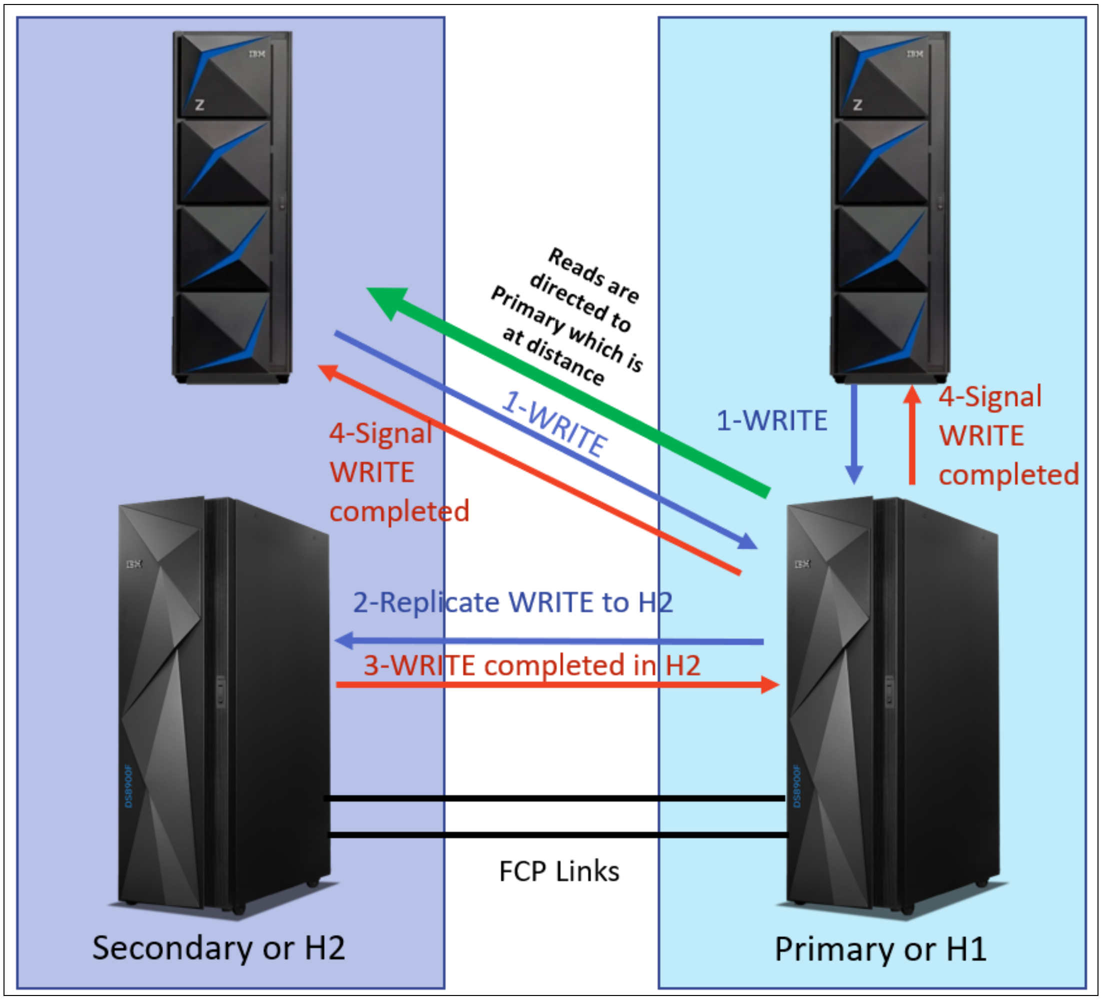


Figure 2-3 Metro Mirror after HyperSwap without Consistent Read from Secondary

With Release 9.2, IBM introduced Consistent Read from Secondary (CRS), which allows applications reads to run through the auxiliary storage server and avoid extra overhead that is caused by the distance between the host server and the primary storage, as shown in Figure 2-4.

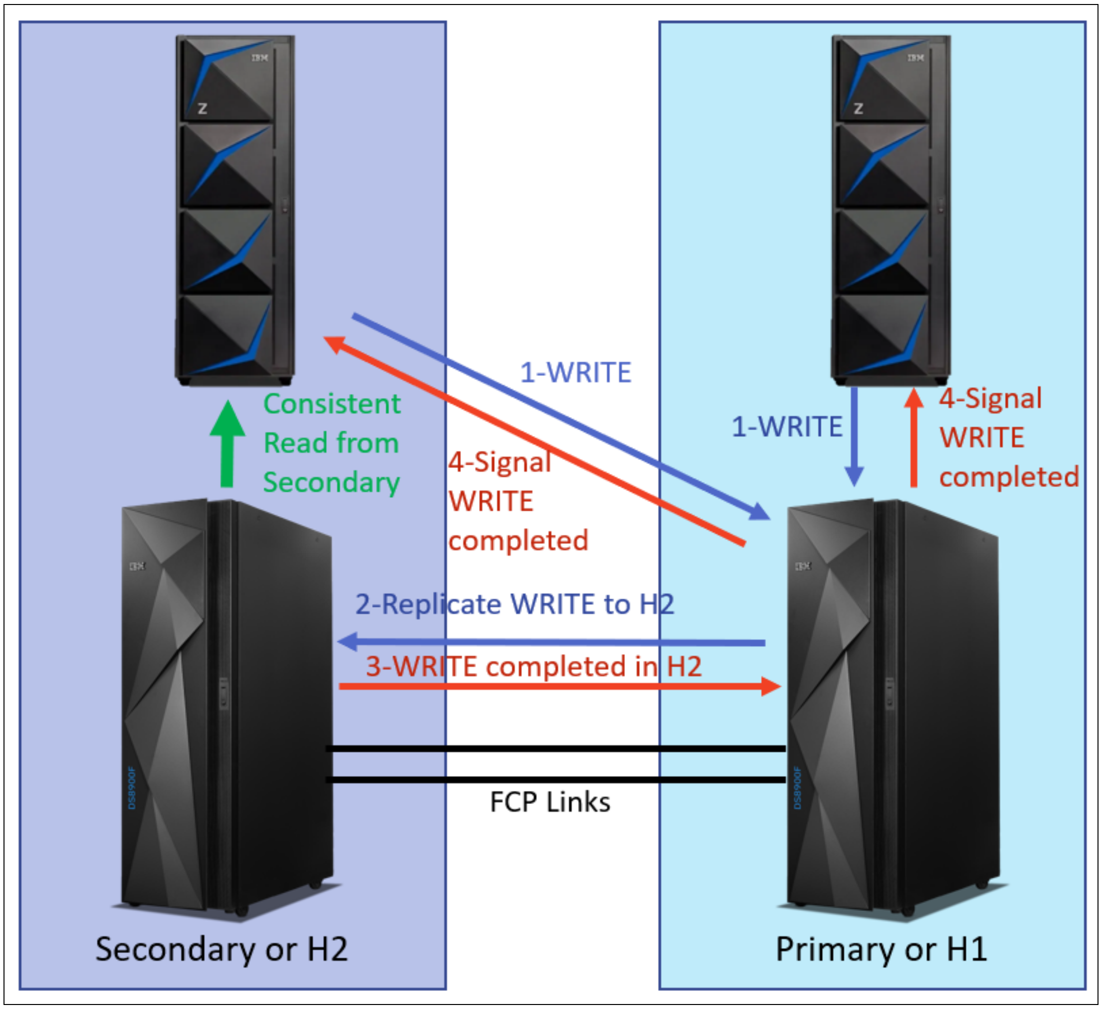


Figure 2-4 Metro Mirror after HyperSwap with Consistent Read from Secondary

For more information about CRS, see 2.3, “Consistent Read from Secondary” on page 20.

2.1.2 Global Mirror 2-site asynchronous volume replication

IBM offers a software-based solution to provide a replication technology that can bridge any distance and still ensure consistent data at the remote site. The solution is known as Global Mirror (GM). The goal is to provide an asynchronous replication technology that can run in an autonomic fashion and provide data currency (within no more than 5 seconds) at a distant site.

The data consistency at a distant site is ensured for a single storage server pair and across as many as 16 storage systems. This consistency is made possible without other constraints, such as imposing timestamps on each write I/O.

The basic operational sequence of GM is shown in Figure 2-5.

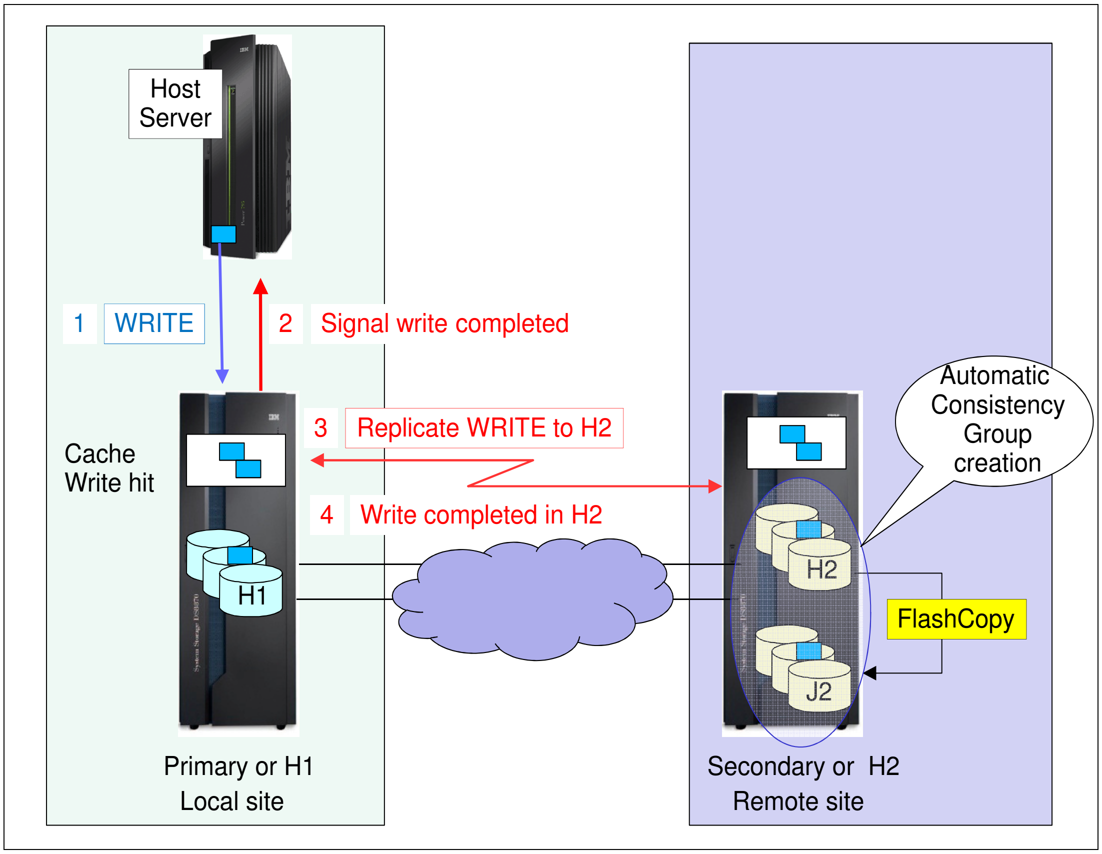


Figure 2-5 Global Mirror basic operation managed out of the storage system

From a host perspective, the write I/O behaves as though it is writing to a nonmirrored volume. The host receives an I/O completion event when the write data arrives in the cache and nonvolatile cache portion of the DS8000 cache. Then, the DS8000 storage system asynchronously replicates the data and sends it to the remote site. The replication I/O is completed when the data is secured in the remote cache and remote non-volatile cache.

GM combines the Global Copy and FlashCopy functions. Global Copy performs the data replication, and FlashCopy secures the previous data from H2 onto J2 before the respective track on H2 is overwritten by the replication I/O. Therefore, J2 behaves as a journal for H2.

The GM consistency group creation process is solely performed within the DS8000 storage systems. Synergy comes into play when managing such a configuration through IBM Copy Services Manager (CSM) or IBM Geographically Dispersed Parallel Sysplex (GDPS).

2.1.3 z/OS Global Mirror 2-site asynchronous volume replication

Important: The IBM DS8900F family is the last platform to support z/OS Global Mirror. New z/OS Global Mirror functions are *not* provided with IBM DS8900F. For more information, see IBM Announcement Letter 920-001.

IBM zGM essentially is z/OS software-based asynchronous volume replication. The design goal for IBM zGM was to not exceed 5 seconds in remaining current with the primary site. IBM zGM relies on timestamped ECKD write I/Os for each write I/O to each IBM zGM primary Count Key Data (CKD) volume. IBM zGM can manage only CKD volumes.

The basic components of IBM zGM operations are shown in Figure 2-6.

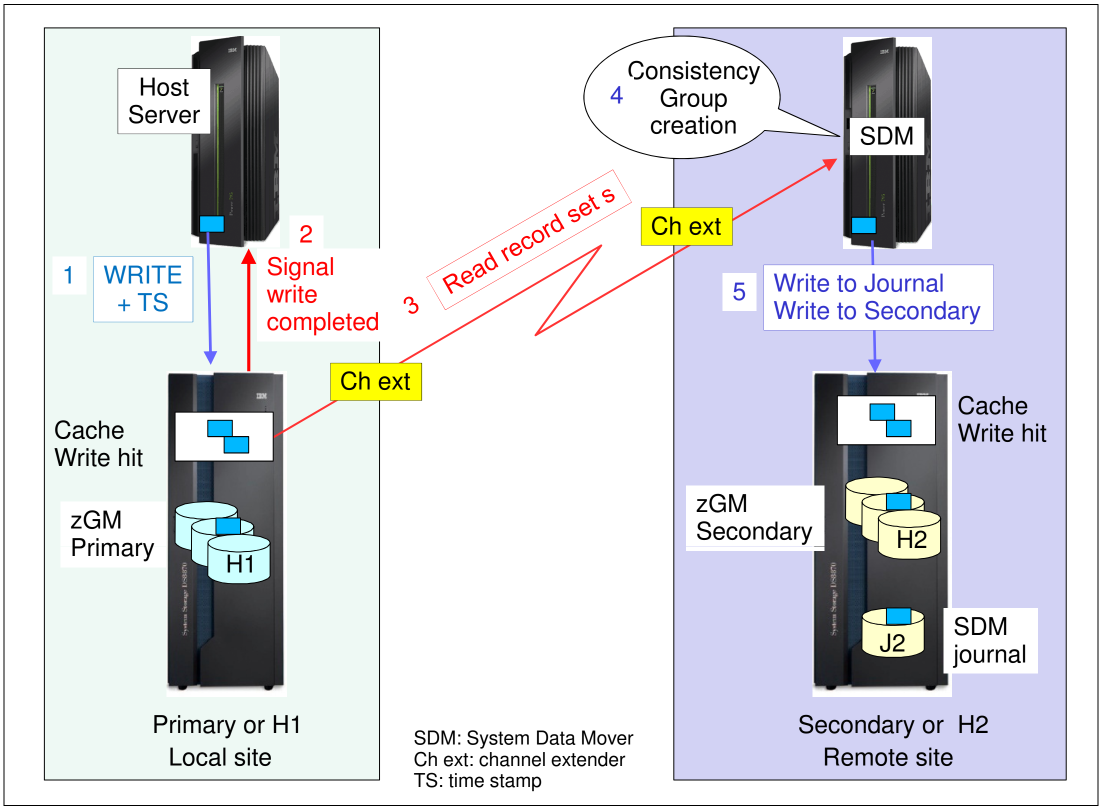


Figure 2-6 IBM zGM basic operation that is managed through IBM Z software

Figure 2-6 shows how closely IBM Z software cooperates with the DS8000 storage system:

1. Application write I/Os perform at the same speed as with writing to an unmirrored volume in H1. Each write I/O also contains a unique timestamp. The IBM Parallel Sysplex® Timer clock is used.
2. Immediately after successfully storing the data to cache and nonvolatile cache storage in the DS8000 storage system, the I/O is complete from an application perspective.
3. System Data Mover (SDM) is a highly parallel working driver that fetches the data from the H1 site as fast as possible by using particular enhancements in z/OS and its Input/Output Supervisor (IOS). Any bandwidth between sites can be used by the SDM and its multiple-reader support.
4. SDM internally sorts all write I/Os according to the applied timestamp during application write I/O processing to ensure the same write order to the secondary volumes as they occurred to the primary volumes.
5. To resume operations after an unplanned outage of any component within the remote site, SDM applies first the consistency groups onto a journal. Next, SDM writes the same consistency group (or groups) to the secondary volumes and then frees the corresponding journal space.

After IBM zGM reaches a balanced system level (combining all involved components), it is a firm solution that runs unnoticed. The key requirement is to provide enough bandwidth between sites and on the storage backend at the recovery site to manage the amount of write data arriving at the local site.

2.1.4 Metro/Global Mirror 3-site solution

Metro/Global Mirror (MGM) is solely based on the DS8000 firmware. It is a cascaded approach that spans over three sites.

The first leg is an MM relationship from site 1 to site 2. Then, the journey continues to a potentially distant site 3 with GM. The role of the site 2 volumes is a cascaded status. The same volume in site 2 is an MM secondary volume in a DUPLEX state while being a Global Copy primary volume with a PENDING status.

The basic components of an MGM configuration are shown in Figure 2-7.

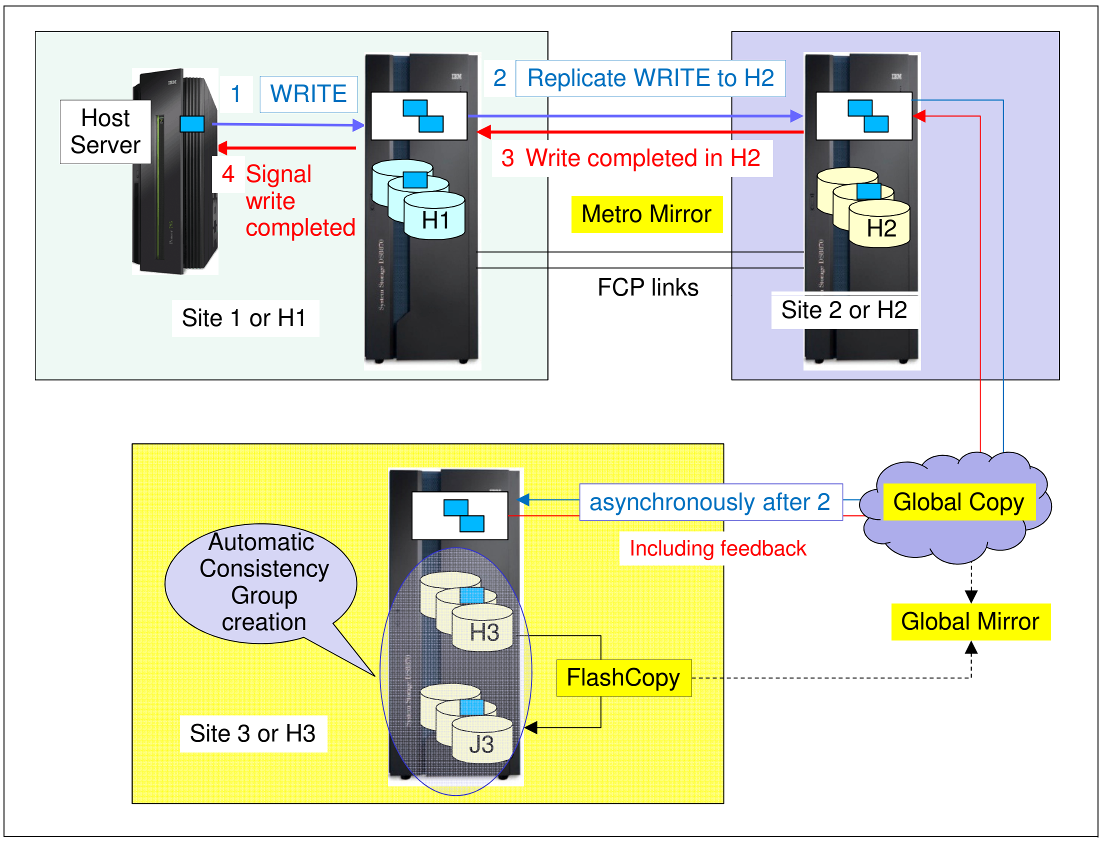


Figure 2-7 Metro/Global Mirror cascaded 3-site solution

Often, site 1 and site 2 are in a campus or metropolitan area within MM acceptable distances. In a typical IBM Z environment, both sites are usually at a distance that is also supported by the Parallel Sysplex architecture and spread the IBM Z server across site 1 and site 2.

Both sites might be only a few kilometers apart from each other to allow an efficient data-sharing approach within the coupling facility. MM acceptable distances are often measured from approximately 100 meters (328 feet) to approximately 5 - 6 km (3.10 - 3.72 miles) because the synergy of this configuration relies on Parallel Sysplex functions and MM in combination with HyperSwap.

When coupling facility-based data sharing is not a potential issue, the distance can be as much as the supported distance by the Parallel Sysplex Timer, which is up to 100 km (62.13 miles) between site 1 and site 2. Then, another synergy item plays a significant role.

When a HyperSwap occurs, and application writes run from site 1 application servers to the site 2 storage systems, the High-Performance FICON for IBM Z (zHPF) Extended Distance (ED) II support improves the performance of large write I/Os.

Site 1 and site 2 fulfill the role of DR covering site 1 and site 2 and HA when site 1 components fail or the storage server (or servers) experience any type of outage. Site 3 is a pure DR site when site 1 and site 2 are no longer available.

Important: A management framework, such as CSM or GDPS, is required to manage a 3- or 4-site volume replication configuration.

2.1.5 Multiple-Target Peer-to-Peer Remote Copy 3-site solutions

Multiple-Target Peer-to-Peer Remote Copy (MT-PPRC) was available since DS8870 firmware levels 7.4 or later. It is also a 3-site Copy Services configuration.

It is called MT-PPRC because it can have two Copy Services relationships that are based on volume copies in site 2 with another relationship in site 3. These Copy Services relationships can be either of the following approaches:

- Two MM relationships of the same site 1 volume
- A combination of an MM relationship between site 1 and site 2 and a second GM or Global Copy relationship from site 1 to another site 3

Both approaches are described next.

MT-PPRC can be used to migrate data from primary or secondary DS8000 storage systems in a PPRC configuration. By using MT-PPRC, you can perform a migration procedure with little or no periods in which the system is not protected by mirroring.

MT-PPRC 3-site with two synchronous targets

The basic mode of operation and configuration of MT-PPRC with two MM relationships is shown in Figure 2-8.

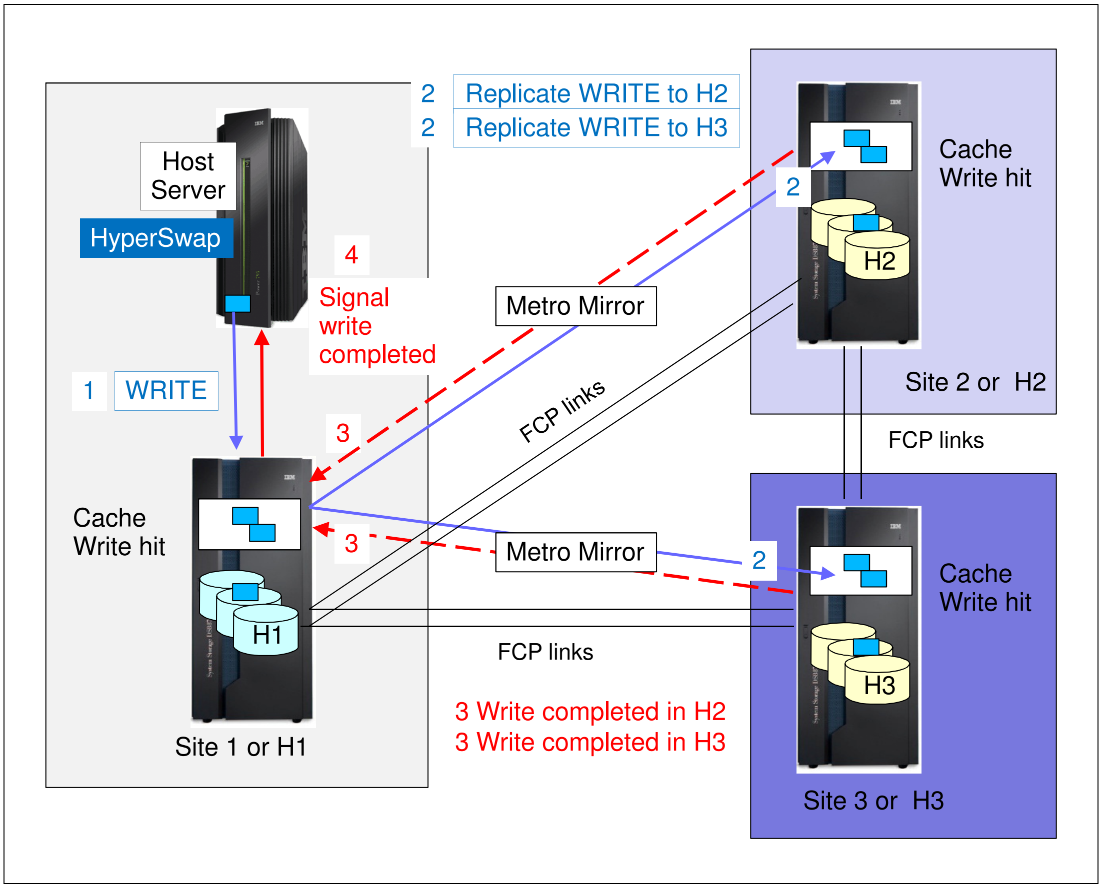


Figure 2-8 MT-PPRC with two synchronous targets

With MT-PPRC and two MM relationships, the DS8000 storage system provides another level of DR and combined with HyperSwap, another level of HA.

The primary storage system schedules two parallel and synchronous replication writes to a target DS8000 storage system in site 2 and another to target DS8000 storage system in site 3. After both replication writes succeed, the application write is considered successfully completed by the host server.

Depending on the available storage area network (SAN) infrastructure, site 2 and site 3 also can be connected to potentially allow synchronous replication from site 2 to site 3 or the opposite configuration if a HyperSwap event occurs in site 1. This configuration is indicated by the HyperSwap action that is shown in Figure 2-8 and requires a Fibre Channel connection (FICON) from the IBM Z server in site 1 to site 2 or site 3.

Managing such a 3-site MT-PPRC configuration is supported by GDPS (GDPS Metro, dual-leg) or by CSM. This support is also proof of how closely IBM Z-based Copy Services software and HyperSwap interact with the connected DS8000 storage systems.

MT-PPRC 3-site configuration with a synchronous and asynchronous target

Another possible MT-PPRC configuration, which implies a Parallel Sysplex configuration across site 1 and site 2, is shown in Figure 2-9. In this configuration, the storage system in site 1 synchronously replicates disk storage volumes over MM to site 2.

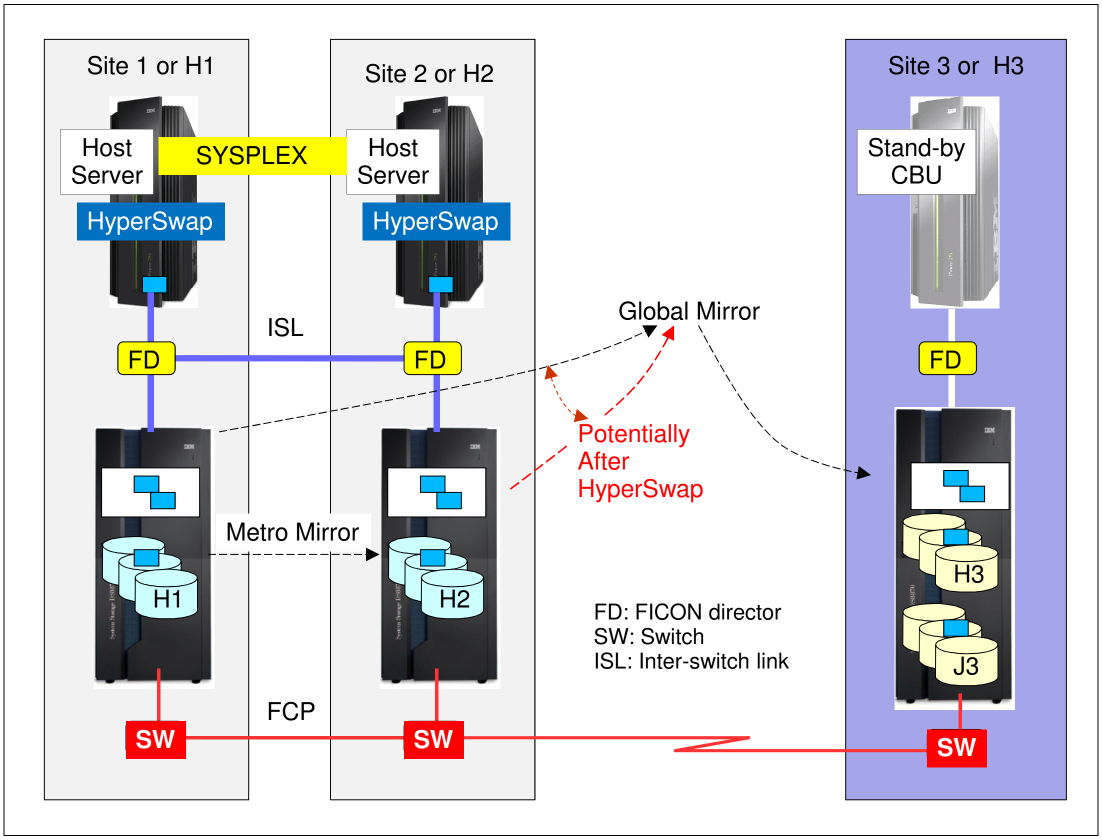


Figure 2-9 MT-PPRC with synchronous and asynchronous target of H1

Although the SAN fabric configuration that is shown in Figure 2-9 also allows a cascaded configuration from site 2 to site 3, the implication here is that GM is the second Copy Services relationship from site 1 to site 3 that is running through a SAN fabric, which might have SAN switches in all three sites. This configuration allows for the highest flexibility.

You must also plan for redundancy in the SAN fabric, which is not shown in Figure 2-9.

Similar considerations apply, as shown in Figure 2-8 on page 16. HyperSwap might transfer the active site from site 1 to site 2 and carry the GM relationship from site 1 to site 2. This configuration can keep the GM relationship active and preserve DR capability in site 3 after a potential HyperSwap event.

When returning the active site to site 1 (if the storage system in site 1 is running again), CSM supports an incremental resynch approach: from site 2 to site 1, and returning to site 1 through a planned HyperSwap operation from site 2 to site 1 when H2 and H1 are back in a FULL-DUPLEX state.

Another subsequent incremental resynch from H1 to H2, and automatically enabling HyperSwap when H1 and H2 are back in a FULL-DUPLEX state, establishes the original configuration, including returning GM back to H1 to H3/J3.

This configuration includes some potential to provide a robust HA/DR configuration.

Again, this configuration combines IBM Z server-based services (such as HyperSwap and hosting Copy Services management software) with the unique DS8000 Copy Services capabilities.

2.1.6 Symmetrical HA/DR 4-site solutions

Many customers conduct regular failover operations to DR sites to comply with regulation requirements, perform data center maintenance at primary locations, or exercise DR capabilities. In such environments, having the ability of still maintaining HA while systems are running on DR sites is a requirement.

The combination of cascading and MT-PPRC on DS8000 machines can provide 4-site solutions. Customers can have sites that span two different metropolitan areas while still maintaining HA/DR capabilities between them and independent of where their systems are running. How this solution can be accomplished is shown in Figure 2-10.

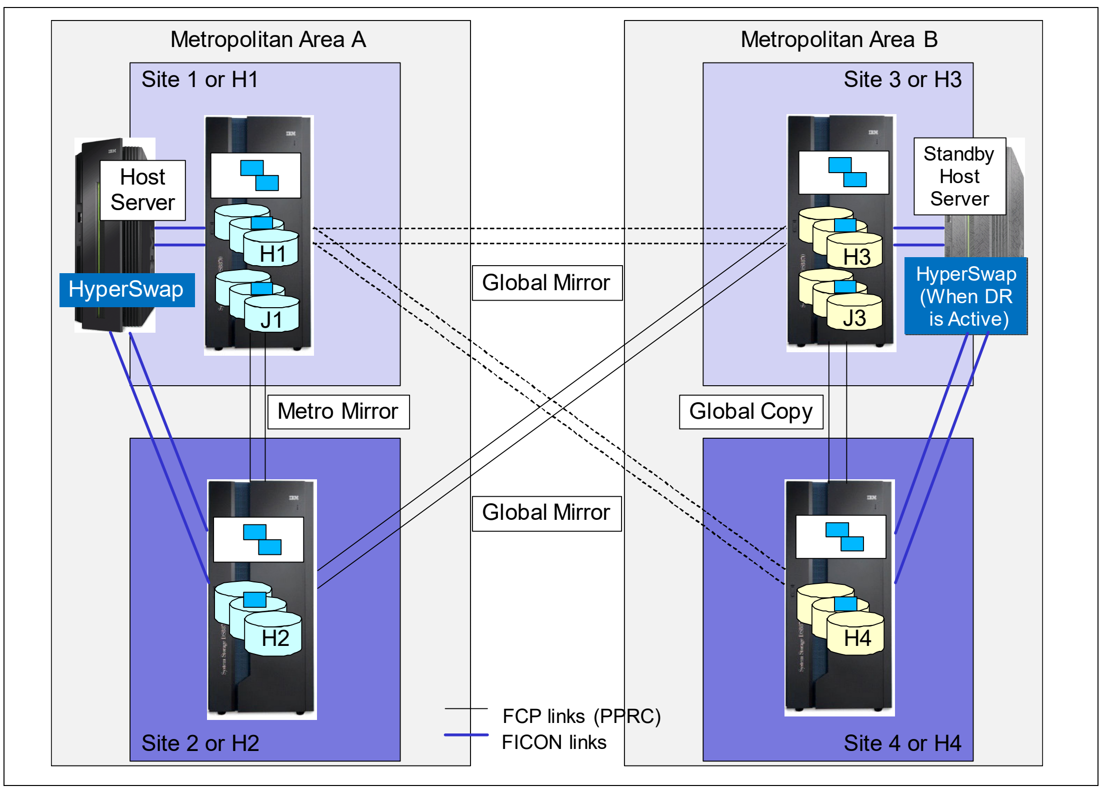


Figure 2-10 Symmetrical HADR 4-site solution

In this example, sites 1 and 2 are running a production workload on “Metropolitan Area A” while Sites 3 and 4 are a DR environment on “Metropolitan Area B”.

H1 includes an active HyperSwap capable MM relationship with H2, and H2 includes an active GM relationship with H3. H1 also features a “stand-by” GM relationship with H3 (denoted with a dotted line in Figure 2-10). In this case, if H2 encounters any issues, the GM relationship can be taken over by H1 without needing full synchronization to H3. H3 also has an active Global Copy relationship with H4, which allows H4 to be close to a “synchronized state” with H3.

If a planned or unplanned DR situation is declared, the systems can be failed over from Metropolitan Area A to Metropolitan Area B. In this case, an H3 to H4 replication relationship can be converted from Global Copy to MM, and HyperSwap can be enabled between them, which provides data HA.

When Metropolitan Area A is ready to be reactivated, H1 becomes the target site for the GM relationship and features a Global Copy relationship that is started from H1 to H2. After the four sites are synchronized, the systems can remain running on Metropolitan Area B with DR protection on Metropolitan Area A, or failed back to Metropolitan Area A, which makes Metropolitan Area B again the DR environment.

This configuration is fully supported by GDPS and CSM. For more information, see *IBM GDPS: An Introduction to Concepts and Capabilities*, SG24-6374.

2.2 z/OS HyperSwap

For many years, z/OS provided the capability to transparently swap the access from one device to another device. The first use of this capability for disk storage volumes occurred with PPRC dynamic address switching (P/DAS). For more information, see the following IBM Documentation web pages:

- Peer-to-Peer Remote Copy dynamic address switching (P/DAS)
- Steps for using P/DAS in a sysplex environment

The z/OS-based swap process that redirects I/Os from device H1 to H2 (if these devices are in an MM relationship and in the FULL DUPLEX state) is shown in Figure 2-11.

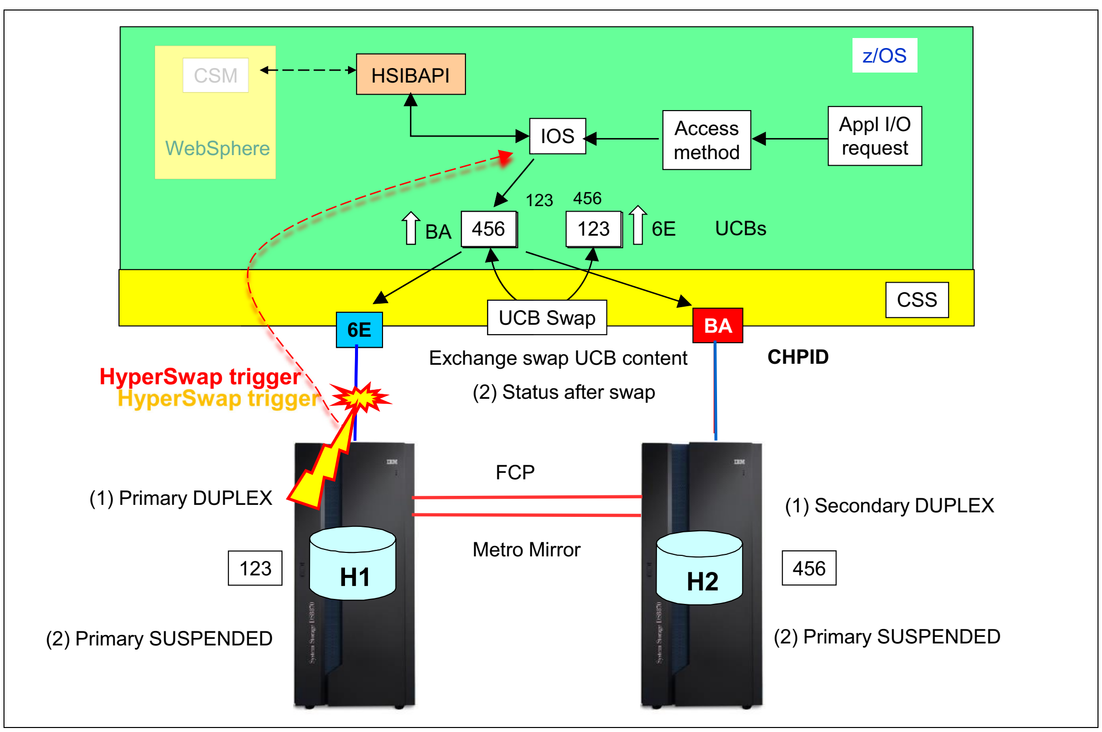


Figure 2-11 z/OS swap process

The core of this process is to exchange (swap) the content of the two unit control blocks (UCBs), which represent the disk storage devices. Among the many details they contain about a device, the UCBs also include one or more channel paths or one or more channel-path identifiers (CHPIDs) that connect to the two devices.

Figure 2-11 on page 19 also shows the status after the UCB swap operation. Before the swap, all I/O to the device on 123 (which is the MM primary device) ran through CHPID 6E.

Eventually, all I/O traffic to 123 is stopped before the swap operation occurs. After all I/O to 123 is quiesced, the swap process exchanges the UCB content of device 123 and device 456. After the swap is completed, IOS resumes I/O operations, and the UCB eventually directs the resumed I/O to CHPID BA, which connects to device 456. An earlier step of the swap process is also to change the MM status of the device on 456 from the SECONDARY DUPLEX to the PRIMARY SUSPENDED state.

IBM enhanced this swap process and raised the number of swap operations that are running in parallel. With today's processor speeds and dedicated highly parallel running swap services within the HyperSwap address space, many thousands of swap operations can occur in a single-digit number of seconds. The key to modernizing this swap process is HyperSwap, which performs in its own address space.

In addition to the actual swap operation that the z/OS HyperSwap service provides, specific DS8000 Copy Services commands can be issued during the swap operation to trigger freeze and failover functions within the DS8000 storage system.

Also, IOS knows that HyperSwap autonomically performs a HyperSwap operation after a trigger is raised that is based on an issue to or within the primary storage server in H1.

Because this HyperSwap service is not an externalized interface, another authorized user must enable this service and work closely with this z/OS based HyperSwap service.

Currently, authorized users of HyperSwap services are CSM and GDPS. Both solutions manage Copy Services configurations and closely interact with the HyperSwap address space to provide a Copy Services configuration to HyperSwap services after the configuration is in a suitable FULL-DUPLEX state.

2.3 Consistent Read from Secondary

In a single-site workload MM configuration, the host processor and primary storage are on the same site, and the auxiliary storage is at some distance from the host processor. All read and write operations are performed locally to the primary storage, and writes are mirrored to the secondary storage.

After a HyperSwap is performed and the mirroring direction is reversed, the secondary is now closer to the processor, and the primary is at distance. With CRS, data is read from the auxiliary storage within an MM relationship as though it was the primary storage, which represents a performance improvement in a z/OS Metro Mirror environment after you shorten the distance between the host and the storage for I/O reads.

To enable consistent reads, z/OS must issue a request to start a consistent read management (CRM) session between the primary and auxiliary storage systems for each participating logical subsystem (LSS).

The CRM session maintains a heartbeat between the primary and auxiliary storage systems. While the secondary receives a heartbeat from the primary within a pre-determined time interval, read operations are allowed to the secondary. If the heartbeat is not received, reads are disallowed from the secondary until the heartbeat is resumed.

CRM becomes active for an LSS when the first system issues a request to start CRM. Subsequent start requests by other systems are ignored.

CRM is stopped only if a PPRC suspension occurs, a specific request is issued to stop CRM, or the HyperSwap configuration is purged.

CRS requirements

This section describes the z/OS requirements to enable CRS:

- The primary and auxiliary storage systems must support CRS. This support is achieved by using DS8900 with microode 9.2 or later.
- The MM relationship is in duplex state, and a HyperSwap configuration is loaded.
- The auxiliary storage system is closer to the processor than the primary. This situation is evaluated on a system-by-system basis.
- z/OS 2.3 or later. APARs OA61131, OA61170, OA61171, and OA61172 are required.
- z/OS has started CRM for the LSS, that is, the heartbeats are transferred between the primary and auxiliary storage.
- The CRS function is enabled in the software.
- A new **READSEC** keyword on **IECIOSxx** parmlib or **SETIOS** command:

  ```
  ZHPFOPTS,MAXSIZE={nnnn|SYSTEM},READSEC={YES|NO}
  ```

  – **YES**: Enables CRS for the current system.
  – **NO**: Disables CRS for the current system.
- zHPF is enabled for the primary and secondary devices.

2.4 Copy Services Manager and HyperSwap

IBM CSM is required to use HyperSwap within z/OS for an MM configuration. This section does not describe CSM beyond the fact that it manages sessions. Such a session contains all MM volume pairs that are set up and defined within a Parallel Sysplex configuration. From a user perspective, the entity of management is the session only.

CSM is server-based and includes two interfaces: a GUI and a command-line interface (CLI). The CSM server is preinstalled on the DS8900F Hardware Management Console (HMC). It also can run on all common server platforms, such as IBM AIX®, Linux, Microsoft Windows, and on z/OS within the UNIX System Services or UNIX System Services shell.

CSM can handle all z/OS-based CKD volumes within its MM session, even when the CSM server is hosted on the HMC or a distributed server. However, a best practice is to use the robust IBM Z server platform and place the CSM server in a z/OS logical partition (LPAR). Also, when possible, you can host the CSM stand-by server in another z/OS LPAR at the other site.

Figure 2-11 on page 19 shows that CSM appears unavailable because CSM is not necessary when z/OS performs an IBM HyperSwap operation. This fact is also true when HyperSwap is enabled within a Parallel Sysplex configuration. Therefore, CSM is also unavailable, as shown in Figure 2-12.

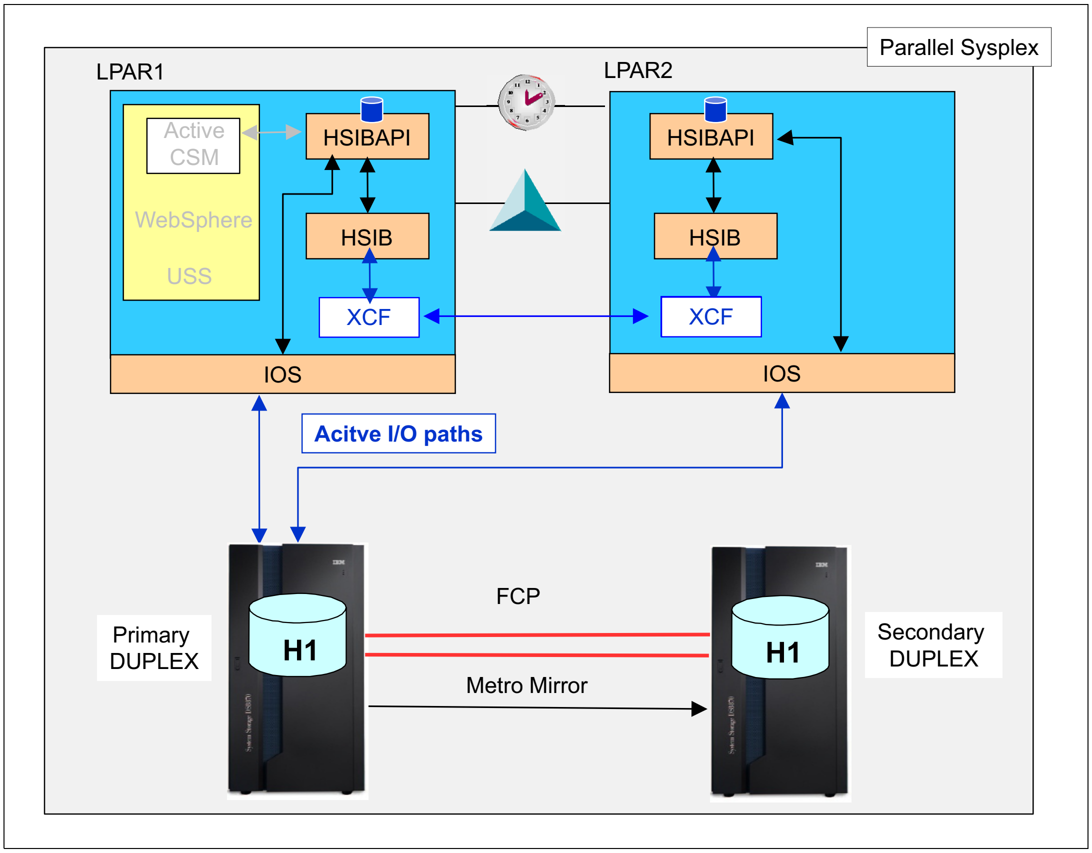


Figure 2-12 HyperSwap enabled: CSM is passive

As shown in Figure 2-12, HyperSwap is represented in an LPAR by the following address spaces:

- The HyperSwap application programming interface (HSIBAPI) address space that handles the swap process.
- The HyperSwap Management (HSIB) address space that is the communication handler for its peers in other Parallel Sysplex members.

Figure 2-12 also shows normal operations with the active I/O paths that are connected to H1 volumes, which are connected through MM to H2 and all in a correctly working FULL-DUPLEX state. This requirement must be met to reach the HyperSwap enabled state.

You can also query the HyperSwap status by using the z/OS **display** command, as shown in Example 2-1.

Example 2-1 Querying the HyperSwap status by using a z/OS system command

```
D HS,STATUS

 IOSHM0303I HyperSwap Status 671
 Replication Session: MGM
 HyperSwap enabled
 New member configuration load failed: Disable
 Planned swap recovery: Disable
 Unplanned swap recovery: Partition
 FreezeAll: Yes
 Stop: No
```

Example 2-2 shows another z/OS system command that can be used to control HyperSwap and disable or enable HyperSwap when a HyperSwap session exists. Because only one potential HyperSwap session is allowed within a Parallel Sysplex, referring to a specific name in the **SETHS** z/OS system command is not necessary.

Example 2-2 z/OS system commands to disable or enable HyperSwap

```
RO *ALL,SETHS DISABLE

RO *ALL,SETHS ENABLE
```

However, it might be necessary to disable HyperSwap in a planned fashion to avoid a HyperSwap operation during a controlled and planned activity that might trigger a HyperSwap operation. After such a controlled activity, the HyperSwap can be reenabled so that z/OS HyperSwap regains control.

Example 2-3 shows another z/OS system command that can be used to query a complete HyperSwap configuration. However, that command might not be helpful when thousands of MM volume pairs are within the HyperSwap session.

Example 2-3 Querying the complete HyperSwap configuration

```
D HS,CONFIG(DETAIL,ALL)
  IOSHM0304I HyperSwap Configuration 495
  Replication Session: MGM
  Prim. SSID  UA   DEV#   VOLSER   Sec. SSID  UA   DEV#  Status
         A0   31  0A031   A#A031         40   31  04031
         A0   7F  0A07F   A#A07F         40   7F  0407F
         A0   AF  0A0AF   A#A0AF         40   AF  040AF
         A0   72  0A072   A#A072         40   72  04072
         A0   A2  0A0A2   A#A0A2         40   A2  040A2
         A0   13  0A013   A#A013         40   13  04013
         .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
```

To see why a HyperSwap session is unavailable, a basic approach is to start with another **D HS** z/OS system command, as shown in Example 2-4.

Example 2-4 Querying HyperSwap for exceptions

```
D HS,CONFIG(EXCEPTION,ALL)
  IOSHM0304I HyperSwap Configuration 840
  Replication Session: MGM
  None Duplex
```

2.4.1 HyperSwap to site H2

A healthy and enabled HyperSwap session is shown in Figure 2-12 on page 22. However, for planned or unplanned reasons, a HyperSwap trigger can change the configuration to what is shown in Figure 2-13.

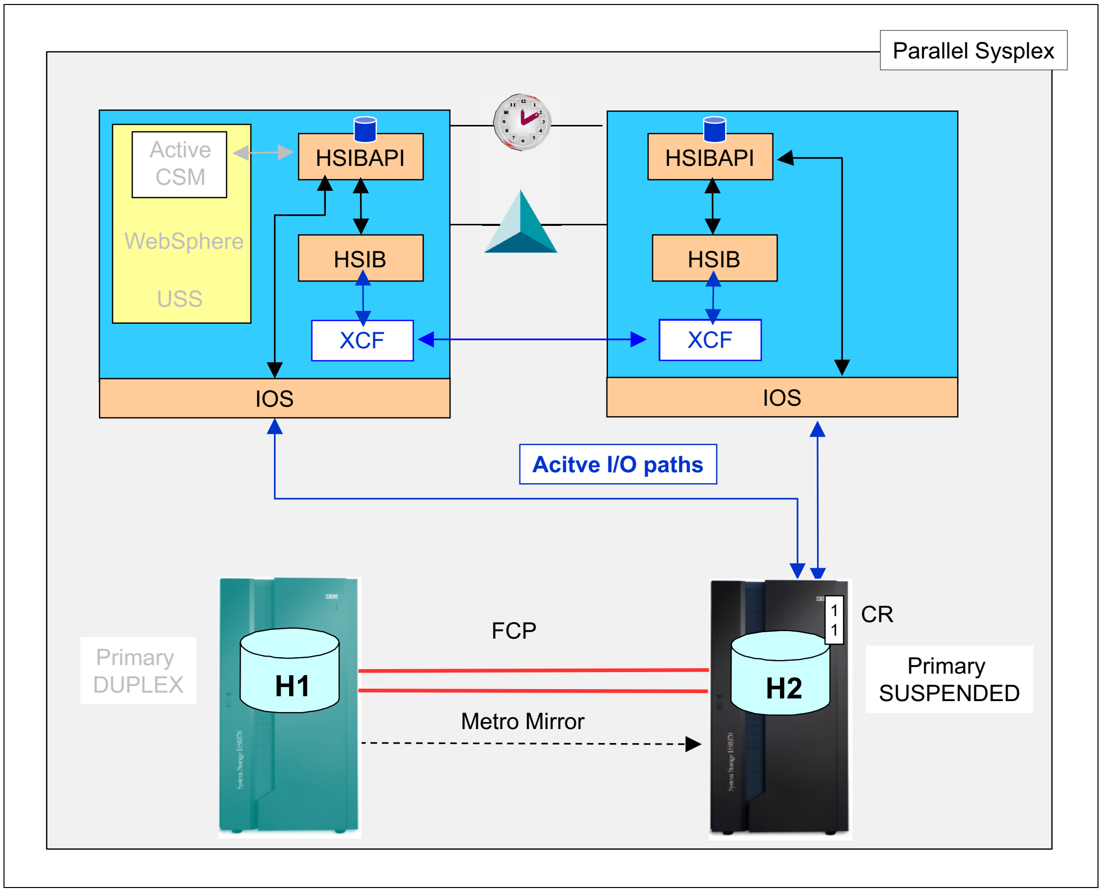


Figure 2-13 After HyperSwap: CSM is still passive

After a planned or unplanned HyperSwap switched the active volumes from H1 to H2, CSM remains passive and not involved. HyperSwap eventually notifies CSM about the session state change after the HyperSwap operation is completed.

During the actual swap operation, HyperSwap is also responsible for issuing all the necessary Copy Services commands to perform the complete failover to H2. This failover also leads to the MM state change of the H2 volumes from secondary DUPLEX to primary SUSPENDED.

2.4.2 Returning to site H1

After the decision is made to return the active site to H1 and if the DS8000 storage system in H1 is recovered and still holds all the data at a level when the HyperSwap occurred, CSM is required to perform the necessary steps.

The first steps to return the active site to H1 are shown in Figure 2-14.

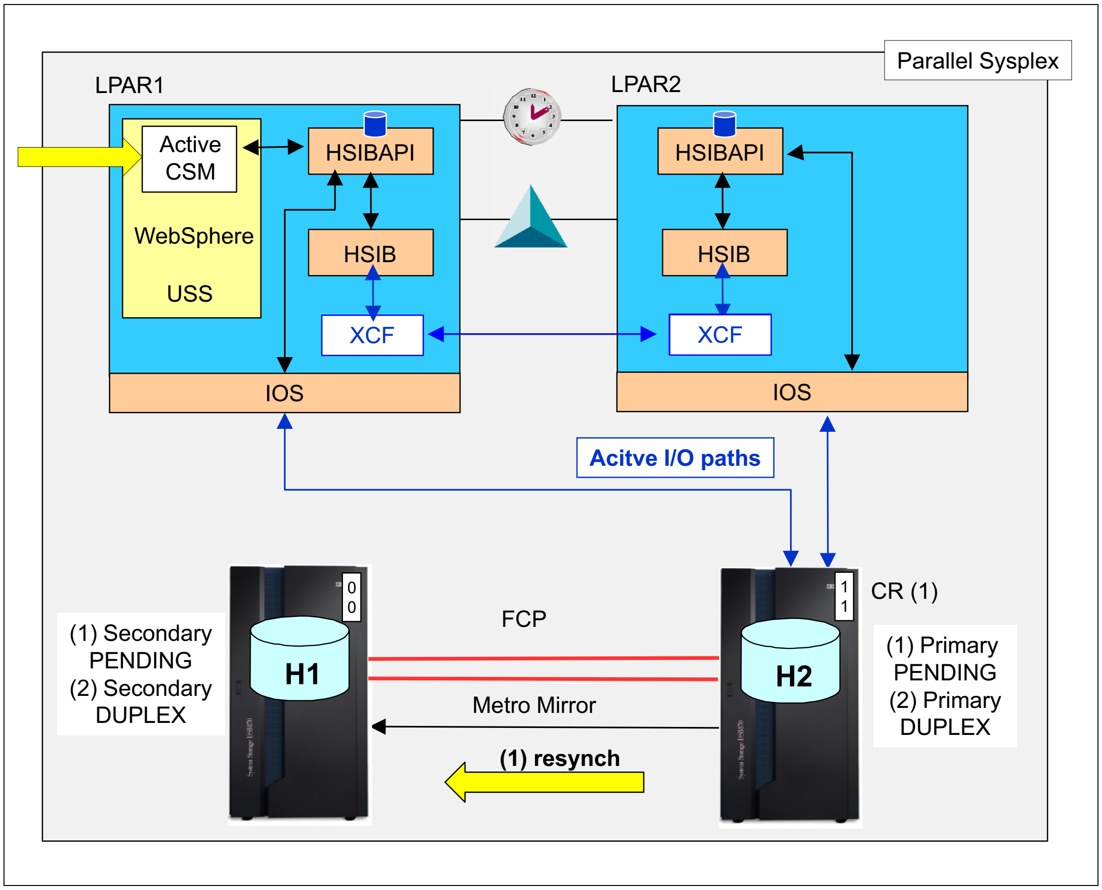


Figure 2-14 Reestablishing Metro Mirror and returning to H1: CSM required

When H2 was the active site, all relevant updates to the H2 configuration were logged within the DS8000 storage system and the corresponding bitmap.

You must return to the corresponding CSM session. There, you discover that the session changed and is no longer in a green OK status.

To start the process and to return the active volumes to H1, complete the following steps:

1. Modify the CSM session to allow the volumes to be replicated in the opposite direction that it was using. This action enables the session to replicate from H2 to H1.
2. Start the replication to resynchronize the volumes incrementally from H2 to H1 by using a CSM **START_H2_H1** command.

After the incremental resynchronization process is completed for all volumes within the session and everything is back to DUPLEX, CSM returns the configuration to HyperSwap, which then switches back to enable the session for HyperSwap ready.

To return the active volumes back to H1, complete the following steps:

1. Issue a planned HyperSwap through the CSM or by using the **SETHS SWAP** z/OS system command. Again, this command performs a swap operation and puts the active volumes back to H1, including the MM status of primary SUSPENDED.
2. CSM performs the following actions as before:
   a. Allows the session to replicate from H1 to H2.
   b. Resynchronizes all the volume pairs from H1 to H2 through another **START_H1_H2** command.

After all MM pairs are in the FULL-DUPLEX state, CSM again signals the new configuration to HyperSwap, enabling HyperSwap ready, and the replication continues now from H1 to H2.

2.4.3 Summary

CSM follows the high IBM standards that also apply for the enhancements that are made to IBM Z and the DS8000 storage system.

CSM is the enabler for z/OS and its HyperSwap function. This synergy allows a 2-, 3-, or 4-site MM-based disk volume configuration to achieve high standards in data availability and DR readiness in a fully transparent fashion to the application I/Os.

2.5 Geographically Dispersed Parallel Sysplex

GDPS is a solution that manages complex multisite HA-DR IBM Z environments. GDPS simplifies DS8000 storage system replication and Parallel Sysplex management while providing end-to-end application business resilience.

To address an entire site failure, GDPS can perform a site switch to another local site or to a remote (out-of-region) location that is based on predefined, automated scripts. Various GDPS offerings are available (see Figure 2-15), and each one addresses specific HADR goals that can be customized to meet various recovery point objective (RPO) and recovery time objective (RTO) requirements.

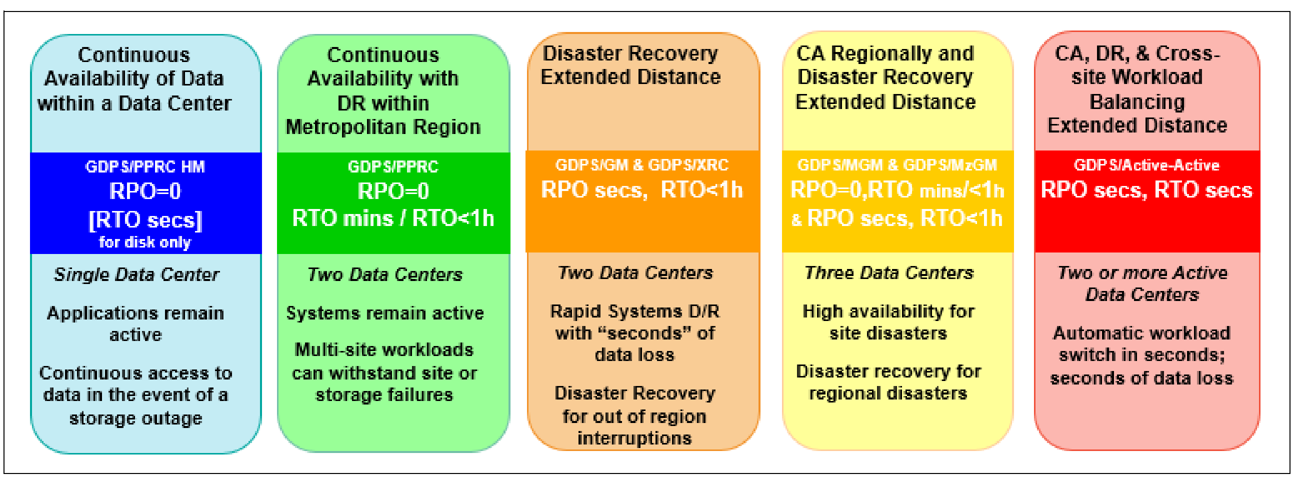


Figure 2-15 GDPS offerings

One difference between options is the type of DS8000 Copy Services that is used as a building block for HADR design. The following Copy Services are used:

- GDPS Metro HyperSwap Manager (GDPS HM) and GDPS Metro (the former GDPS/PPRC): Based on DS8000 synchronous data replication MM (known as PPRC).
- GDPS Global - GM (GDPS GM): Based on the DS8000 GM, which is an asynchronous form of remote copy.
- GDPS Global - Extended Remote Copy (GDPS XRC): Uses asynchronous data replication XRC (also known as IBM zGM).
- GDPS Metro Global - GM (GDPS MGM): Uses MM and GM disk replication for a 3-site or 4-site HADR environment.
- GDPS Metro (dual-leg): Supports Multi-Target Metro Mirror (MTMM) on DS8000 storage systems. GDPS Metro dual-leg provides similar capabilities as the available capabilities in GDPS Metro single-leg while extending PPRC management and HyperSwap capabilities to cover the two replication legs.
- GDPS Metro Global - XRC (GDPS MzGM): Uses MM and XRC or IBM zGM disk replication for a 3-site or 4-site HADR environment.
- GDPS Continuous Availability (GDPS AA): A multisite HADR solution at almost unlimited distances. This solution is based on software-based asynchronous mirroring between two active production sysplexes that are running the same applications with the ability to process workloads in either site.

For more information about GDPS and each option, see *IBM GDPS: An Introduction to Concepts and Capabilities*, SG24-6374.

2.5.1 GDPS and DS8000 synergy features

Almost all GDPS solutions (except for GDPS AA) rely on IBM disk replication technologies that are used in the DS8000 storage family. This section provides more information about the key DS8000 technologies that GDPS supports and uses.

Metro Mirror (PPRC) failover/failback support

When a primary disk failure occurs and the disks are switched to the secondary devices, failover/failback support eliminates the need to perform a full copy when reestablishing replication in the opposite direction. Because the primary and secondary volumes are often in the same state when the freeze occurred, the only differences between the volumes are the updates that occur to the secondary devices after the switch.

Failover processing sets the secondary devices to primary suspended status and starts change-recording for any subsequent changes that are made. When the mirror is reestablished with failback processing, the original primary devices become secondary devices, and changed tracks are resynchronized.

GDPS Metro transparently uses the failover/failback capability. This support mitigates RTO exposures by reducing the amount of time that is needed to resynchronize mirroring after a HyperSwap. The resynchronization time depends on how long the mirroring was suspended and the number of changed tracks that must be transferred.

If an entire central processor complex (CPC) fails, GDPS can run the **`CECFAIL_cpcname`** recovery script, which can contain any valid statements, such as **`CAPACITY`** statements. This script can be used to schedule the capacity changes that needed on the backup CPCs.

GDPS also supports MTMM on the IBM DS8900F, DS8880, and DS8870 storage systems. Initial support is for two synchronous copies from a single primary volume, also known as an *MTMM configuration*. GDPS Metro (dual-leg) provides similar capabilities to the ones that are available in single-leg GDPS Metro while extending PPRC management and HyperSwap capabilities to cover the two replication legs.

Global Copy

Global Copy (initially known as PPRC-XD) is an asynchronous form of the DS8000 advanced copy functions. GDPS uses Global Copy rather than synchronous MM (PPRC) to reduce the performance effect of certain remote copy operations that potentially involve a large amount of data. The replication links are typically sized for steady-state update activity, but not for bulk synchronous replication, such as initial volume copy or resynchronization.

Initial copy or resynchronizations by using synchronous copy do not need to be performed because the secondary disks cannot be made consistent until all disks in the configuration reach the duplex state. Therefore, GDPS supports initial copy and resynchronization by using asynchronous Global Copy.

When GDPS starts copy operations in asynchronous copy mode, GDPS monitors the progress of the copy operation. When the volumes are near full duplex state, GDPS converts the replication from the asynchronous copy mode to synchronous. Initial copy or resynchronization by using Global Copy eliminates the performance effect of synchronous mirroring on production workloads.

The use of asynchronous copy allows clients to establish or resynchronize mirroring during periods of high production workload. It also might reduce the time during which the configuration is exposed.

DS8000 Health Message Alert

An unplanned HyperSwap is started automatically by GDPS if a primary disk failure occurs.

In addition to a disk problem being detected as a result of an I/O operation, a primary disk subsystem can proactively report that it is experiencing an acute problem. The DS8000 storage system features a special microcode function that is known as the *Storage Controller Health Message Alert* capability. It alerts z/OS when hardware events occur and generates a message and Event Notification Facility (ENF) signal, as shown in Example 2-5.

Example 2-5 DS8000 Health Message Alert

```
IEA074I STORAGE CONTROLLER HEALTH,MC=20,TOKEN=1004,SSID=AB01, DEVICE
NED=2107.961.IBM.75.0000000ABCD1.0100,PPRC SECONDARY CONTROLLER RECOVERY ACTION
```

The DS8000 storage system reports problems of different severities. Those problems that are classified as *acute* are also treated as HyperSwap triggers. After systems are swapped to use the secondary disks, the disk subsystem and operating system can attempt to perform recovery actions on the former primary without affecting the applications that use those disks.

One main benefit of the Health Message Alert function is to reduce false freeze events. GDPS Freeze and Conditional Stop actions query the secondary disk subsystem to determine whether systems can be allowed to continue in a freeze event.

Metro Mirror (PPRC) suspension

An MM suspension generates a message aggregation that is also known as *Summary Event Notification*. This aggregation dramatically reduces host interrupts and operator messages when an MM volume pair is suspended.

When GDPS performs a freeze, all primary devices in the MM configuration suspend. This suspension can result in significant *state change interrupt* (SCI) traffic and many messages in all systems. GDPS supports reporting suspensions in a summary message per DS8000 logical control unit (LCU) instead of at the individual device level.

When compared to reporting suspensions on a per devices basis, the Summary Event Notification dramatically reduces the message traffic and extraneous processing that is associated with MM suspension events and freeze processing. Examples exist where 10,000 operator messages were reduced to under 200.

Soft Fence feature

After a GDPS HyperSwap or an unplanned site switch, potential exposures exist to systems that are connected to the original primary MM (PPRC) volumes. Figure 2-16 shows that after a planned or unplanned HyperSwap, the GDPS changes the secondary volumes to primary suspended, but the former primary volumes' statuses remain unchanged. Therefore, these devices remain accessible and usable to any system within or outside the sysplex. In this case, possibilities exist to update or perform an IPL accidentally from the wrong set of disks, which can result in potential data integrity or a data loss problem.

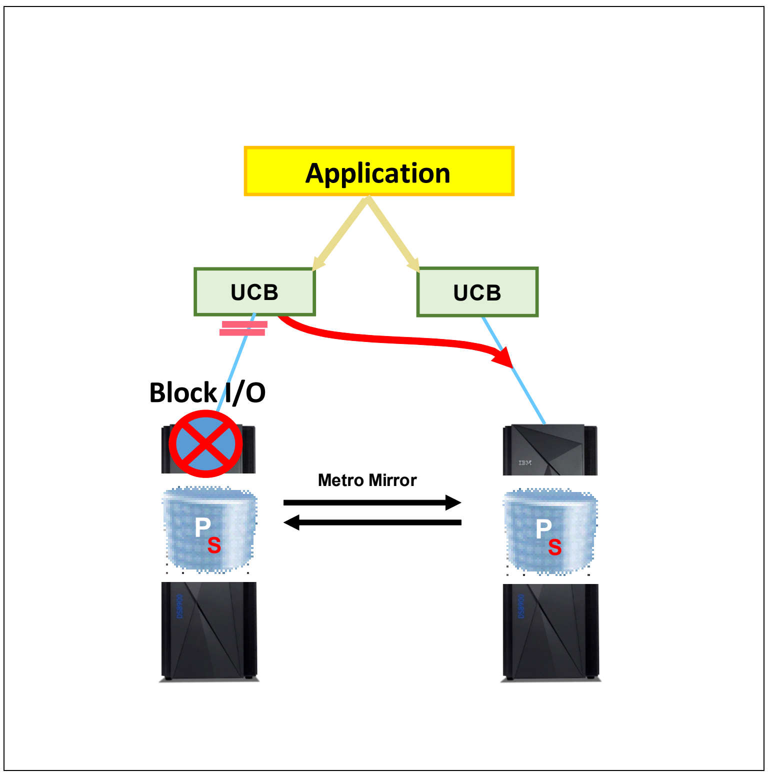


Figure 2-16 GDPS and DS8000 Soft Fence feature

GDPS uses a DS8000 capability that is called *Soft Fence* to fence (block access to a selected device). GDPS uses Soft Fence when suitable to fence devices that otherwise might be exposed to accidental update; for example, after a GDPS HyperSwap event, as shown in Figure 2-16.

Although GDPS includes built-in protection features that prevent an IPL of the systems from the incorrect set of disks, the DS8000 Soft Fence function is more protection. If an IPL of any system is done manually (without GDPS), the attempt of an IPL from the wrong set of disks (fenced former primary MM volumes) is prohibited.

Also, other systems that are outside the sysplex (and therefore outside GDPS control) can access the former primary MM volumes. Soft Fence protection blocks any attempt to update these volumes.

On-Demand Dump

When problems occur with disk systems, such as problems that result in an unplanned HyperSwap, a mirroring suspension, performance issues, or a lack of diagnostic data from the time that the event occurs, can result in difficulties in identifying the root cause of the problem. Taking a full *statesave* can lead to temporary disruption to host I/O.

The On-Demand Dump (ODD) capability of the DS8000 storage system facilitates taking a nondisruptive statesave (NDSS) at the time such an event occurs. The DS8000 microcode performs this statesave automatically for specific events, such as generating a memory dump of the primary disk system that triggers a freeze event and allows an NDSS to be requested by a user. This feature enables first-failure data capture (FFDC) and ensures that diagnostic data can help with problem determination efforts.

GDPS supports taking an NDSS that uses the remote copy pages (or web GUI). In addition to this support, GDPS autonomically takes an NDSS if an unplanned freeze or HyperSwap event occurs.

Query Host Access

When an MM (PPRC) disk pair is being established, the target (secondary) device must not be used by any system. The same is true when establishing a FlashCopy relationship with a target device. If the target is in use, the establishment of the PPRC or FlashCopy relationship fails.

When such failures occur, identifying which system is delaying the operation can be a tedious task. The DS8000 Query Host Access function provides the means to query and identify which system uses a selected device. This function is used by the IBM Device Support Facilities (ICKDSF) utility (for more information, see 4.5, “Volume formatting overwrite protection” on page 55) and by GDPS.

GDPS features the following capabilities:

- Query Host Access identifies the LPAR by using the selected device through the CPU serial number and LPAR number. For the operations staff to convert this information to a system or CPU and LPAR name is still a tedious job. GDPS performs this conversion and presents the operator with more readily usable information, which avoids this extra conversion effort.
- When GDPS is requested to perform a PPRC or FlashCopy establish operation, GDPS first performs Query Host Access to determine whether the operation is expected to succeed or fail as a result of one or more target devices being in use. GDPS alerts the operator if the operation is expected to fail and identifies the target devices in use and the LPARs holding them.
- GDPS continually monitors the target devices that are defined in the GDPS configuration and alerts operations that target devices are in use when they should not be in use. This alert allows operations to fix the reported problems in a timely manner.
- With GDPS, the operator can perform *ad hoc* Query Host Access to any selected device by using the GDPS pages (or GUI).

The GDPS QHA monitor was optimized with GDPS V4.4 to significantly reduce the time that is taken to query each CPC that is defined in the System Automation policy.

IBM DS8000 Easy Tier Heat Map Transfer

IBM DS8000 Easy Tier Heat Map Transfer (HMT) can transfer the Easy Tier learning from an MM (PPRC) primary to the secondary disk system. The secondary disk system can also be optimized based on this learning and have similar performance characteristics in the HyperSwap event. For more information, see 5.6, “Easy Tier” on page 92.

GDPS integrates support for HMT. The suitable HMT actions (such as the starting and stopping of processing and reversing transfer direction) are incorporated into the GDPS managed processes. For example, if MM is temporarily suspended by GDPS for a planned or unplanned secondary disk outage, HMT is also suspended. If the MM direction is reversed as a result of a HyperSwap, the HMT direction is also reversed.

GDPS HMT support was integrated with the running of GDPS MGM procedures to ensure that the transfer direction of Easy Tier learning information reflects the replication direction at the successful completion of the procedure. With GDPS V3.12 and later, HMT is supported for all available GDPS options (2-, 3-, and 4-site environments).

Logical Corruption Protection

GDPS Logical Corruption Protection (LCP) is a set of GDPS capabilities that is provided in response to the growing number of requests for a GDPS managed Continuous Data Protection capability. It is aimed at helping customers to recover from logical corruption events, whether caused by internal attacks or cyberattacks.

LCP can capture multiple, secure point-in-time copies of critical production data and restore the data to production. LCP also can recover a specific point-in-time copy to another set of devices that can be used to start one or more isolated recovery systems to analyze the scope of a specific logical corruption event. The protection copies can be captured by using the Safeguarded Copy technology or the FlashCopy technology.

Support is introduced for multiple recovery copy sets, which provides greater flexibility when GDPS LCP Manager is used. Starting with GDPS V4.1, the LCP feature of GDPS Metro enables the capture of multiple and secure point-in-time copies that can later be used for identification, repair, or replacement of production data that was compromised by cyberattacks or internal attacks, or corrupted by system failures or human error. Up to 10 recovery copy sets can be defined for use. The total of FlashCopy sets and recovery copy sets cannot exceed 11.

With GDPS V4.3, the LCP Manager is extended to support the DS8000 Safeguarded Copy function. Profile characteristics can be defined, such as the retention period before the capture is classed as expired or the minimum interval between captures, which prevents flooding the storage with bad captures after a corruption event.

GDPS V4.4 instructs DS8000 to create and maintain a persistent Safeguarded Copy recovery relationship so that the user can query the persistent Safeguarded Copy recovery relationship and percentage that is copied through the GDPS LCP Manager interface.

For more information about Safeguarded Copy, see *IBM DS8000 Safeguarded Copy (Updated for DS8000 R9.2.1)*, REDP-5506.

For more information how IBM DS8000® Storage with Safeguarded Copy is integrated in the IBM Z Cyber Vault solution, see *Getting Started with IBM Z Cyber Vault*, SG24-8511.

2.5.2 GDPS and DS8000 synergy summary

GDPS is designed for complex multi-site or single-site IBM Z environments. It can manage disk remote copy, automate Parallel Sysplex operation tasks, and perform failure recovery from a single point of control easily and efficiently. Continuous collaboration over many years between IBM Z, GDPS, and DS8000 development teams delivered a robust HA-DR design that is commonly used among IBM Z clients.

With its HyperSwap capability, GDPS is the ideal solution when targeting 99.99999% (seven nines) availability.

Moreover, it also allows clients to run DR tests more frequently without affecting production. The more that you practice your DR process, the more confident you become in recovering your systems and applications if a real disaster strikes.

3

Data protection and backup

This chapter describes FRBACKUP and other backup solutions that use FlashCopy with Db2 and IMS, adding to the list of synergy items between IBM Z server and DS8000 storage systems.

This chapter includes the following topics:

3.1 Introduction

Traditional data protection and backup typically rely on a point-in-time copy, which allows restoring data only from when that copy or backup was created. As the name implies, *continuous data protection,* in contrast to traditional data protection and backup, has no defined schedule. However, it is often an asynchronously created copy, which does not necessarily ensure consistent data at any one time. Utilities and other software or middleware-based functions are available to overcome this shortcoming of continuous data backup.

Massive data backups still rely on a point-in-time backup copy that is combined with serialization efforts. For example, database subsystems can temporarily quiesce databases and related data sets, which forces a buffer flush to permanent storage.

So, the challenge is to create a backup copy as quickly as possible to make the database data sets available again. Several solutions were developed over time, but the basic and simplest approach is still to create a consistent backup as quickly as possible that also is instantly reusable.

With the DS8000 storage system, the foundation for this simple backup approach is based on the FlashCopy function. The next step was to develop software that intelligently uses FlashCopy to have a fast process to flash or snap a volume or data set.

IBM Fast Replication Backup (FRBACKUP) enables storage administrators to create backup policies for IBM Db2 and IBM Information Management System (IMS) databases with minimum effect on applications. Db2 uses this function for Db2 system-level backup and to control the serialization effort that ensures a consistent set of the complete system-level backup, including all Db2 table space and logs. Db2 also maintains its repository, for example, by updating the bootstrap data set (BSDS) to reflect all backup activity.

3.2 FRBACKUP

FRBACKUP is a storage-based solution that uses storage software components of the z/OS Data Facility Storage Management Subsystem (DFSMS) with its components:

- Data Facility Storage Management Subsystem Data Facility Product (DFSMSdfp)
- Data Facility Storage Management Subsystem Hierarchical Storage Manager (DFSMShsm)
- Data Facility Storage Management Subsystem Data Set Services (DFSMSdss)

These components can interact and use DS8000 Copy Services functions, and here in particular, DS8000 FlashCopy and its various characteristics.

All of the involved storage components from a hardware perspective and z/OS DFSMS software perspective are shown in Figure 3-1.

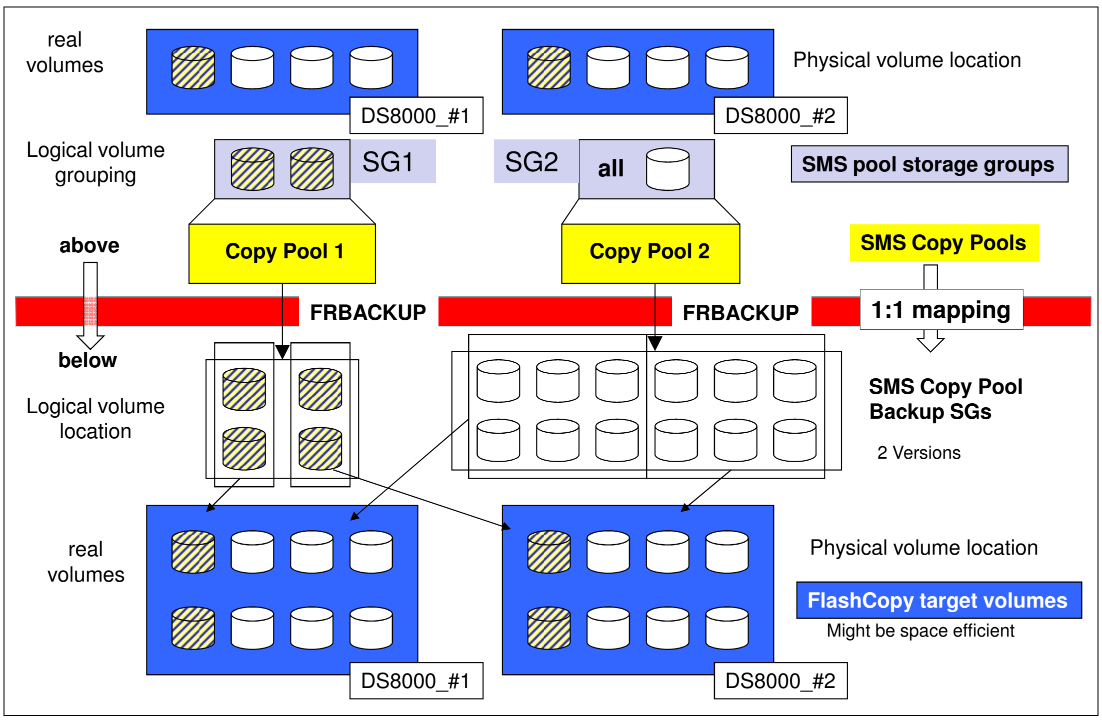


Figure 3-1 FRBACKUP: Storage associations

The configuration shows all important components from a storage perspective. Although the configuration might appear to be a complex solution, it is simple rather than complex. It is logically divided into two parts by a horizontal bar (showing FRBACKUP in Figure 3-1).

For more information, see *DFSMShsm Fast Replication Technical Guide*, SG24-7069.

3.2.1 Storage Management Subsystem construct of the copy pool to support FRBACKUP as input

Different types of Storage Groups (SGs) are used to support FRBACKUP I/O. On the input site to FRBACKUP is a copy pool. Another SG type, the Copy Pool Backup Storage Group (CPBSG), supports FRBACKUP on the output site of FRBACKUP.

The Storage Management Subsystem (SMS) construct of the copy pool is shown in Figure 3-2. The conventional z/OS volumes are grouped in conventional SMS pool SGs. Two DS8000 storage systems and SMS with two SGs that contain all the relevant logical volumes are used.

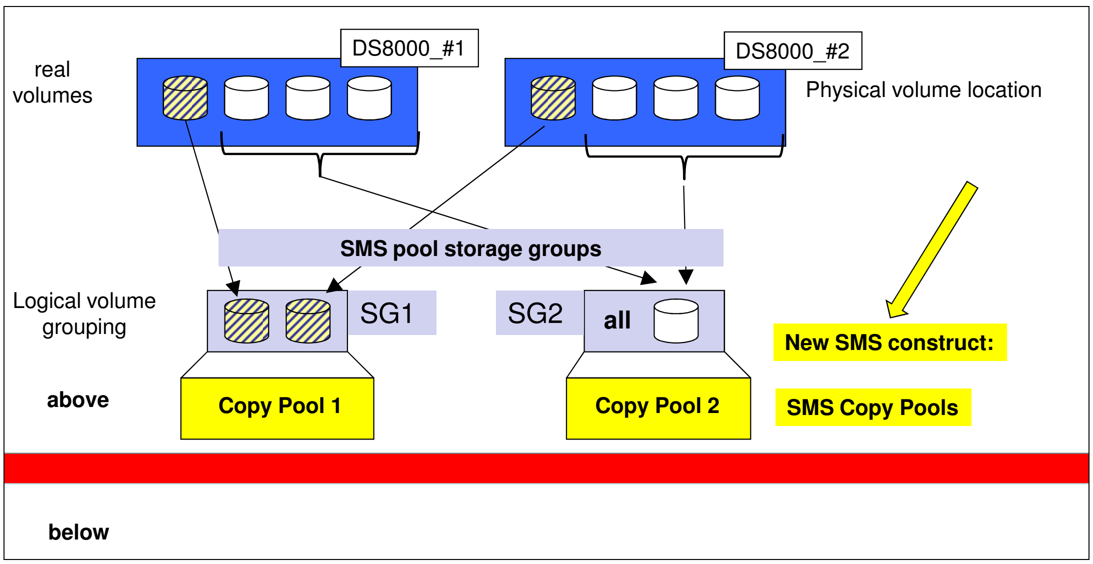


Figure 3-2 New SMS construct: SMS copy pool

The example that is shown in Figure 3-2 might represent a Db2 environment with log volumes and database volumes. The log volumes are grouped in SG SG1. All database volumes are grouped in SG SG2.

Tip: Simplify your SMS configuration whenever possible. Typically, reduce the number of SGs to approximately 4 - 6 SGs, the Storage Class (SC) number to no more than 10, and the Management Class (MC) number to less than 20.

Copy Pool 1 incorporates only a single SMS SG (SG SG1). A copy pool can contain as many as 256 SMS pool SGs, and all of these SGs are processed collectively during FRBACKUP processing. In this simple configuration, Copy Pool 2 is defined with another single SG2 SG. Copy pools serve as input to the FRBACKUP process.

For more information about defining copy pools, see *z/OS DFSMSdfp Storage Administration*, SC23-6860.

3.2.2 Copy Pool Backup Storage Group for output to FRBACKUP

A second construct is another SMS SG type. The CPBSG is a logical bracket around candidate target FlashCopy volumes. All volumes within a CPBSG are reserved for DFSMShsm FRBACKUP requests and for DFSMShsm use only.

CPBSG is associated with copy pools only. When preparing for an FRBACKUP process, DFSMShsm interfaces with SMS to assign to each FlashCopy source volume from the copy pool a corresponding FlashCopy target volume in the CPBSG. Therefore, you must have enough eligible FlashCopy target volumes that are available to map all volumes from a copy pool to the CPBSG.

As shown in Figure 3-3, the DS8000 storage systems need enough disk space for FlashCopy target volumes to accommodate all FlashCopy source volumes from the copy pool that originates from the same DS8000 storage system. All volumes in Copy Pool 1 and in Copy Pool 2 from DS8000_#1 need at least another FlashCopy target volume in the same DS8000_#1 within a corresponding CPBSG when only one copy is required. FRBACKUP supports up to 85 copies.

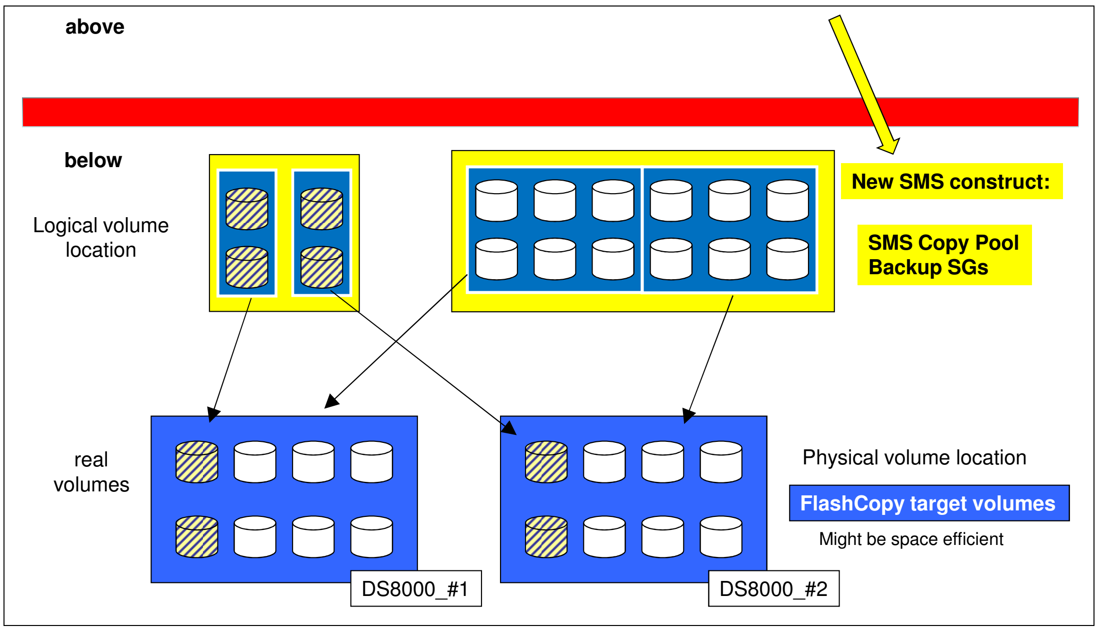


Figure 3-3 New SMS construct: CPBSG

When you define a copy pool, consider keeping backup versions on tape. DFSMShsm automatic memory dump processing can create these tape copies.

For more information, see the following publications:

- Copy pool Backup Storage Group: *z/OS DFSMSdfp Storage Administration,* SC23-6860
- Mirrored FlashCopy targets (also known as Preserve Metro Mirror [Preserve MM]): *z/OS DFSMS Advanced Copy Services*, SC23-6847

3.2.3 Combining copy pool and Copy Pool Backup Storage Group

Combining a copy pool and CPBSG is required. A 1:1 mapping of two copy pools into two corresponding CPBSGs is shown in Figure 3-4.

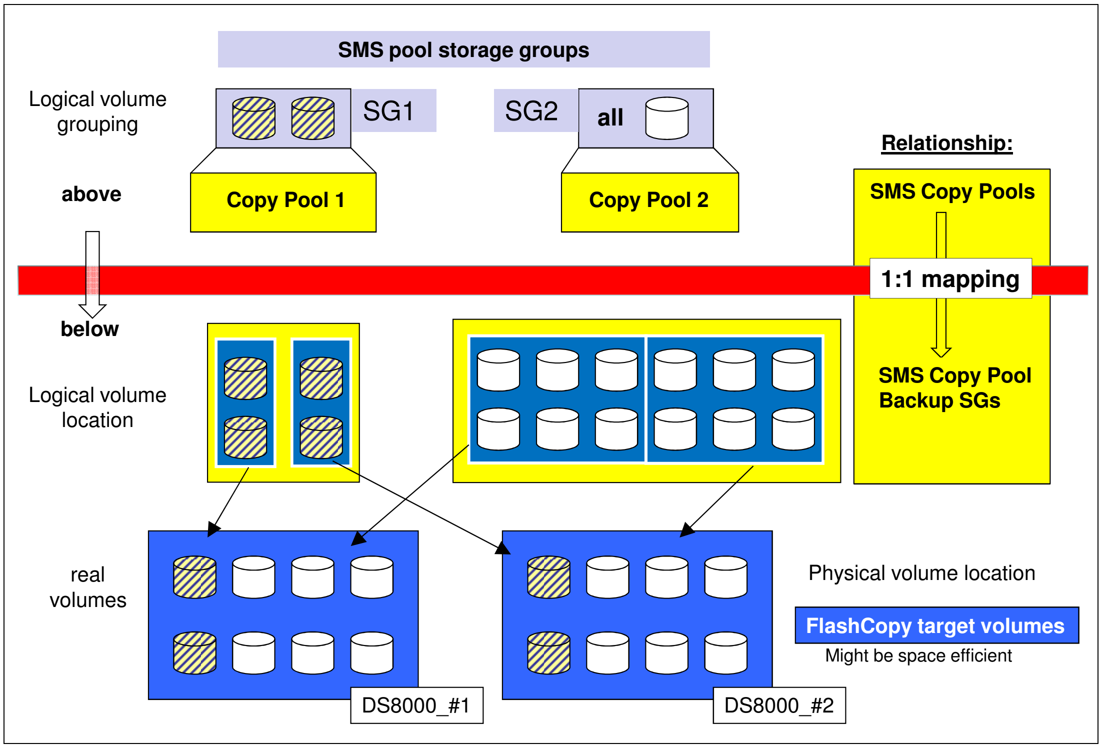


Figure 3-4 Combining SMS constructs

To keep the example simple, two copy pools are defined in Figure 3-4, each with its own CPBSG. In reality, more flexibility is available, if needed. For example, an SMS pool SG might need to belong to more than one copy pool, or a CPBSG might need to be shared by multiple copy pools.

3.2.4 The frbackup command

The DFSMShsm **frbackup** command starts the process to flash all volumes within the relevant copy pool to assign dynamically FlashCopy target volumes in the CPBSG. Figure 3-5 shows this process, which can be started only through the DFSMShsm FRBACKUP interface.

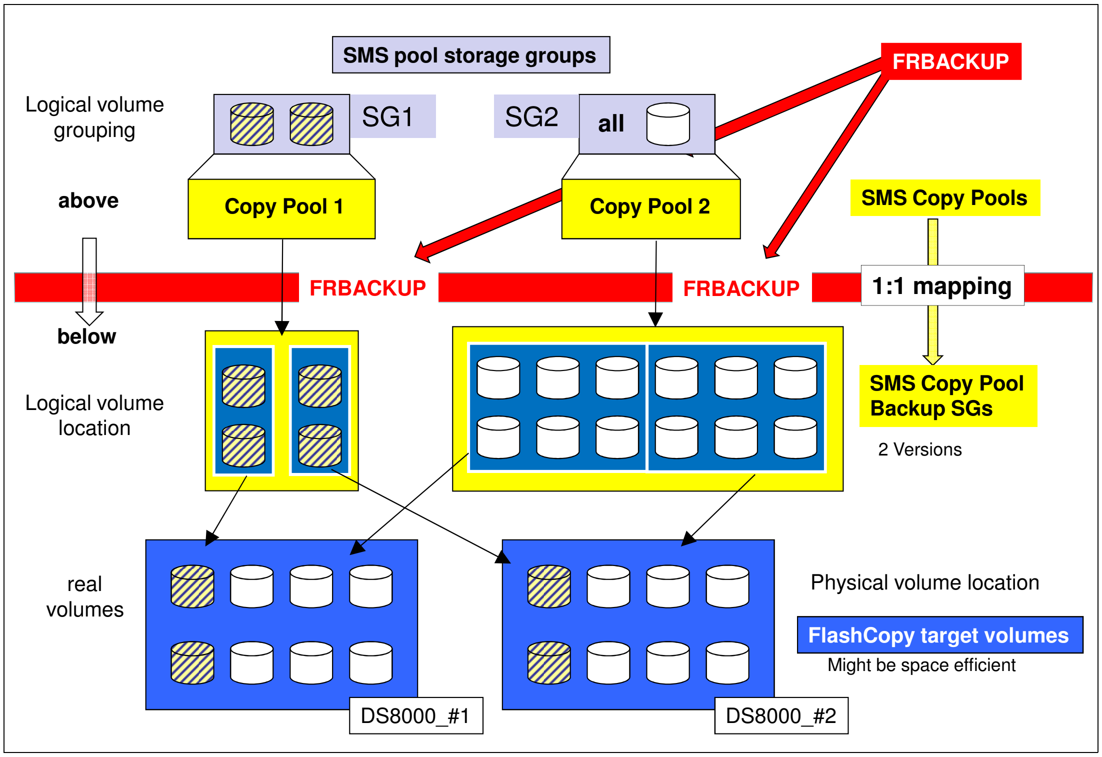


Figure 3-5 FRBACKUP is the interface to trigger FlashCopy based backups

DFSMShsm interfaces with SMS to select FlashCopy target volumes for each source volume from the copy pool. To shorten this FlashCopy volume-mapping process, the **frbackup** command can be issued with the **prepare** parameter.

The **frbackup** command that is issued through TSO is shown in Example 3-1. The **prepare** parameter triggers DFSMShsm to select a FlashCopy target volume for each copy pool-based volume and preserves this mapping until the FRBACKUP process issues all the FlashCopy copies.

Example 3-1 Issued frbackup command with the prepare parameter

```
hsend frbackup copypool (cp1) prepare

 ARC1000I COPY POOL CP1 FRBACKUP PROCESSING ENDED
 ARC1007I COMMAND REQUEST 00000582 SENT TO DFSMShsm
```

Another DFSMShsm **frbackup** command that is issued again through TSO and performs the FlashCopy operations is shown in Example 3-2.

Example 3-2 Issued frbackup command with the execute parameter

```
hsend frbackup copypool (cp1) execute

 ARC1000I COPY POOL CP1 FRBACKUP PROCESSING ENDED
 ARC1007I COMMAND REQUEST 00000587 SENT TO DFSMShsm
```

The example that is shown in Figure 3-1 on page 35 implies two versions exist for each FRBACKUP process with six FlashCopy target volumes in each DS8000 storage system for Copy Pool 2 and two FlashCopy target volumes in each DS8000 storage system for Copy Pool 1.

3.2.5 FRBACKUP software-based interaction

The software-based interaction within z/OS is shown in Figure 3-6. DFSMShsm has the leading role in managing the complete FRBACKUP process and communicates with the key address spaces, which are required to provide support to create successfully FlashCopy based volume copies.

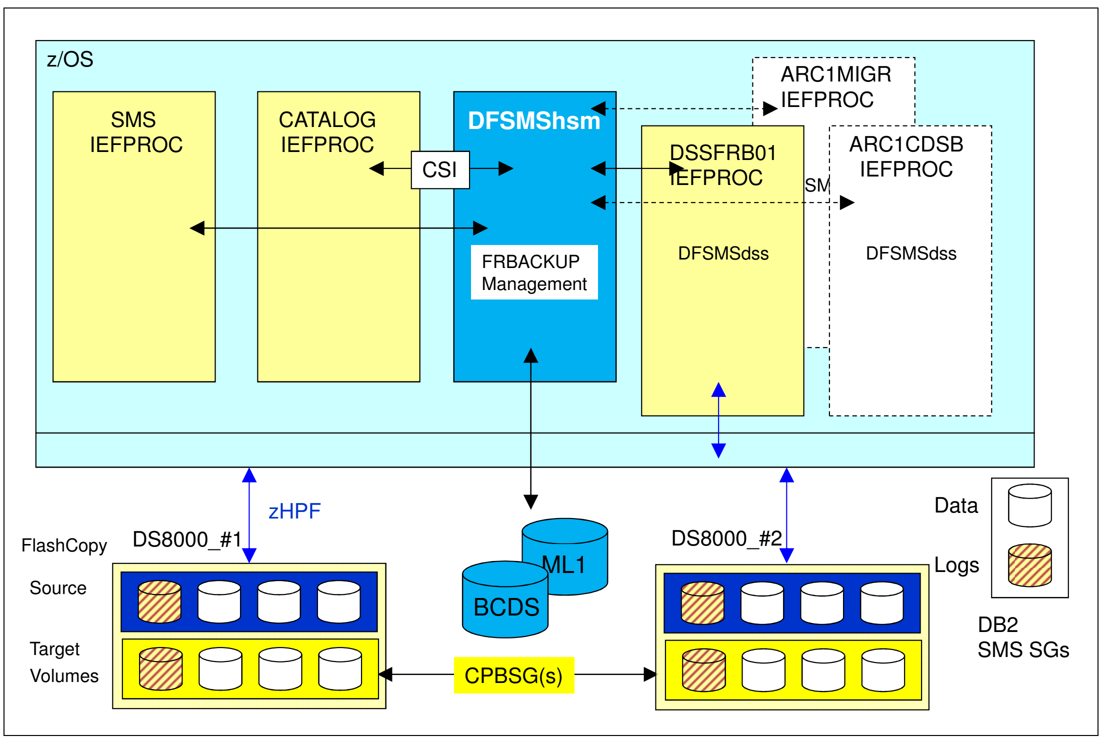


Figure 3-6 FRBACKUP software interaction within z/OS

The interaction that is shown in Figure 3-6 shows a Db2 based environment with logging volumes and table spaces within a single copy pool and its related CPBSG. Because FRBACKUP has a counterpart, Fast Replication Recover (FRRECOV), this FRBACKUP volume set in the CPBSG allows for a recovery on an individual table space (data set level). This capability requires getting all related catalog information during the FRBACKUP process.

To achieve a highly parallel process, DFSMShsm starts several DFSMSdss instances because DFSMSdss is the actual interface to FlashCopy in the DS8000 storage system. For more information, see *DFSMShsm Fast Replication Technical Guide*, SG24-7069.

3.3 Use of IBM FlashCopy with Db2 and IMS

The IBM FlashCopy (point-in-time copy) function in the DS8000 storage system has an extraordinary synergy with IBM Z storage management, which is delivered through a few powerful interfaces.

IBM FlashCopy opens many possibilities and several interesting use cases for backups of your entire system. It enhances testing capabilities by quickly cloning environments without application disruption. It also increases the availability of Db2 and IMS utilities.

The following interfaces can be used to configure and control IBM FlashCopy process for z/OS users:

- TSO commands
- IBM Device Support Facilities (ICKDSF)
- ANTRQST application programming interface (API)
- DFSMSdss

This section describes how Db2 and IMS databases use the DS8000 FlashCopy function.

3.3.1 IBM FlashCopy and Db2 for z/OS synergy

IBM Db2 for z/OS utilities frequently are enhanced to use more of the DS8000 FlashCopy functions. The following new Db2 versions included enhancements to use Db2, DS8000 storage systems, and z/OS synergy:

- V10 and V11:
  - Data set FlashCopy for COPY
  - Data set FlashCopy for inline copy in REORG TABLESPACE, REORG INDEX, REBUILD INDEX, and LOAD
  - FlashCopy image copies with consistency and no application outage (SHRLEVEL CHANGE)
  - FCIC accepted as input to RECOVER, COPYTOCOPY, DSN1COPY, DSN1COMP, and DSN1PRNT
- V12 and V13:
  - Prevent leaving the page set COPY-pending when REORG utility creates inline FlashCopy with no sequential inline image copy and FlashCopy fails.
  - Use the `COPY_FASTREPLICATION` parameter to identify whether fast replication is required, preferred, or not needed during the creation of FlashCopy image copy by the COPY utility.
  - System-level backup supports for multiple copy pools in which you can keep extra system-level backups on disk during upgrades.

Db2 BACKUP and RESTORE utility

The Db2 BACKUP SYSTEM utility is used to take a fast and minimally disruptive system-level backup. It backs up the entire data sharing group with all Db2 catalog and application data. The advantage of this utility is that you do not need to suspend or quiesce the Db2 logs and hold off the write I/Os. During the backup process, all the update activities are available. However, you cannot modify several items; for example, you cannot create, delete, and extend a data set.

The BACKUP SYSTEM calls the DFSMShsm **`frbackup`** command (as described in 3.1, "Introduction" on page 34), which is converted into DFSMSdss copy commands that start FlashCopy at a volume level.

The Db2 BACKUP SYSTEM supports incremental FlashCopy. The **`ESTABLISH FCINCREMENTAL`** keyword is used for that purpose. This keyword establishes the persistent FlashCopy relationship, and you use it only the first time that you start the BACKUP SYSTEM utility.

The next time that you start the BACKUP SYSTEM utility, it takes an incremental FlashCopy from the physical background copy. This persistent incremental FlashCopy relationship can be stopped by using the BACKUP SYSTEM **`END FCINCREMENTAL`** parameter.

After the BACKUP SYSTEM completes, the restore is done by using the RESTORE SYSTEM utility, which can restore a complete system. Like the BACKUP SYSTEM utility, RESTORE SYSTEM calls the DFSMShsm **`FRRECOV`** command, which is converted into DFSMSdss copy commands that start FlashCopy in the background.

Another possibility is to restore an individual Db2 object (table space or index) from the full system backup. Unlike the full BACKUP SYSTEM and RESTORE SYSTEM that use the FlashCopy on the volume level, the data-set-level FlashCopy is used when the single Db2 object is recovered. The **`SYSTEM_LEVEL_BACKUPS`** parameter must be set to `YES` (default is `NO`).

DFSMShsm is used in the background to extract the data set from the volume-level FlashCopy that is taken with BACKUP SYSTEM.

Db2 FlashCopy image copy utility

Db2 features an option to start data-set-level FlashCopy to create image copies. Although the Image Copy utility supports image copies, data-set-level FlashCopy can be used by any Db2 utility that creates an inline image copy, such as a Db2 reorg, load, rebuild index, and reorg index.

By using FlashCopy in the background, the IBM Z processing is offloaded to the DS8000 FlashCopy process. In addition, the Image Copy utility is completed when the FlashCopy relationship is established.

A FlashCopy image copy can be taken concurrently with share-level reference Image Copies, and without any effect on the applications.

Unlike the BACKUP SYSTEM, the FlashCopy image copy starts DFSMSdss directly. Because DFSMShsm is not involved, you do not need to plan your data placement (for example, Integrated Catalog Facility [ICF] catalogs). The only requirement is that all Db2 volumes must be SMS-managed.

When the FlashCopy image copy is created, it is available for restoration. When the recover utility restores the FlashCopy image copy object, it starts FlashCopy through DFSMSdss.

Check Index, Check Data, and Check large object utilities

Although the Check Index, Check Data, and Check large object (LOB) utilities are not used frequently, they are some of the most important utilities regarding data corruption analysis. These utilities can help you find the level of corruption and the best recovery processes.

Here, FlashCopy synergy with these Check utilities plays an important role. In a situation where you suspect data corruption, a prompt investigation is critical. Running the Check utilities while the application is running can save enormous amounts of time.

FlashCopy provides an option to run a Check utility nondisruptively with share-level change (by using the **`SHRLEVEL CHANGE`** parameter) included. The use of this option creates a shadow data set of data, indexes, and LOBs that can be checked by the utility that uses the DSFSMdss copy command (FlashCopy).

While the FlashCopy relationship is being established (which takes several seconds), the Check utility starts the use of the shadow data set (FlashCopy target data set). The shadow data set is in read-only mode only while the FlashCopy relationship is being established. After DFSMSdss confirms that the FlashCopy relationship is established, full read and write access to the data set is restored.

Therefore, without FlashCopy and the Check utility with share-level access, the shadow data set is created with standard I/O by using IBM Z resources rather than the DS8000 storage system. This process can be disruptive for the application. To avoid this situation, a best practice is to update the ZPARM FlashCopy **`fastreplication`** keyword to `REQUIRED` instead of `PREFERRED`, which is the default value in Db2 V10 and later. This parameter ensures that FlashCopy is used every time. If problems occur while FlashCopy is started, the Check utility fails, which is a better outcome than an application outage.

3.3.2 FlashCopy and IMS for z/OS synergy

Various IMS utilities use DFSMSdss to start the DS8000 FlashCopy function, which is the case with Db2. This section provides information about IMS High-Performance Image Copy (IMS HP IC), which is used for IMS backup and restore.

IMS High-Performance Image Copy

IMS HP IC provides fast back up and recovery of database data. Although it includes many services, we focus here on Advanced Image Copy Services, which allow IMS HP IC to produce faster image copies and increase availability time for IMS databases.

Advanced Image Copy Services use the DFSMSdss cross-memory API, ADRXMAIA, to process the DFSMSdss **`DUMP`** and **`COPY`** commands. These commands allow IMS HP IC to use the DS8000 FlashCopy advanced point-in-time copy function.

By using FlashCopy and IMS synergy, you can back up a database or any collection of data at a point in time and with minimum downtime for the database. The database is unavailable only long enough for DFSMSdss to initialize a FlashCopy relationship for the data (that is, data-set-level FlashCopy), which is a small fraction of the time that is required for a complete backup.

The copy that is made does not include any update activity. When the FlashCopy relationship is established, DFSMSdss releases all the serialization that it holds on the data, informs the Advanced Image Copy services about it so that update activity can resume, and begins reading the data.

The following Advanced Image Copy Services types are shown in Figure 3-7:

- COPY creates the clone data sets of DB data sets by using FlashCopy.

  The COPY process is used to create Image Copy Data Set (ICDS) quickly and to recover the database faster. This process is possible because FlashCopy is started as a background task.

- FDUMP accesses DB data sets by using FlashCopy and can create ICDS on the tape.

  Like the COPY process, FDUMP can be used when you must create a quick shadow of ICDS and dump it to tape.

- DUMP accesses DB data sets by using Concurrent Copy and can create ICDS on the tape.

  The DUMP process does not use FlashCopy, and it is limited only to the Concurrent Copy function.

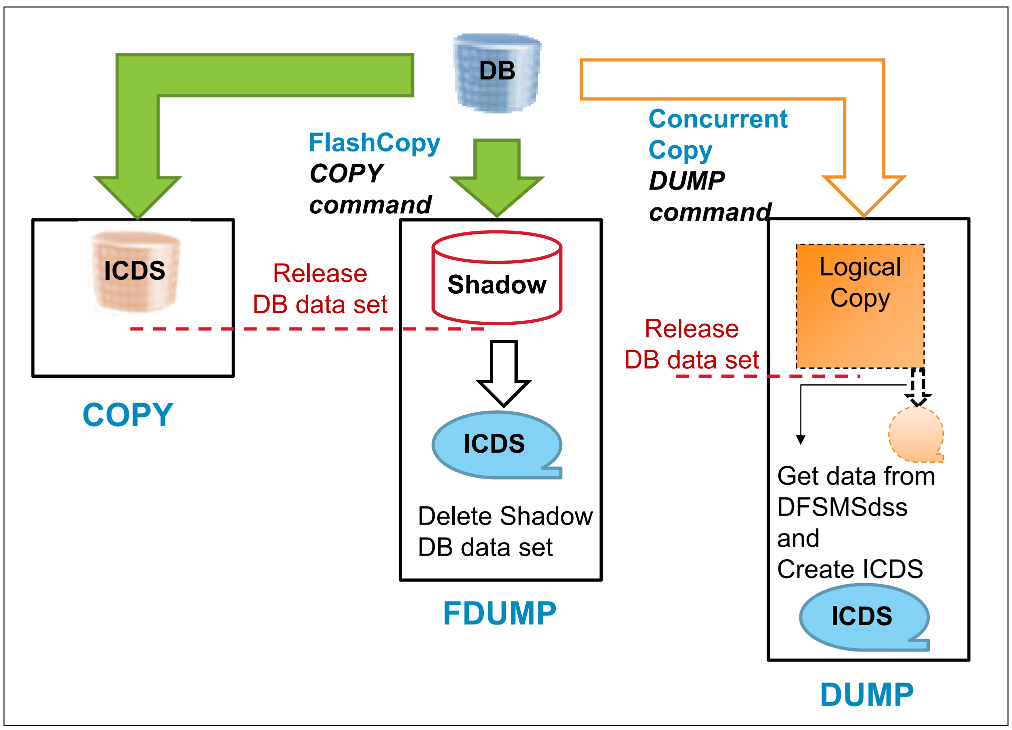


Figure 3-7 IMS Advanced Image Copy Services types

For more information about IMS and High-Performance Image Copy, see *IBM IMS High Performance Image Copy for z/OS User's Guide*, SC19-2756.

3.3.3 z/OS Distributed Data Backup

z/OS Distributed Data Backup (zDDB) is a licensed feature on the base frame that allows hosts, which are attached through a Fibre Channel connection (FICON) interface, to access data on Fixed Block (FB) volumes through a device address on FICON interfaces.

If zDDB is installed and enabled and a volume group type specifies FICON interfaces, this volume group features implicit access to all FB logical volumes that are configured in addition to all Count Key Data (CKD) volumes that are specified in the volume group. Then, with the suitable software, a z/OS host can complete backup and restore functions for FB logical volumes that are configured on a storage system image for open systems hosts.

4

Management and configuration

This chapter describes enhancements in z/OS that help to simplify storage management so that the system can manage storage automatically and autonomically. Enhancements over the years to today's z/OS Version 2 perfected the storage software within z/OS.

This chapter focuses on storage pool designs and specific enhancements that were jointly developed by the z/OS and DS8000 storage system development teams to achieve more synergy between both units.

This chapter includes the following topics:

4.1 Storage pool design considerations

Storage pool design considerations are a source of debate. Discussions about it originated in the early days when customers discovered that the growing number of their disk-based volumes was unmanageable.

> **Note:** The information in this section is important for former DS8000 hybrid configurations. Although still valid, it is not so relevant for all-flash D8000 systems.

IBM responded to this challenge by introducing system-managed storage and its corresponding system storage software. The approach was to no longer focus on volume awareness, but instead turn to a pool concept. The pool was the container for many volumes, and disk storage was managed at a pool level.

Eventually, storage pool design considerations also evolved with the introduction of storage systems, such as the DS8000 storage system, which offered other possibilities.

This section covers the system-managed storage and DS8000 views and how both are combined to contribute to the synergy between the IBM Z server and DS8000 storage systems.

4.1.1 Storage pool design considerations within z/OS

System storage software, such as Data Facility Storage Management Subsystem (DFSMS), manages information by creating a file or data set, setting its initial placement in the storage hierarchy, and managing it through its entire lifecycle until the file is deleted. The z/OS Storage Management Subsystem (SMS) can automate management storage tasks and reduce related costs. This automation is achieved through policy-based data management, availability management, space management, and even performance management, which DFSMS provides autonomically.

Ultimately, DFSMS is complemented by the DS8000 storage system and its rich variety of functions that work well with DFSMS and its components, as described in 4.1.4, “Combining SMS Storage Groups and DS8000 extent pools” on page 48.

This storage management starts with the initial placement of a newly allocated file within a storage hierarchy. It includes consideration for storage tiers in the DS8000 storage system where the data is stored. SMS assigns policy-based attributes to each file. Those attributes might change over the lifecycle of that file.

In addition to logically grouping attributes, such as the SMS Data Class (DC), SMS Storage Class (SC), and SMS Management Class (MC) constructs, SMS uses the concept of Storage Group (SG).

Finally, files are assigned to an SG or set of SGs. Ideally, the criteria for placing the file are solely dictated by the SC attributes and controlled by the last Automatic Class Selection (ACS) routine. A chain of ACS routines begins with an optional DC ACS routine, and a mandatory SC ACS routine that is followed by another optional MC ACS routine. This chain of routines is concluded by the SG ACS routine. The result of the SG ACS routine culminates in a list of candidate volumes where this file is placed.

Behind this candidate volume list is sophisticated logic to create the volume list. The z/OS components and measurements are involved in reviewing this volume list to identify the optimal volume for the file.

Therefore, the key construct from a pooling viewpoint is the SMS SG. The preferred approach is to create enough SMS SGs and populate each SMS SG with as many volumes as needed. Consider the volume size because you can create volumes with 27, 54, or larger extent numbers in the box.

Plan bigger volumes to huge data sets, and as much as possible, do not mix large and small data sets on bigger volumes. “Small” and “large” are abstract concepts that depend on each installation. This approach delegates to the system software the control for how to use and populate each single volume within the SMS SG.

The system software includes all the information about the configuration and the capabilities of each storage technology within the SMS SG. Performance-related service-level requirements can be addressed by SC attributes.

For more information about SMS constructs and ACS routines, see *z/OS DFSMSdfp Storage Administration*, SC23-6860.

> **Note:** In the newer DS8000 storage systems, the common practice for multitier extent pools is to enable Easy Tier. By doing so, the DS8000 microcode handles the data placement within the multiple storage tiers that belong to the extent pools for best performance results.

4.1.2 z/OS DFSMS class transition

Starting with z/OS 2.1, DFSMS (and specifically Data Facility Storage Management Subsystem Hierarchical Storage Manager [DFSMShsm]) is enhanced to support a potential class transition in an automatic fashion, which enables relocating a file within its L0 storage level from an SG into another SG. This relocation is performed during DFSMShsm automatic space management and is called *class transition* because ACS routines are exercised again during this transition process.

Based on a potentially newly assigned SC, MC, or a combination of both, the SG ACS routine then assigns an SG. This group might then be an SG that is different from the previous group that is based on the newly assigned SC within this transition process. This new policy-based data movement between SCs and the SG is a powerful function that is performed during DFSMShsm primary space management and by on-demand migration and interval migration.

> **Note:** During initial file allocation, SMS selects an SG that is defined in the ACS routine. However, this file’s performance requirements might change over time, and a different SG might be more suitable.

Various **`migrate`** commands were enhanced to support class transitions at the data set, volume, and SG level. For more information, see the following publications:

- *z/OS DFSMShsm Storage Administration*, SC23-6871
- *z/OS DFSMS Using the New Functions*, SC23-6857

4.1.3 DS8000 storage pools

The DS8000 architecture also uses the concept of pools or SGs as *extent pools*.

The concept of extent pools within the DS8000 storage system also evolved over time from homogeneous drive technologies within an extent pool to what today is referred to as a *hybrid extent pool*, which uses heterogeneous storage technology within the same extent pool. The use of a hybrid extent pool is made possible and efficient through the Easy Tier functions.

With Easy Tier, you can autonomically use up to three storage tiers in the most optimal fashion, based on workload characteristics. Although this goal is ambitious, Easy Tier for the DS8000 storage system evolved over the years.

In contrast to SMS, where the granularity or entity of management is a file or data set, the management entity is a DS8000 extent within Easy Tier. The Count Key Data (CKD) extent size is the equivalent of a 3390-1 volume with 1113 cylinders or approximately 0.946 GB as a DS8000 extent size when large extents are used.

When small extents are used, these extents are 21 cylinders (or approximately 17.85 MB), but are grouped in extent groups of 64 small extents for Easy Tier management. Automatic Easy Tier management within the DS8000 storage system migrates extents between technologies within the same extent pool or within an extent pool with homogeneous storage technology.

For more information, see *IBM DS8000 Easy Tier (Updated for DS8000 R9.0)*, REDP-4667.

4.1.4 Combining SMS Storage Groups and DS8000 extent pools

When comparing SMS storage tiering and DS8000 storage tiering, each approach has its own strength. SMS-based tiering addresses application and availability needs through initial data set or file placement according to service levels and policy-based management. It also gives control to the user and application regarding where to place the data within the storage hierarchy.

With DFSMS class transition, files can automatically move between different SMS SGs. As of this writing, this capability requires that the file is not open to an active application. From this standpoint, Easy Tier can relocate DS8000 extents without affecting a file that might be in use by an application.

Easy Tier does not understand an application’s needs when the data set is created. It intends to use the available storage technology in an optimal fashion by using faster types of flash for those parts of an application that are in heavy access and benefit most from better response times.

SMS storage tiering and DS8000 storage tiering are compared in Table 4-1.

Table 4-1 Comparing SMS tiering and DS8000 tiering

	SMS-based	DS8000-based
Movement entity	Data set or file level	Physical DS8000 extent
Management scope	Across DS8000 storage systems, but within a Parallel Sysplex	Within a DS8000 storage system, extent pool
Management level	Policy-based	I/O activity-based
Access	Closed files only	Open and closed files
Impact	File must be quiesced	Transparent (no impact)
Costs	Host MIPS	No host MIPS

Combining both tiering approaches is possible, and it might lead to the best achievable results from a total systems perspective. This combination provides tight integration between IBM Z and DS8000 storage system to achieve the best possible results economically with highly automated and transparent functions serving IBM Z customers.

4.2 Extended address volume enhancements

Today's large storage facilities tend to expand to larger CKD volume capacities. Some installations are nearing or are beyond the z/OS addressable unit control block (UCB) 64 KB limitation disk storage. Because of the four-digit, device-addressing limitation, larger CKD volumes must be defined by increasing the number of cylinders per volume.

As of this writing, an extended address volume (EAV) supports volumes with up to 1,182,006 cylinders (approximately 1 TB).

With the introduction of EAVs, the addressing changed from track to cylinder addressing. The partial change from track to cylinder addressing creates the following address areas on EAVs:

- Track-Managed Space: The area on an EAV that is within the first 65,520 cylinders. The usage of the 16-bit cylinder addressing allows a theoretical maximum address of 65,535 cylinders. To allocate more cylinders, you must have a new format to address the area above 65,520 cylinders.

  For 16-bit cylinder numbers, the track address format is `CCCCHHHH`, where:

  – `HHHH`: 16-bit track number
  – `CCCC`: 16-bit track cylinder

- Cylinder-Managed Space: The area on an EAV that is above the first 65,520 cylinders. This space is allocated in multicylinder units (MCUs), which of this writing is 21 cylinders. A new cylinder-track address format addresses the extended capacity of an EAV.

  For 28-bit cylinder numbers, the format is `CCCCcccH`, where:

  – `CCCC`: The low order 16 bits of a 28-bit cylinder number
  – `ccc`: The high order 12 bits of a 28-bit cylinder number
  – `H`: A 4-bit track number (0 - 14)

The following z/OS components and products now support 1,182,006 cylinders:

- DS8000 storage system and z/OS V1.R12 or later support CKD EAV volume sizes (3390 Model A: 1 - 1,182,006 cylinders [approximately 1004 TB addressable storage]).
- Configuration granularity:
  - 1-cylinder boundary sizes: 1 - 56,520 cylinders
  - 1113-cylinder boundary sizes: 56,763 (51 x 1113) - 1,182,006 (1062 x 1113) cylinders

The size of a Mod 3/9/A volume can be increased to its maximum supported size by using dynamic volume expansion (DVE). For more information, see 4.3, “Dynamic volume expansion” on page 54.

The volume table copy (VTOC) allocation method for an EAV volume was changed compared to the VTOC that was used for traditional smaller volumes. The size of an EAV VTOC index was increased four-fold and now has 8192 blocks instead of 2048 blocks.

Because no space remains inside the Format 1 data set control block (DSCB), new DSCB formats (Format 8 and Format 9) were created to protect programs from seeing unexpected track addresses. These DSCBs are known as *extended attribute DSCBs* (EADSCBs). Format 8 and 9 DSCBs are new for EAV. The Format 4 DSCB also was changed to point to the new Format 8 DSCB.

4.2.1 Data set type dependencies on an EAV

EAV includes several data set type dependencies. In all Virtual Storage Access Method (VSAM) sequential data set types, the following components can be placed on the extended addressing space (EAS):

- Extended, Basic, and Large format
- Basic direct-access method (BDAM)
- Partitioned data set (PDS)
- Partitioned data set extended (PDSE)
- VSAM volume data set (VVDS)
- Basic catalog structure (BCS)

This EAS is the cylinder-managed space of an EAV volume that is running on z/OS V1.12 and later.

EAV includes all VSAM data types, including the following examples:

- Key-sequenced data set (KSDS)
- Relative record data set (RRDS)
- Entry-sequenced data set (ESDS)
- Linear data set (LDS)
- IBM Db2
- IBM Information Management System (IMS)
- IBM CICS®
- IBM z/OS File System (zFS) data sets

The VSAM data sets that are placed on an EAV volume can be SMS or non-SMS managed.

For an EAV volume, the following data sets might exist but are not eligible to have extents in the EAS (cylinder-managed space):

- VSAM data sets with incompatible control area sizes
- VTOC (it is still restricted to the first 64 K - 1 tracks)
- VTOC index

- Page data sets
- A VSAM data set with **IMBED** or **KEYRANGE** attributes is not supported
- Hierarchical file system (HFS) file system
- SYS1.NUCLEUS

All other data sets can be placed on an EAV EAS.

You can expand all Mod 3/9/A volumes to a large EAV by using DVE. For a sequential data set, VTOC reformat is run automatically if REFVTOC=ENABLE is enabled in the **DEVSUPxx** parmlib member.

The data set placement on EAV as supported on z/OS is shown in Figure 4-1.

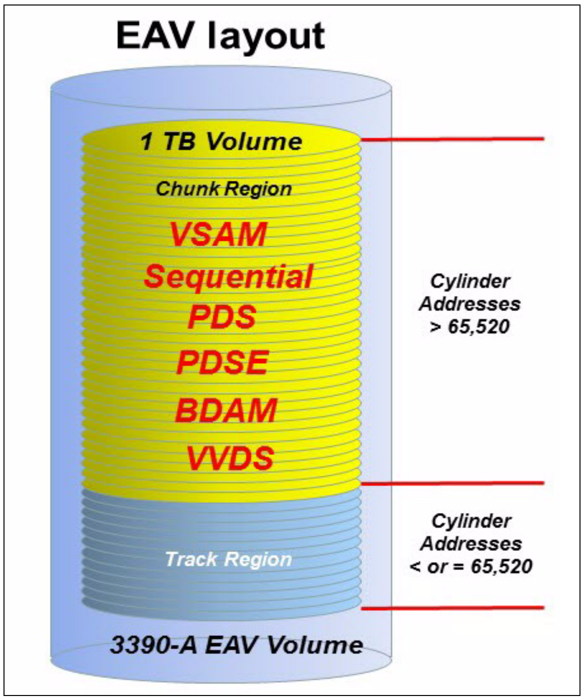


Figure 4-1 Data set placement on EAV

4.2.2 EAV volumes on z/OS

Consider the following points about EAV volumes:

- EAV volumes with 1 TB sizes are supported on z/OS V1.12 and later. A non-VSAM data set that is allocated with an EADSCB on z/OS V1.12 cannot be opened on earlier versions of z/OS V1.12.
- After a large volume is upgraded to 3390 Model A volume (an EAV with up to 1,182,006 cylinders) and the system is granted permission, an automatic VTOC refresh and index rebuild are run. The permission is granted by REFVTOC=ENABLE in parmlib member **DEVSUPxx**. The trigger to the system is a state change interrupt (SCI) that occurs after the volume expansion, which is presented by the storage system to z/OS.
- No other hardware configuration definition (HCD) considerations are available for the 3390 Model A definitions.

- On parmlib member **IGDSMSxx**, the **USEEAV(YES)** parameter must be set to allow data set allocations on EAV volumes. The default value is NO and prevents allocating data sets to an EAV volume. Example 4-1 shows a message that you receive when you are trying to allocate a data set on an EAV volume, and **USEEAV(NO)** is set.

Example 4-1 Message IEF021I with USEEVA set to NO

```
IEF021I TEAM142 STEP1 DD1 EXTENDED ADDRESS VOLUME USE PREVENTED DUE TO SMS USEEAV
(NO)SPECIFICATION.
```

- The new Break Point Value (BPV) parameter determines which size the data set must be allocated on a cylinder-managed area. The default for that the data set on parmlib member **IGDSMSxx** and in the SG definition (SG BPV overrides system-level BPV). The BPV value can be 0 - 65520, where 0 means that the cylinder-managed area is always preferred and 65520 means that a track-managed area is always preferred.

4.2.3 Identifying an EAV volume

Any EAV has more than 65,520 cylinders. To address this volume, the Format 4 DSCB was updated to x'FFFE', and DSCB 8+9 is used for cylinder-managed address space. Most of the eligible EAV data sets were modified by software with EADSCB=YES.

An easy way to identify any EAV that is used is to list the VTOC summary in TSO/ISPF option 3.4. Figure 4-2 shows the VTOC summary of a 3390 Model A with a size of 1 TB CKD usage.

```
Menu   Reflist   Refmode    Utilities    Help
 Volume . : SL9F05
 Command ===>

 Unit . . : 3390               Free Space

  VTOC Data                    Total            Tracks        Cyls
  Tracks  . :     2,069        Size  . . :   14,732,210     982,147
  %Used . . :         1        Largest . :   13,751,640     916,776
  Free DSCBS:   103,386        Free
                               Extents . :        3

  Volume Data                  Track Managed      Tracks        Cyls
  Tracks . :    17,730,090     Size  . . :       980,570      65,371
  %Used  . :            16     Largest . :       979,070      65,271
  Trks/Cyls:            15     Free
  F1=Help    F3=Exit   F12=Cancel
```

Figure 4-2 TSO/ISPF 3.4 panel for a 1 TB EAV volume: VTOC summary

When the data set list is displayed, enter one of the following commands:

- / in the data set list command field for the CLI, an ISPF line command, the name of a TSO command, CLIST, or REXX exec
- = to run the previous command

Important: Before EAV volumes are implemented, apply the latest z/OS maintenance levels. For more information, see this IBM Support web page.

4.2.4 EAV migration considerations

When you are planning to migrate to EAV volumes, consider the following items:

- Assistance

  Migration assistance is provided by the z/OS Generic Tracker Facility.
- Suggested actions:
  - Review your programs and look for calls for the following macros:
    - `OBTAIN`
    - `REALLOC`
    - `CVAFDIR`
    - `CVAFSEQ`
    - `CVAFDSM`
    - `CVAFFILT`

    These macros were modified, and you must update your program to reflect those changes.
  - Look for programs that calculate volume or data set size by any means, including reading a VTOC or VTOC index directly with a basic sequential access method (BSAM) or `EXCP` DCB. This task is important because now you have new values that are returning for the volume size.
  - Review your programs and look for the `EXCP` and `STARTIO` macros for direct access storage device (DASD) channel programs and other programs that examine DASD channel programs or track addresses. Now that a new addressing mode exists, programs must be updated.
  - Look for programs that examine any of the many operator messages that contain a DASD track, block address, data set, or volume size. The messages now show new values.
- Migrating data:
  - Define new EAVs by creating them on the DS8000 storage system or expanding volumes by using DVE.
  - Add new EAV volumes to SGs and storage pools, and update ACS routines.
  - Copy data at the volume level:
    - IBM Transparent Data Migration Facility (IBM TDMF)
    - Data Facility Storage Management Subsystem Data Set Services (DFSMSdss)
    - IBM DS8000 Copy Services Metro Mirror (MM) (formerly known as Peer-to-Peer Remote Copy (PPRC))
    - Global Mirror (GM)
    - Global Copy
    - FlashCopy
  - Copy data at the data set level:
    - DS8000 FlashCopy
    - SMS attrition
    - IBM z/OS Dataset Migration Facility (IBM zDMF)
    - DFSMSdss
    - DFSMShsm

4.3 Dynamic volume expansion

DVE simplifies management by enabling easier online volume expansion for IBM Z to support application data growth. It also supports data center migration and consolidation to larger volumes to ease addressing constraints.

The size of a Mod 3/9/A volume can be increased to its maximum supported size by using DVE. The volume can be dynamically increased in size on a DS8000 storage system by using the GUI or DS command-line interface (DS CLI).

Example 4-2 shows how the volume can be increased by using the DS8000 DS CLI.

Example 4-2 Dynamically expanding a CKD volume

```
dscli> chckdvol -cap 262268 -captype cyl 9ab0
CMUC00022I chckdvol: CKD Volume 9AB0 successfully modified.
```

DVE can be done while the volume remains online to the z/OS host system. When a volume is dynamically expanded, the VTOC and VTOC index must be reformatted to map the extra space. With z/OS V1.11 and later, an increase in volume size is detected by the system, which then performs an automatic VTOC and rebuilds the index.

The following options are available:

- **DEVSUPxx** parmlib options.

  The system is informed by SCIs, which are controlled by the following parameters:

  - `REFVTOC=ENABLE`

    With this option, the device manager causes the volume VTOC to be automatically rebuilt when a volume expansion is detected.

  - `REFVTOC=DISABLE`

    This parameter is the default. An IBM Device Support Facilities (ICKDSF) batch job must be submitted to rebuild the VTOC before the newly added space on the volume can be used. Start ICKDSF with **REFORMAT/REFVTOC** to update the VTOC and index to reflect the real device capacity.

    The following message is issued when the volume expansion is detected:

    `IEA019I dev, volser, VOLUME CAPACITY CHANGE,OLD=xxxxxxxx,NEW=yyyyyyyy`

- Use the **SET DEVSUP=xx** command to enable automatic VTOC and index reformatting after an IPL or disabling.
- Use the **F DEVMAN,ENABLE(REFVTOC)** command to communicate with the device manager address space to rebuild the VTOC. However, update the **DEVSUPxx** parmlib member to ensure it remains enabled across subsequent IPLs.

Note: For the DVE function, volumes cannot be in Copy Services relationships (point-in-time copy or FlashCopy, MM, GM, Metro/Global Mirror [MGM], or IBM z/OS Global Mirror [IBM zGM]) during expansion. Copy Services relationships must be stopped until the source and target volumes are at their new capacity and then, the Copy Service pair can be reestablished.

4.4 Quick initialization

Whenever the new volumes are assigned to a host, any new capacity that is allocated to it must be initialized. On a CKD logical volume, any CKD logical track that is read before it is written is formatted with a default record 0, which contains a count field with the physical cylinder and head of the logical track, record (R) = 0, key length (KL) = 0, and data length (DL) = 8. The data field contains 8 bytes of zeros.

A DS8000 storage system supports the quick volume initialization function (Quick Init) for IBM Z environments. Quick Init makes the newly provisioned CKD volumes accessible to the host immediately after they are created and assigned to it. The Quick Init function is automatically started whenever the volume is created or the existing volume is expanded.

It dynamically initializes the newly allocated space, which allows logical volumes to be configured and placed online to host more quickly. Therefore, manually initializing a volume from the host side is *not* necessary.

If the volume is expanded by using the DS8000 DVE function, normal read and write access to the logical volume is allowed during the initialization process. Depending on the operation, the Quick Init function can be started for the entire logical volume or for an extent range on the logical volume.

Quick Init improves device initialization speeds, simplifies the host storage provisioning process, and allows a Copy Services relationship to be established soon after a device is created.

4.5 Volume formatting overwrite protection

ICKDSF is the main z/OS utility to manage disk volumes (for example, for the initialize and reformat actions). In the complex IBM Z environment with many logical partitions (LPARs) in which volumes are assigned and accessible to more than one z/OS system, it is easy to erase or rewrite mistakenly the contents of a volume that is used by another z/OS image.

The DS8900F addresses this exposure through the Query Host Access function, which is used to determine whether target devices for specific script verbs or commands are online to systems where they should not be online. Query Host Access provides more useful information to ICKDSF about every system (including various sysplexes, virtual machine (VM), Linux, and other LPARs) that has a path to the volume that you are about to alter by using the ICKDFS utility.

The ICKDSF **VERIFYOFFLINE** parameter was introduced for that purpose. It fails an **INIT** or **REFORMAT** job if the volume is being accessed by any system other than the one performing the **INIT** or **REFORMAT** operation (as shown in Figure 4-3). The **VERIFYOFFLINE** parameter is set when the ICKDSF reads the volume label.

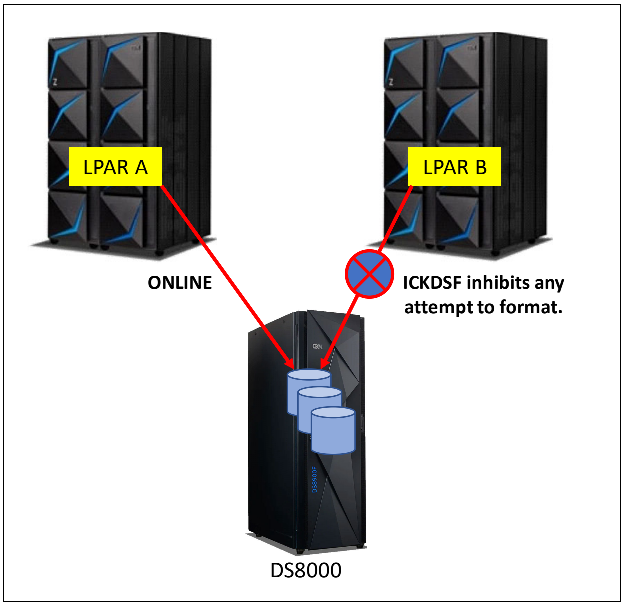


Figure 4-3 ICKDSF volume formatting overwrite protection

Messages that are generated soon after the ICKDSF **REFORMAT** starts and the volume is found to be online to some other system are shown in Example 4-3.

Example 4-3 ICKDSF REFORMAT volume

```
REFORMAT UNIT(8000) NVFY VOLID(DS8000) VERIFYOFFLINE
ICK00700I DEVICE INFORMATION FOR 8000IS CURRENTLY AS FOLLOWS:
PHYSICAL DEVICE = 3390
STORAGE CONTROLLER = 2107
STORAGE CONTROL DESCRIPTOR = E8
DEVICE DESCRIPTOR = 0E
ADDITIONAL DEVICE INFORMATION = 4A00003C
TRKS/CYL = 15, # PRIMARY CYLS = 65520
ICK04000I DEVICE IS IN SIMPLEX STATE
ICK00091I 9042 NED=002107.900.IBM.75.0000000xxxxx
ICK31306I VERIFICATION FAILED: DEVICE FOUND TO BE GROUPED
ICK30003I FUNCTION TERMINATED. CONDITION CODE IS 12
```

If this condition is found, the Query Host Access command from ICKDSF (**ANALYZE**) or **DEVSERV** (with the **QHA** option) can be used to determine what other z/OS systems have the volume online.

Example 4-4 shows the result of **DEVSERV** (or **DS** for short) with the **QHA** option.

Example 4-4 DEVSERV with the QHA option

```
-DS QD,037DF,QHA
 IEE459I 10.43.52 DEVSERV QDASD 027
  UNIT VOLSER SCUTYPE DEVTYPE       CYL  SSID SCU-SERIAL DEV-SERIAL EFC
 037DF XXY011 2107988 2107900     64554  80E0 0175-EYQ91 0175-EYQ91 *OK
   QUERY HOST ACCESS TO VOLUME
  PATH-GROUP-ID          FL STATUS  SYSPLEX  SYSTEM    MAX-CYLS
 80044604D73906DA43E1DA  50 ON      PLEXAA   AAAA       1182006
 80033509A73906DA219FA9  50 ON      PLEXBB   BBBB       1182006
 8003351E273906DA216966  50 ON      PLEXCC   CCCC       1182006
 80033503F73906DA2179BF  50 ON      PLEXDD   DDDD       1182006
 80033504D73906DA215B8D* 50 ON      PLEXEE   EEEE       1182006
 80044504D73906DA43FA7A  00 OFF     PLEXFF   FFFF       1182006
 80044503F73906DA449F33  50 ON      PLEXGG   GGGG       1182006
 8003361E273906DA2198AC  50 ON      PLEXHH   HHHH       1182006
 ****      8 PATH GROUP ID(S) MET THE SELECTION CRITERIA
 ****      1 DEVICE(S) MET THE SELECTION CRITERIA
 ****      0 DEVICE(S) FAILED EXTENDED FUNCTION CHECKING
```

This synergy between a DS8000 storage system and ICKDSF prevents accidental data loss and some unpredictable results. In addition, it simplifies the storage management by reducing the need for manual control.

The DS8000 Query Host Access function is used by IBM Geographically Dispersed Parallel Sysplex (GDPS), as described in 2.5, “Geographically Dispersed Parallel Sysplex” on page 26.

4.6 Channel paths and a control-unit-initiated reconfiguration

In the IBM Z environment, the standard practice is to provide multiple paths from each host to a storage system. Typically, four or eight paths are installed. The channels in each host that can access each logical control unit (LCU) in the DS8000 storage system are defined in the HCD or I/O configuration data set (IOCDS) for that host.

Dynamic Path Selection (DPS) allows the channel subsystem to select any available (nonbusy) path to start an operation to the disk subsystem. Dynamic Path Reconnect (DPR) allows the DS8000 to select any available path to a host to reconnect and resume a disconnected operation; for example, to transfer data after disconnection because of a cache miss.

These functions are part of IBM z/Architecture® and are managed by the channel subsystem on the host and the DS8000 storage system.

A physical Fibre Channel connection (FICON) path is established when the DS8000 port sees light on the fiber; for example, a cable is plugged into a DS8000 host adapter, a processor or the DS8000 storage system is powered on, or a path is configured online by z/OS.

Now, logical paths are established through the port between the host and some or all the LCUs in the DS8000 are controlled by the HCD definition for that host. This configuration occurs for each physical path between an IBM Z host and the DS8000 storage systems.

Multiple system images can be in a CPU. Logical paths are established for each system image. The DS8000 storage system then knows which paths can be used to communicate between each LCU and each host.

Control-unit-initiated reconfiguration (CUIR) varies off a path or paths to all IBM Z hosts to allow service for an I/O enclosure or host adapter. Then, it varies on the paths to all host systems when the host adapter ports are available. This function automates channel path management in IBM Z environments in support of selected DS8000 service actions.

CUIR is available for the DS8000 when it operates in the z/OS and IBM z/VM environments. CUIR provides automatic channel path vary on and off actions to minimize manual operator intervention during selected DS8000 service actions.

CUIR also allows the DS8000 storage system to request that all attached system images set all paths that are required for a specific service action to the offline state. System images with the suitable level of software support respond to such requests by varying off the affected paths and notifying the DS8000 that the paths are offline or that it cannot take the paths offline.

CUIR reduces manual operator intervention and the possibility of human error during maintenance actions and reduces the time that is required for the maintenance. This function is useful in environments in which many z/OS or z/VM systems are attached to a DS8000 storage system.

4.7 CKD thin provisioning

DS8000 storage systems allow CKD volumes to be formatted as thin-provisioned extent space-efficient (ESE) volumes. These ESE volumes perform physical allocation only on writes and only when another new extent is needed to satisfy the capacity of the incoming write block.

The allocation granularity and the size of these extents is 1113 cylinders or 21 cylinders, depending on how the extent pool was formatted. The use of small extents makes more sense in the context of thin provisioning.

One scenario to use such thinly provisioned volumes is for FlashCopy target volumes or GM Journal volumes so that the volumes can be space efficient while maintaining standard (thick) volume sizes for the operational source volumes.

Another scenario is to create *all* volumes as ESE volumes. In PPRC relationships, this idea has the advantage that on initial replication, extents that are not yet allocated in a primary volume do not need to be replicated, which also saves on bandwidth.

In general, thin provisioning requires tight control over the capacity that is free physically in the specific extent pool, especially when over-provisioning is performed. These controls are available along with respective alert thresholds and alerts that can be set.

4.7.1 Advantages

Thin provisioning can make storage administration easier. You can provision large volumes when you configure a new DS8900F storage system. You do not have to manage different volume sizes when you use a 3390 Model 1 size, Model 9, Model 27, and so on. All volumes can be large and of the same size.

A volume or device address is sometimes required to communicate with the control unit (CU), such as the utility device for Extended Remote Copy (XRC). Such a volume can include a minimum capacity. With thin provisioning, you still can use a large volume because less data is written to such a volume, its size in physical space remains small, and no storage capacity is wasted.

For many z/OS customers, migrating to larger volumes is a task they avoid because it involves substantial work. As a result, many customers have too many small 3390 Model 3 volumes. With thin provisioning, they can define large volumes and migrate data from other storage systems to a DS8900F storage system that is defined with thin-provisioned volumes and likely use even less space. Most migration methods facilitate copying small volumes to a larger volume. You refresh the VTOC of the volume to recognize the new size.

4.7.2 Advanced Volume Creation from GUI

The Management GUI now offers an easier way to create a single volume or multiple volume sets at the same time (of the same type). Thin-provisioned volumes can now be created without the use of the former Custom mode.

4.7.3 Space release

Space is released when either of the following conditions is met:

- A volume is deleted.
- A full FlashCopy Volume relationship is withdrawn and reestablished.

> **Note:** This condition is not true when working with data-set- or extent-level FlashCopy with z/VM minidisks. Therefore, use caution when you are working with the DFSMSdss utility because it might use data-set-level FlashCopy, depending on the parameters that are used.

A space release is also done on the target of an MM or Global Copy if the source and target are thin provisioned and the relationship is established.

Introduced with DS8000 code releases R8.2 and APAR OA50675 is the ability for storage administrators to perform an extent-level space release by using the DFSMSdss **SPACEREL** command. This volume-level command is used to scan and release free extents from volumes back to the extent pool.

The **SPACEREL** command can be issued for volumes or SGs and uses the following format:

```
SPACERel
   DDName(ddn)
   DYNam(volser,unit)
   STORGRP(groupname)
```

An enhancement was provided with the R8.3 code to release space on the secondary volume when the **SPACEREL** command is issued to an MM duplex primary device.

If issued to an MM suspended primary device, the space first is released in the primary device only. When the PPRC relationship is reestablished, the extents that were freed on the primary device also are released on the secondary device. The pair remains in the DUPLEX PENDING state until the extents on the secondary device are freed; the sync process resumes later.

Note: At the time of this writing, Global Copy primary devices must be suspended to allow the use of the `SPACEREL` command. Global Copy Duplex Pending devices are not supported for the `SPACEREL` command and devices in FlashCopy relationships.

For Multiple Target Peer-to-Peer Remote Copy (MT-PPRC) relationships, each relationship on the device must allow the `SPACEREL` command to run for the release to be allowed on the primary device (that is, the primary must be in a Suspended state for Global Copy or GM relationships, and in a Duplex or Suspended state in an MM relationship).

Suspended primary devices in a GM session are supported. Cascaded devices follow the same rules as noncascaded devices, although space is not released on the target because these devices are FlashCopy source devices.

4.7.4 Overprovisioning controls

Overprovisioning a storage system with thin provisioning brings the risk of running out of space in the storage system, which causes a loss of access to the data when applications cannot allocate space that was presented to the servers. To avoid this situation, clients typically use a policy regarding the amount of overprovisioning that they allow in an environment and monitor the growth of allocated space with predefined thresholds and warning alerts.

The current DS8000 codes provide clients with an enhanced method of enforcing such policies so that overprovisioning is capped to an *overprovisioning ratio* (see Figure 4-4), which does not allow further space allocations in the system.

$$Overprovisioning\ Ratio = \frac{Allocated\ Capacity\ (TP\ \&\ standard\ volumes)}{Total\ Capacity - Overhead\ capacity - Reserved\ Capacity}$$

Figure 4-4 Overprovisioning ratio formula

As part of the implementation project or permanently in a production environment, some clients might want to enforce an overprovisioning ratio of 100%, which means that no overprovisioning is allowed. (The use of this ratio does not risk affecting production because of running out of space on the DS8000 storage system.) By doing so, the Easy Tier and replication benefits of thin provisioning can be realized without the risk of accidentally overprovisioning the underlying storage. The overprovisioning ratio can be changed dynamically later, if wanted.

Implementing overprovisioning control results in changes to the standard behavior to prevent an extent pool from exceeding the overprovisioning ratio. As a result of these changes, the following actions are prevented:

- Volume creation, expansion, and migration
- Rank depopulation
- Pool merging
- Turning on Easy Tier space reservation

Overprovisioning controls can be implemented at the extent pool level, as shown in Example 4-5.

Example 4-5 Creating an extent pool with a 350% overprovisioning ratio limit

```
dscli> mkextpool -rankgrp 0 -stgtype fb -opratiolimit 3.5 -encryptgrp 1
test_create_fb
CMUC00000I mkextpool: Extent pool P8 successfully created.
```

An extent pool's overprovisioning ratio can be changed by running the **chextpool** DS CLI command, as shown in Example 4-6.

Example 4-6 Changing the overprovisioning ratio limit on P3 to 3.125

```
dscli> chextpool -opratiolimit 3.125 p3
CMUC00001I chextpool: Extent pool P3 successfully modified.
```

To display the overprovisioning ratio of an extent pool, run the **showextpool** DS CLI command, as shown in Example 4-7.

Example 4-7 Displaying the current overprovisioning ratio of extent pool P3 and the limit set

```
dscli> showextpool p3

...
%limit                100
%threshold            15
...
opratio               0.96
opratiolimit          3.13
%allocated(ese)       2
%allocated(rep)       0
%allocated(std)       77
%allocated(over)      0
%virallocated(ese)    -
%virallocated(tse)    -
%virallocated(init)   -
...
```

4.8 DS CLI on z/OS

Another synergy item between IBM Z and a DS8000 storage system is that you can install the DS CLI along with IBM Copy Services Manager (CSM) 6.1.4 and later on a z/OS system. It is a regular SMP/E for z/OS installation.

The DS CLI runs under UNIX System Services for z/OS and includes a separate FMID HIWN61K. You can also install the DS CLI separately from CSM.

For more information, see the *IBM DS CLI on z/OS Program Directory*, GI13-3563.

After the installation is complete, access your UNIX System Services for z/OS, which can vary among installations. One common way to access these services is by using TSO option 6 (ISPF Command Shell) and the **OMVS** command. For more information, contact your z/OS System Programmer.

Access to DS CLI in z/OS is shown in Figure 4-5. It requests the same information that you supply when you are accessing DS CLI on other platforms.

```
$ cd /opt/IBM/CSMDSCLI
$ ./dscli
Enter the primary management console IP address: <enter-your-DS8K-machine-ip-address>
Enter the secondary management console IP address:
Enter your username: <enter-your-user-name-as-defined-on-the-machine>
Enter your password: <enter-your-user-password-to-access-the-machine>
dscli> ver -l
...
dscli>
INPUT
ESC=¢     1=Help      2=SubCmd    3=HlpRetrn  4=Top       5=Bottom    6=TSO       7=BackScr   8=Scroll
9=NextSess 10=Refresh 11=FwdRetr  12=Retrieve
```

Figure 4-5 Accessing DS CLI on z/OS

Some DS CLI commands that are run in z/OS are shown in Figure 4-6.

```
dscli> lssi
Name          ID            Storage Unit    Model WWNN           State  ESSNet
==================================================================================
IBM.2107-75ACA91 IBM.2107-75ACA91 IBM.2107-75ACA90 980  5005076303FFD13E Online Enabled

dscli> lsckdvol -lcu EF
Name    ID  accstate datastate configstate deviceMTM voltype  orgbvols extpool cap (cyl)
=====================================================================================
ITSO_EF00 EF00 Online  Normal   Normal    3390-A   CKD Base -      P1      262668
ITSO_EF01 EF01 Online  Normal   Normal    3390-9   CKD Base -      P1       10017
ITSO_EF02 EF02 Online  Normal   Normal    3390-3   CKD Base -      P1        3339
dscli> rmckdvol EF02
CMUC00023W rmckdvol: The alias volumes that are associated with a CKD base volume are automatically
deleted before deletion of the CKD base volume. Are you sure you want to delete CKD volume EF02? Ÿy/n¨:
y
CMUC00024I rmckdvol: CKD volume EF02 successfully deleted.
dscli>
```

Figure 4-6 Common commands on DS CLI

With this synergy, you can use all z/OS capabilities to submit batch jobs and perform DS CLI functions, such as creating disks or LCUs.

4.9 Lightweight Directory Access Protocol authentication

The DS8000 storage systems allow directory services-based user authentication.

By default, the DS8000 authentication is based on local user management. Maintaining local repositories of users and their permissions is convenient and straightforward when dealing with only a few users and only a few DS8000 servers or other systems.

However, as the number of users and interconnected systems grows, the benefits of a centralized user management approach can be substantial. Here, the DS8000 works together with external Lightweight Directory Access Protocol (LDAP) servers.

As shown in Figure 4-7, the DS8900F can connect to a wider variety of LDAP server types natively, starting with code release 9.1. These types include Resource Access Control Facility (IBM RACF®) and CA Top Secret for z/OS.

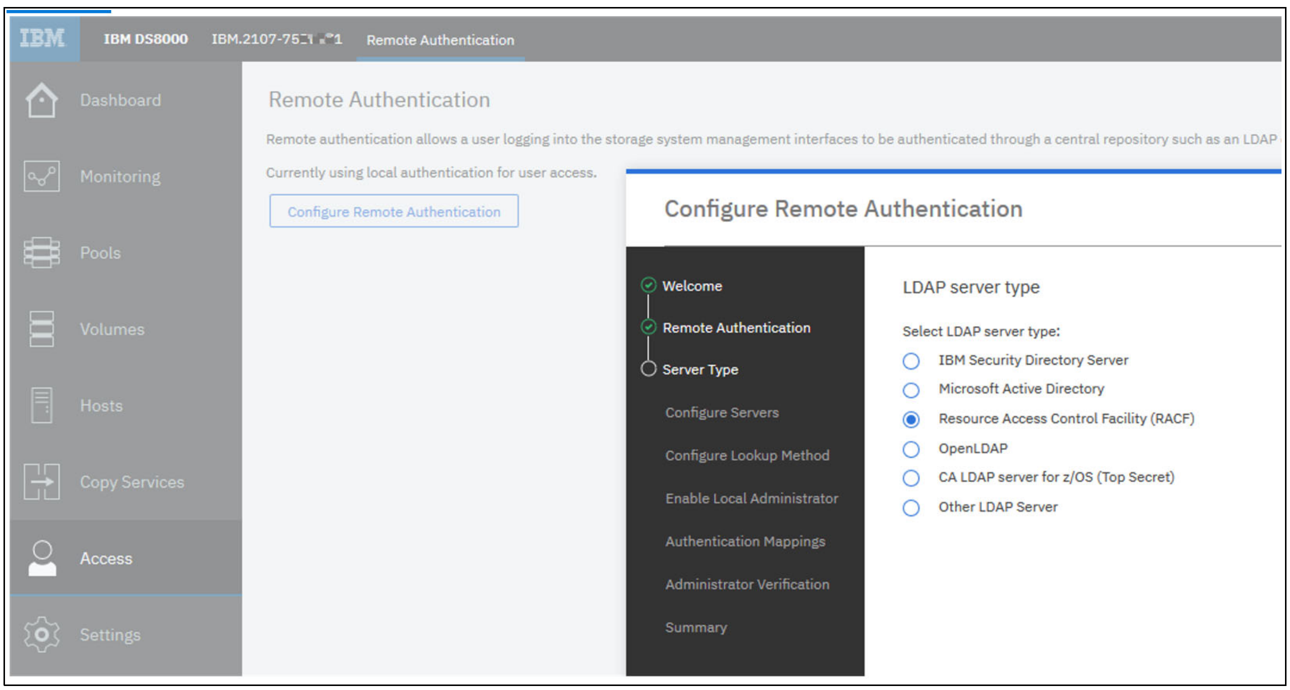


Figure 4-7 Connecting the DS8900F to LDAP servers

For more information about the DS8900F using LDAP, see *LDAP Authentication for IBM DS8000 Systems: Updated for DS8000 Release 9.1*, REP-5460.

4.10 GUI performance reporting

The DS8000 Storage Management GUI allows storage performance reporting on the following levels:

- Full-system
- Pool level
- Array
- Port
- LSS/LCU
- Volume

Volume-level performance reporting provides a detailed performance metrics view per volume within the management GUI. This feature helps administrators with simplifying and optimizing tasks to improve their ability to solve problems.

The newer DS8000 codes make it easy to access data that is related to I/Os per second (IOPS), latency, and bandwidth from the logical system and host performance. This configuration enables faster troubleshooting for host applications that are based on graphics, and easy data analysis in the management GUI.

From the management GUI and DS CLI, it is possible to access report files that can be exported and uploaded into the IBM Storage Modeller tool for performance modeling and simulation by the IBM and IBM Business Partners storage technical sales and specialists teams.

Performance Graphs report

The Performance Graphs report is available on the Management GUI. It displays the performance metrics for one or more resources on the storage. It is possible to select the resource and metrics in the graph for the past 7 days. The graphs are refreshed automatically as new data becomes available.

Figure 4-8 shows the Performance Graphs report.

Figure 4-8 Performance Graphs report

This report includes the following features:

- Legend

  Resources. such as system, pool, I/O port, metrics, and measurements, are displayed in the upper right of the graph, as shown in Figure 4-8. More than one resource or metric can be displayed in different colors to facilitate visualization.

- Timeline

  The timeline is the horizontal line at the bottom of the graph. You can see the performance history by dragging the timeline control. You also can zoom in and out by using the mouse wheel.

- Split-screen view

  The split-screen view allows two graphs to be shown simultaneously. By clicking the split-screen icon at the right, you can adjust the timelines independently. To return to a single graph mode, click the unlink icon at the right.

- Preset graphs

  The performance page includes the following preset graphs:

  – System IOPS: Displays read, write, and total requests in thousand I/O per second (KIOPS) averaged over 1 minute.

  – System Latency: Displays the response time in milliseconds (ms) for read/write operations averaged over 1 minute.

> **Note:** The System Latency graph for the storage system that is connected to an IBM Z host includes only partial end-to-end response times on a system level for CKD volumes. For detailed system level response times, see the Resource Measurement Facility (RMF) on the host.

- System Bandwidth: Displays the number of megabytes per second (MBps) for read, write, and total bandwidth averaged over 1 minute.
- System Cache: Displays the percentage of total read I/O operations that were fulfilled from the storage system cache and write I/O operations that were delayed because of write cache limitations. Both metrics are averaged over 1 minute.

For more information about volume-level and other levels of performance reporting, see this IBM Documentation web page.

4.11 Fibre Channel connectivity report

Since DS8000 Release 9.2, the management GUI can generate a report for the attached port, all host logins and replication logins for selected ports in the standard configuration, and the security status for each IBM Fibre Channel Endpoint Security login.

The Fibre Channel Connectivity report is a CSV file that contains information about all login FICON connections, including open systems host connections, IBM Z, LinuxONE servers, or other DS8000 storage systems. This report helps the system administrator with the optimization of system resources. It also better understands the storage area network (SAN) configuration and helps with problem determination.

The CSV file contains the following information:

- Local Port ID
- Local Port Fibre Channel ID
- Local Port WWPN
- Local Port WWNN
- Local Port Security Capability
- Local Port Security Configuration
- Local Port Logins
- Local Port Security Capable Logins
- Local Port Authentication Only Logins
- Local Port Encrypted Logins
- Attached Port WWPN
- Attached Port WWNN
- Attached Port Interface ID
- Attached Port Type
- Attached Port Model
- Attached Port Manufacturer
- Attached Port S/N
- Remote Port WWPN
- Remote Port WWNN
- Remote Port Fibre Channel ID
- Remote Port PRLI Complete
- Remote Port Login Type
- Remote Port Security State
- Remote Port Security Config
- Remote Port Interface ID

- Remote Port Type
- Remote Port Model
- Remote Port S/N
- Remote Port Manufacturer
- Remote Port System Name

For more information about the Fibre Channel connectivity report, see this IBM Documentation web page.

4.12 Encryption

IBM Z features the concept of *pervasive encryption*, which is the idea of having encryption everywhere. Applying it to storage, it covers encrypting at-rest and in-flight data.

By using the self-encrypting drives (SEDs), the DS8000 for many years offered data at-rest (DAR) encryption. Traditionally, this encryption was combined with a key manager software and servers for storing the encryption keys.

For data-in-flight encryption, the DS8000 supports the IBM Fibre Channel Endpoint Security between host and DS8000 ports. For more information, see 5.11, “IBM Fibre Channel Endpoint Security since IBM z15” on page 115.

For offloading data to tape or to a cloud tier (see 5.13, “Transparent Cloud Tiering” on page 116), encryption also can be used.

With the TCT Secure Data Transfer (SDT) option, encryption of Data-in-Flight (EDiF) can occur while the data is moved over to a tape Grid network. It also can be decrypted again when arriving at the TS7700 cluster.

TCT SDT does *not* require an external key manager. Hardware acceleration is used in the POWER CECs of the DS8900F to offload CPU cycles to a crypto engine.

When TCT is used, an option is available for client-side encryption of all data; that is, data lands in the object store as encrypted and is decrypted only when it is recalled by DS8900F. This method requires an external key manager, such as IBM Security Guardium Key Lifecycle Manager.

4.12.1 Encryption without external key manager servers

For more information about external key managers, see *IBM DS8000 Encryption for Data at Rest, Transparent Cloud Tiering, and Endpoint Security (DS8000 Release 9.2)*, REDP-4500.

Since DS8000 Release 9.2, a new optional license supports local encryption. Running on the DS8000 Processor Complexes, a local key manager performs the enablement and key management for encryption at rest, as shown in Figure 4-9.

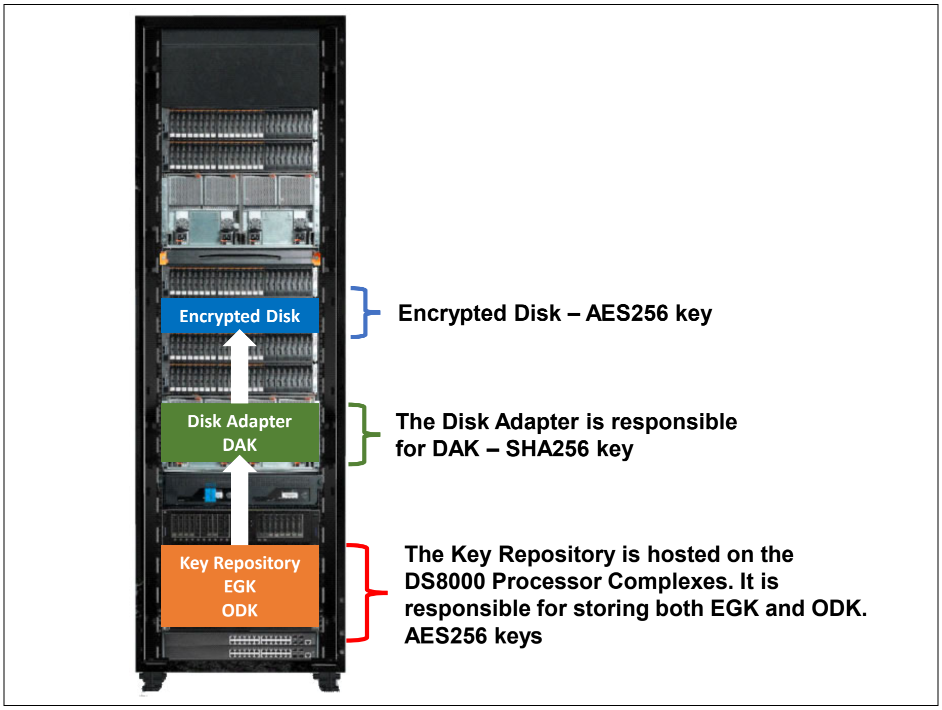


Figure 4-9 Internal key management

> **Note:** An external key manager is still required for TCT data-at-rest encryption, or for IBM Fibre Channel Endpoint Security.

Local encryption on DS8000 works by storing an Encrypted Group Key (EGK) and Obfuscated Data Key (ODK), DS8000 decrypts the ODK by using a known key. The DS8000 uses a Data Key (DK) to decrypt the Group Key (GK). The disk adapters use the GK to derive the Drive Access Key (DAK) that is unique to each drive. DAK is used to decrypt the Drive Encryption Key (DEK) and read data from the drive.

With the internal key management, DS8000 generates a DK, obfuscates it, and stores it internally rather than storing the key in an external server.

4.12.2 Encryption with GKLM and EKMF Web Integration

Beginning with GKLM 4.1.1, GKLM can store the GKLM master key in ICSF by way of EKMF Web 2.1. The GKLM master key is generated by the HSM by using EKMF Web REST APIs, which also are used to wrap and unwrap the data keys that are stored in Db2 or Postgres.

With Release 9.3, DS8900 now supports GKLM Integration with Redundant EKMF Web Instances. It protects the GKLM Master Key in ICSF and Crypto Express Cards.

With enhanced Connectivity to GKLM Containerized Edition (CE), DS8900 now is ready for Fibre Channel Endpoint Security (FCES) and is prepared for future full support that is planned with IBM Z for GKLM CE.

Deployed as a Docker container in a z/OS Container Extension (zCX), the GKLM CE adds KMIP capabilities to z/OS. z/OS teams can continue to use IBM Security Key Lifecycle Manager for z/OS to manage storage devices and add GKLM for zCX to manage new storage devices or use KMIP required capabilities.

For more information, see the following resources:

- *IBM Fibre Channel Endpoint Security for IBM DS8900F and IBM Z*, SG24-8455
- IBM Security Guardium Key Lifecycle Manager documentation

4.13 Transparent Cloud Tiering multi-cloud support

Since DS8000 Release 9.2, TCT can be configured to support up to eight clouds, which can be on-premises, off-premises, and on an IBM TS7700.

The definition of multiple SMS cloud network connection constructs must be done on DFSMShsm by specifying which types of data must be moved to the TS7700 object store and which other types of data must be moved to on-premises or off-premises clouds. This capability can also be used to separate test data, development data, and reduction data through service-level agreements to various cloud targets without reconfiguring the system.

The first cloud that is created is automatically activated. All clouds that are created after the first cloud must be manually activated by using the `managecloudserver` command.

Figure 4-10 shows the TCT multi-cloud support.

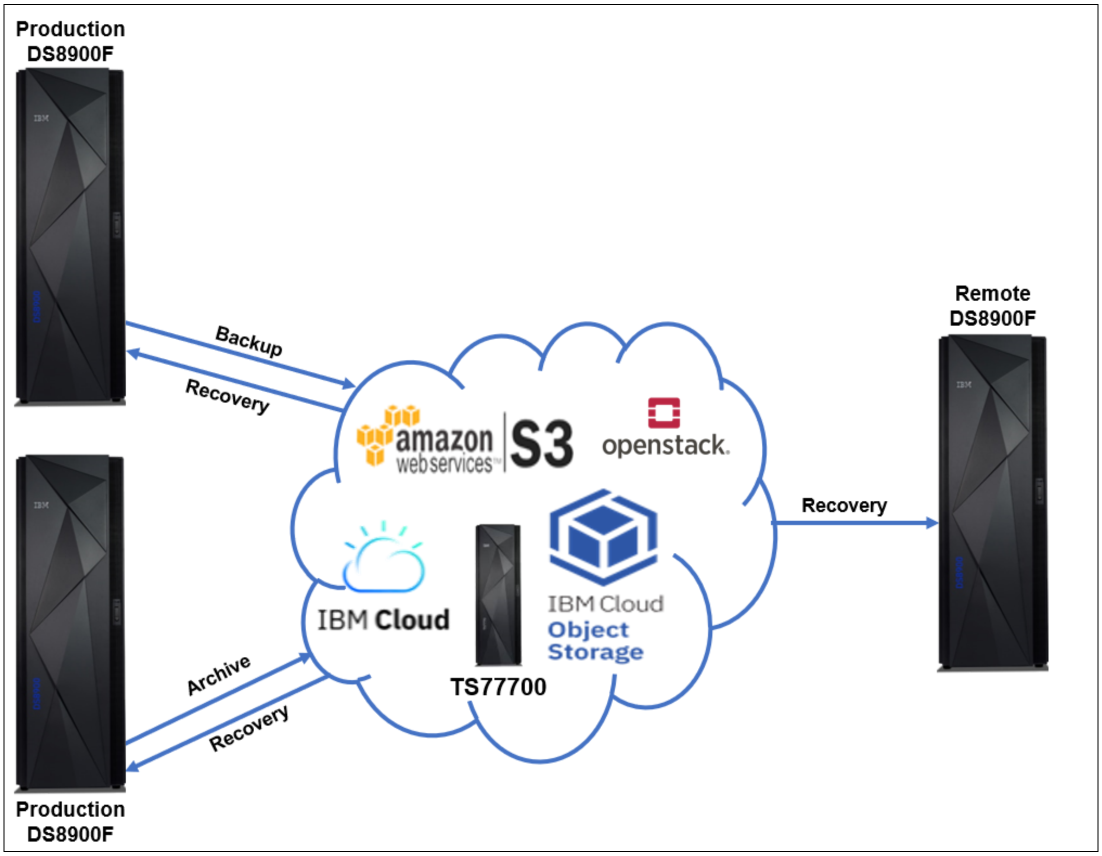


Figure 4-10 Transparent Cloud Tiering multi-cloud support

The following types of clouds are supported:

- Swift: A DS8000 system can use cloud solutions that are based on OpenStack Swift to encrypt authentication credentials and connect to a storage target.
- Swift-Keystone: You can use Swift-Keystone to encrypt authentication credentials and connect to a cloud storage target. Authentication is done by using root or system certificates with Secure Sockets Layer (SSL) or Transport Layer Security (TLS).
- TS7700: Used for authentication and data stores on an IBM TS7700 system.
- AWS-S3: A DS8000 system can authenticate and connect to S3 storage through the Amazon Simple Storage Service (Amazon S3) protocol.
- IBM Cloud® Object Storage: You can use IBM Cloud Object Storage for data protection through backup and recovery.
- S3: The DS8000 system can authenticate and store data on any other S3 Cloud that is a storage target.

For more information, see the following resources:

- *IBM DS8000 Transparent Cloud Tiering (DS8000 Release 9.2)*, SG24-8381
- 5.13, “Transparent Cloud Tiering” on page 116

5

IBM Z and DS8000 performance

This chapter describes the IBM Z and DS8000 synergy features from a performance perspective. It also discusses how these features contribute to a robust and efficient overall mainframe storage infrastructure.

This chapter includes the following topics:

5.1 Parallel access volume, HyperPAV, and SuperPAV

Parallel access volume (PAV) is a function of the DS8000 storage system for the z/OS and z/VM operating systems to concurrently share logical volumes. This ability to handle multiple I/O requests to the same volume nearly eliminates Input/Output Supervisor queue (IOSQ) delay time, which is one of the major components that affect z/OS response time.

Traditionally, access to highly active volumes involved manual tuning, splitting data across multiple volumes, and more. With PAV and the Workload Manager (WLM), you can almost forget about manual performance tuning. WLM also manages PAVs across all the members of a sysplex.

Traditional z/OS behavior without PAV

Traditional storage disk systems (which allow for only one channel program to be active to a volume at a time to ensure that data that is accessed by one channel program) cannot be altered by the activities of another channel program.

The traditional z/OS behavior without PAV, where subsequent simultaneous I/Os to volume 100 are queued while volume 100 is still busy with a previous I/O, is shown in Figure 5-1.

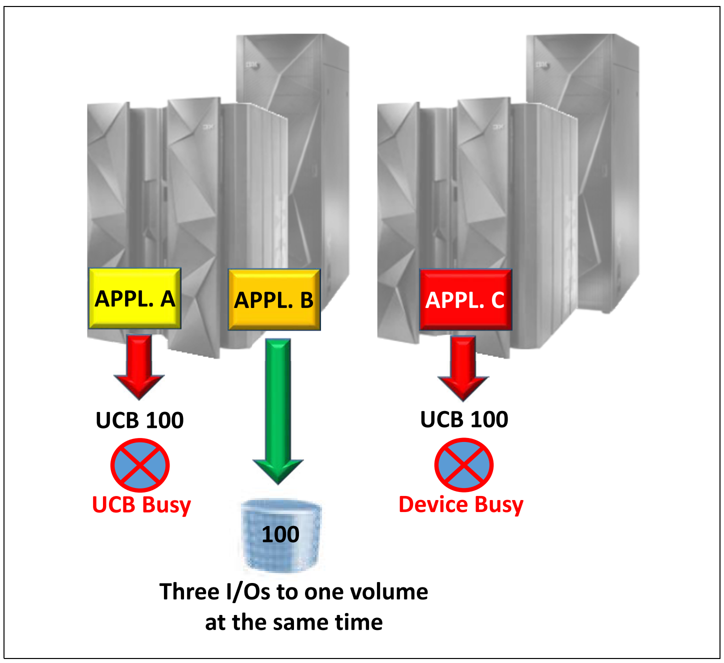


Figure 5-1 Traditional z/OS behavior

From a performance perspective, sending more than one I/O at a time to the storage system did not make sense because the hardware processes only one I/O at a time. With this information, the z/OS systems did not try to issue another I/O to a volume (which, in z/OS, is represented by a unit control block [UCB]) while an I/O was active for that volume, which was indicated by a UCB busy flag.

Not only did the z/OS systems process only one I/O at a time, but the storage systems accepted only one I/O at a time from different system images to a shared volume.

Parallel I/O capability: z/OS behavior with PAV

The DS8000 storage system runs more than one I/O to a Count Key Data (CKD) volume. By using the alias address and the conventional base address, a z/OS host can use several UCBs for the same logical volume instead of one UCB per logical volume. For example, base address 100 might include alias addresses 1FF and 1FE, which allows for three parallel I/O operations to the same volume, as shown in Figure 5-2.

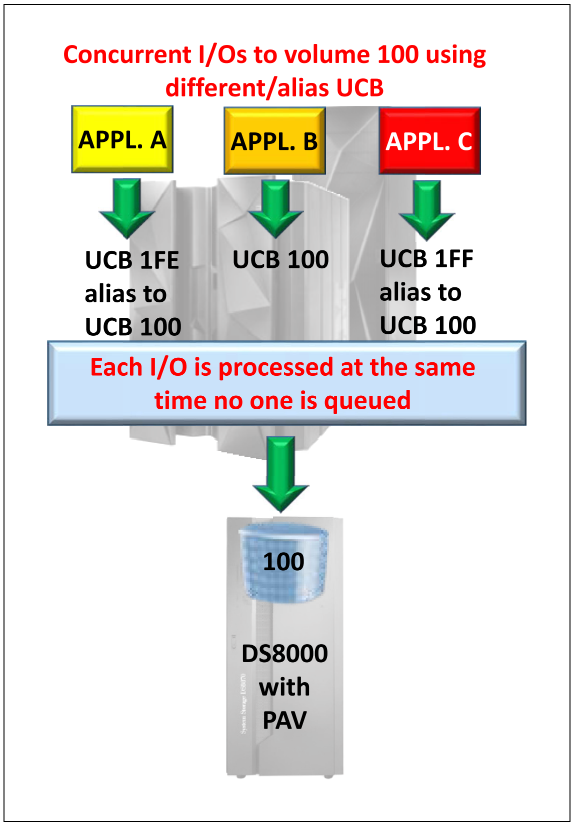


Figure 5-2 z/OS behavior with PAV

PAV allows parallel I/Os to a volume from one host. The following basic concepts are featured in PAV functions:

- Base device address

  The base device address is the conventional unit address of a logical volume. Only one base address is associated with any volume.

- Alias device address

  An alias device address is mapped to a base address. I/O operations to an alias run against the associated base address storage space. No physical space is associated with an alias address. You can define more than one alias per base.

Alias addresses must be defined to the DS8000 storage system and to the I/O definition file (IODF). This association is predefined, and you can add aliases nondisruptively. In the static PAV, the relationship between the base and alias addresses is static. Dynamically assigning alias addresses to your base addresses reduces the number of aliases that are required in your system.

For more information about PAV definition and support, see *IBM System Storage DS8000: Host Attachment and Interoperability*, SG24-8887.

5.1.1 Dynamic PAV tuning with z/OS Workload Manager

Predicting which volumes must include an alias address that is assigned and how many alias addresses are included is not always easy. Your software can automatically manage the aliases according to your goals. z/OS can use automatic PAV tuning if you use the z/OS WLM in goal mode.

The WLM can dynamically tune the assignment of alias addresses. The WLM monitors the device performance and can dynamically reassign alias addresses from one base to another base if predefined goals for a workload are not met.

z/OS recognizes the aliases that are initially assigned to a base during the nucleus initialization program phase. If dynamic PAVs are enabled, the WLM can reassign an alias to another base by instructing the Input/Output Supervisor (IOS) to do so when necessary, as shown in Figure 5-3.

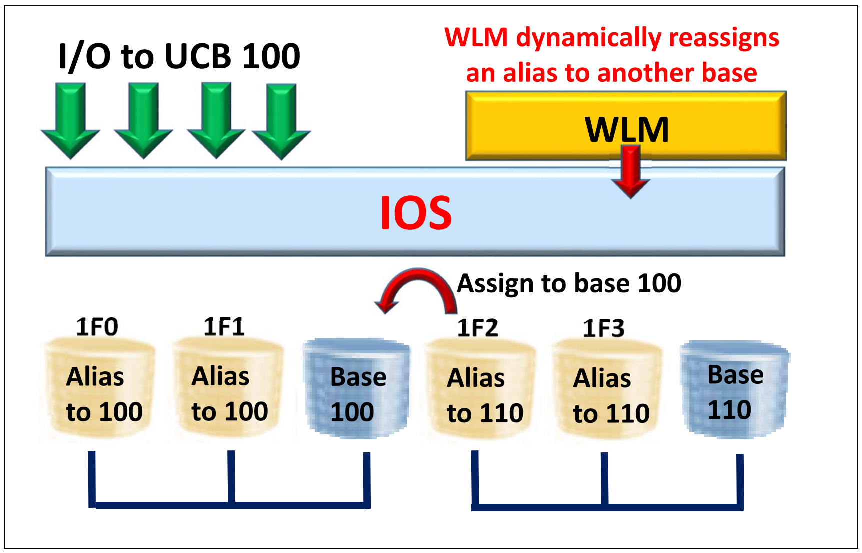


Figure 5-3 WLM assignment of alias addresses

z/OS WLM in goal mode tracks system workloads and checks whether workloads are meeting their goals as established by the installation.

WLM also tracks the devices that are used by the workloads, accumulates this information over time, and broadcasts it to the other systems in the same sysplex.

If WLM determines that any workload is not meeting its goal because of IOSQ time, WLM attempts to find an alias device that can be reallocated to help this workload achieve its goal.

As Figure 5-4 shows, WLM checks the IOSQ time of volume 100 and allocates free aliases from volume 110.

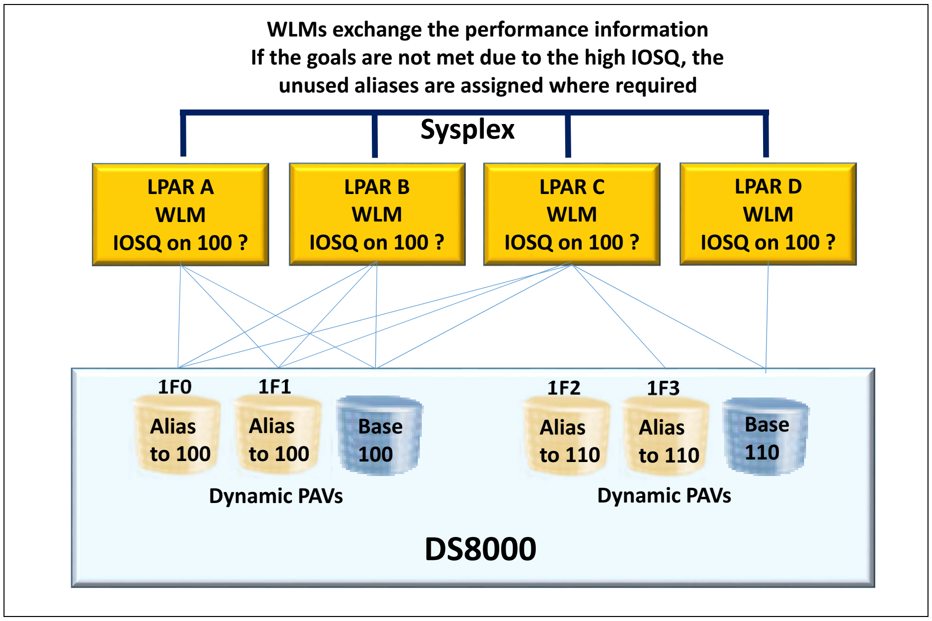


Figure 5-4 Dynamic PAVs in a sysplex

5.1.2 HyperPAV

Dynamic PAV requires the WLM to monitor the workload and goals. The process of the WLM detecting an I/O bottleneck and then coordinating the reassignment of alias addresses within the sysplex and the DS8000 storage system can take time. In addition, if the workload is fluctuating or has bursts, the job that caused the overload of one volume might end before the WLM reacts. In these cases, the IOSQ time was not eliminated.

With HyperPAV, an on-demand proactive assignment of aliases is possible, as shown in Figure 5-5.

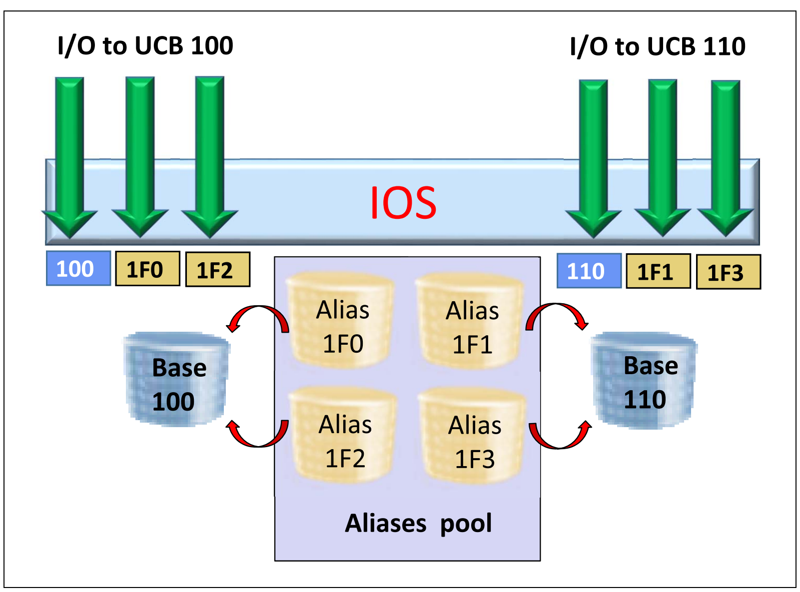


Figure 5-5 Basic operational characteristics of HyperPAV

With HyperPAV, the WLM is no longer involved in managing alias addresses. For each I/O, an alias address can be automatically picked from a pool of alias addresses within the same logical control unit (LCU).

This capability also allows multiple HyperPAV hosts to use one alias to access different bases, which reduce the number of alias addresses that are required to support a set of bases in an IBM Z environment, with no latency in assigning an alias to a base. This function is also designed to enable applications to achieve better performance than was possible with the original PAV feature alone, and the same or fewer operating system resources are used.

Benefits of HyperPAV

HyperPAV offers the following benefits:

- Provides a more efficient PAV function.
- Helps with the implementation of larger volumes because I/O rates per device can be scaled up without needing more PAV alias definitions.
- Can reduce the number of PAV aliases that are needed, which affects the 64-K device limitation less and leaves more devices for capacity use.
- Enables a more dynamic response for changing workloads.
- Simplifies alias management.

HyperPAV alias considerations on an extended address volume

HyperPAV provides a far more agile alias management algorithm because aliases are dynamically bound to a base during the I/O for the z/OS image that issued the I/O. When I/O completes, the alias is returned to the pool in the LCU, and then it becomes available to subsequent I/Os.

The general rule is that the number of aliases that are required can be estimated by the peak of the following multiplication:

```
I/O rate x Average response time = Number of aliases required
```

For example, if the peak of the calculation that occurs when the I/O rate is 2000 I/Os per second (IOPS) and the average response time is 4 ms (which is 0.004 sec), the result of the calculation is as follows:

```
2000 IOPS x 0.004 sec/I/O = 8
```

This result means that the average number of I/O operations that are running at one time for that LCU during the peak period is eight. Therefore, eight aliases should handle the peak I/O rate for that LCU. However, because this calculation is based on the average during the IBM Resource Measurement Facility (RMF) period, multiply the result by two to accommodate higher peaks within that RMF interval. Hence, in this case, the advised number of aliases is 16 (2 x 8 = 16).

A more precise approach to know how many PAVs are needed is to look at the RMF I/O Queuing Activity LCU report. The report shows the following values:

- HyperPAV Wait Ratio: Ratio of the number of times an I/O did not start and the total number of I/O requests
- HyperPAV Maximum: Maximum number of concurrently in-use HyperPAV aliases

If a value exists for the HyperPAV Wait Ratio, more PAVs are needed. If no waiting (no value) exists, the maximum is the number of PAVs that is needed. Those values must be monitored and evaluated over time, looking for peaks and comparing values for various logical partitions (LPARs).

Depending on the workload, a large reduction in PAV-alias UCBs with HyperPAV occurs. The combination of HyperPAV and extended address volume (EAV) allows you to reduce the constraint on the 64-K device address limit and in turn increase the amount of addressable storage that is available on z/OS. With multiple-subchannel sets (MSSs) on IBM Z, even more flexibility is available in the device configuration.

5.1.3 RMF reporting on PAV

RMF reports the number of exposures for each device in the following reports:

- Monitor/Direct Access Storage Device (DASD) Activity report
- Monitor II and Monitor III Device reports

If the device is a HyperPAV base device, the number is followed by the letter “H” (for example, 5.4H). This value is the average number of HyperPAV volumes (base and alias) in that interval. RMF reports all I/O activity against the base address, not by the individual base and associated aliases. The performance information for the base includes all base and alias I/O activity.

PAV and HyperPAV help minimize the IOSQ Time. You still see IOSQ Time for one of the following reasons:

- More aliases are required to handle the I/O load when compared to the number of aliases that are defined in the LCU.
- A Device Reserve is issued against the volume. A Device Reserve makes the volume unavailable to the next I/O, which causes the next I/O to be queued. This delay is recorded as IOSQ Time.

5.1.4 PAV and HyperPAV in z/VM environments

z/VM provides PAV and HyperPAV support in the following ways:

- As traditionally supported for virtual machine (VM) guests as dedicated guests by using the **`CP ATTACH`** command or **`DEDICATE`** user directory statement.
- z/VM supports PAV minidisks:
  - The base and its aliases can be dedicated to only one guest.
  - The base must be dedicated first, and then all required aliases devices.

PAV on a z/VM environment provides linkable minidisks for guests that use PAV (that is, z/OS and Linux), as shown in Figure 5-6.

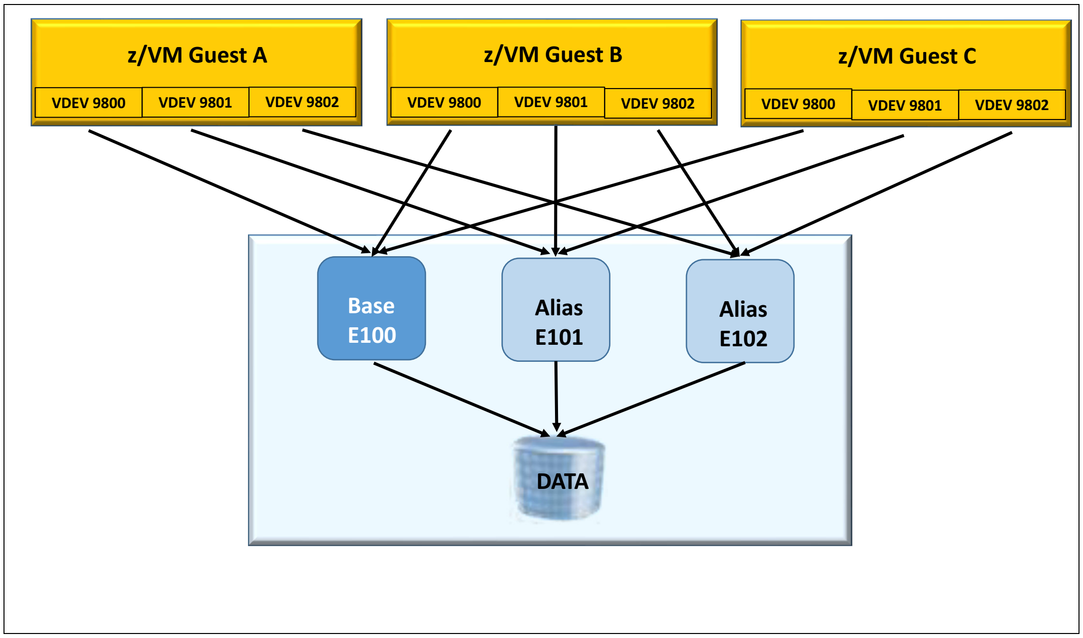


Figure 5-6 z/VM support of PAV minidisk

Base minidisks are defined by using the existing **MDISK** or **LINK** user directory statements. Aliases are defined with the **PAVALIAS** parameter of the **DASDOPT** and **MINIOPT** user directory statements or with the **CP DEFINE PAVALIAS** command. z/VM also provides workload balancing for guests that do not use PAV (such as Conversational Monitor System [CMS]). The real I/O dispatcher queues the minidisk I/O across system-attached aliases.

To the z/VM environments, PAV provides the benefit of a greater I/O performance (throughput) by reducing I/O queuing.

Starting with z/VM V5.4, z/VM supports HyperPAV for dedicated DASD and minidisks.

For more information about PAV and z/VM, see *IBM System Storage DS8000: Host Attachment and Interoperability*, SG24-8887.

5.1.5 SuperPAV

SuperPAV was introduced with DS8000 R8.1. DS8000 SuperPAV technology takes PAV technology one step further: With PAV, the base-to-alias bindings are static. HyperPAV allows dynamic alias-to-base bindings to occur, but only from within one LCU/logical subsystem (LSS).

SuperPAV is an extension of HyperPAV in the sense that it uses aliases in an on-demand fashion and allows them to be shared among similar CUs across a set of CU images that are defined with the same set of host channels in the storage system; that is, *even* LCUs go with other *even* LCUs; *odd* LCUs go with other *odd* LCUs.

Aliases are first selected from the home CU. When no more aliases are available, they are borrowed from a peer CU. Although the total LCU addresses are still limited to 256, you eventually have more than 256 addresses with SuperPAV because of the SuperPAV internal borrowing process.

A volume 2002 that borrows an alias from another LCU is shown in Figure 5-7.

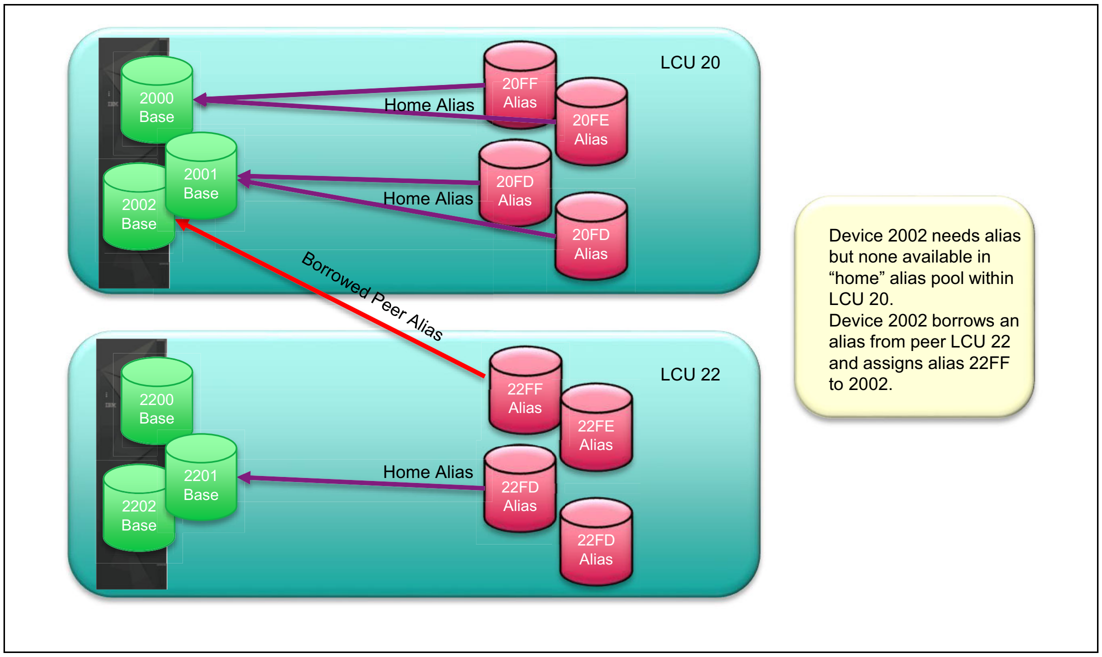


Figure 5-7 SuperPAV in the DS8000 storage system

SuperPAV allows the z/OS operating system to amortize the use of PAV-alias devices across a larger set of resources, which effectively eliminates PAV-alias exhaustion occurrences and the IOS queue time that can cause I/O response time to grow.

The following advantages also are realized:

- SuperPAV complements the thin provisioning capability of the DS8000 by allowing the growth in physical data to also grow in I/O rates without the client needing to redefine the I/O configuration, move data between CUs, or add hardware resources. This process is accomplished through enhanced virtualization techniques and allowing PAV-alias devices to be shared across CUs in the same storage system.
- SuperPAV is autonomically managed by z/OS. The client enables the function by using a setting in `SYS1.PARMLIB` and the IOS of z/OS dynamically discovers all the shared resources and auto-configures the system for their use.
- The WLM workload management for I/O ensures that whenever an I/O request finishes for a PAV-alias device, the next I/O started is the highest priority request for all CU images that share a pool of PAV-aliases.
- With the increased number of PAV-aliases that are available on average, any workload spikes can be more easily processed with SuperPAV. Also, if any hardware failures and associated I/O recovery processes in the storage area network (SAN) that can delay production work occur, the increased number of PAV-aliases that are available to process the backlog can reduce the mean-time-to-recovery (MTTR) and mitigate the effect of the failure. Thus, system resilience is improved.

Running the **D M=DEV** command shows **XPAV**, which is the new SuperPAV function when it is enabled for a device. In the **D M=DEV** output result, find the line starting with “CU NUMBER” and use the first CU on the command **D M=CU(4E02)**. A device that is named 4E02 with peers from other CUs is shown in Example 5-1.

Example 5-1 SuperPAV peers example: XPAV-enabled

```
D M=CU(4E02)
IEE174I 11.26.55 DISPLAY M 511
CONTROL UNIT 4E02
CHP                   65   5D   34   5E
ENTRY LINK ADDRESS    98   ..   434B ..
DEST LINK ADDRESS     FA   0D   200F 0D
CHP PHYSICALLY ONLINE Y    Y    Y    Y
PATH VALIDATED        Y    Y    Y    Y
MANAGED               N    N    N    N
ZHPF - CHPID          Y    Y    Y    Y
ZHPF - CU INTERFACE   Y    Y    Y    Y
MAXIMUM MANAGED CHPID(S) ALLOWED = 0
DESTINATION CU LOGICAL ADDRESS = 56 (see note below)
CU ND           = 002107.998.IBM.75.0000000LBN71.0231
CU NED          = 002107.998.IBM.75.0000000LBN71.9100
TOKEN NED       = 002107.900.IBM.75.0000000LBN71.9100
FUNCTIONS ENABLED = ZHPF, XPAV
XPAV CU PEERS     = 4802, 4A02, 4C02, 4E02
DEFINED DEVICES
  04E00-04E07
DEFINED PAV ALIASES
  14E40-14E47
```

To enable SuperPAV, use Release 8.1 (88.11 bundles) or later code levels. For z/OS (2.1+), APARs OA49090 and OA49110 are needed, and for RMF (2.1+), APAR OA49415 is needed.

In SYS1.PARMLIB, set HYPERPAV=YES and HYPERPAV=XPAV in **IECIOSxx**, or set the **SETIOS HYPERPAV=YES and SETIOS HYPERPAV=XPAV** commands. Specifying **YES** *and* **XPAV** in a window where SuperPAV support is not available on all storage systems ensures that at least HyperPAV is in effect.

Note: In Example 5-1, the marker refers to the real LCU/LSS that is inside the DS8000.

New sections of the RMF I/O activity report

Among several new fields, the following important fields in RMF reports can help you investigate performance problems:

- Alias Management Groups (AMGs)

  For each defined AMG, this field shows performance measurements for all channel paths that are connected to the LCUs that are grouped into the AMG.

- LCUs

  For each LCU with online devices, this field shows performance measurements for all channel paths that are connected to the LCU.

- `HPAV WAIT` and `HPAV MAX`:
  - `HPAV WAIT` is the ratio of the number of I/O requests that did not start because no HyperPAV aliases were available.
  - `HPAV MAX` is the maximum number of concurrently used HyperPAV alias devices (including borrowed aliases) for that LCU or AMG during that interval.

Many other fields were updated. For more information, see *z/OS 2.5 Resource Measurement Facility Report Analysis*, SC34-2665.

Part of the I/O Queuing Activity is shown in Figure 5-8.

```
                                  I/O   Q U E U I N G   A C T I V I T Y
                                                                                                 PAGE    1

-------------------------------------------------------------------------------------------------------------
                                       ALIAS MANAGEMENT GROUPS
-------------------------------------------------------------------------------------------------------------

                                                AVG   AVG                  DELAY  AVG              AVG   DATA
AMG          DCM GROUP  CHAN    CHPID   % DP  % CU  CUB   CMR CONTENTION     Q    CSS     HPAV     OPEN  XFER
            MIN MAX DEF PATHS   TAKEN   BUSY  BUSY  DLY   DLY    RATE      LNGTH  DLY  WAIT   MAX  EXCH  CONC
                                                          ...
00000016                65      0.537   0.00  0.00  0.0   3.6
                        34      0.538   0.00  0.00  0.0   4.5
                        99      0.532   0.00  0.00  0.0   1.3
                        *       1.607   0.00  0.00  0.0   3.1    0.000      0.00  0.8  0.000    0
00000017                65      0.361   0.00  0.00  0.0   2.1
                        34      0.356   0.00  0.00  0.0   3.8
                        99      0.357   0.00  0.00  0.0   1.2
                        *       1.073   0.00  0.00  0.0   2.4    0.000      0.00  0.7  0.000    0
                                                          ...

-------------------------------------------------------------------------------------------------------------
                                       LOGICAL CONTROL UNITS
-------------------------------------------------------------------------------------------------------------

                                                     AVG   AVG                  DELAY  AVG              AVG   DATA
LCU/    CU     DCM GROUP  CHAN    CHPID   % DP  % CU  CUB   CMR CONTENTION     Q    CSS     HPAV     OPEN  XFER
AMG           MIN MAX DEF PATHS   TAKEN   BUSY  BUSY  DLY   DLY    RATE      LNGTH  DLY  WAIT   MAX  EXCH  CONC
                                                          ...
0049    5100              43      0.026   0.00  0.00  0.0   0.0
                          55      0.024   0.00  0.00  0.0   0.0
                          44      0.024   0.00  0.00  0.0   0.0
                          56      0.026   0.00  0.00  0.0   0.0
                          *       0.100   0.00  0.00  0.0   0.0    0.000      0.00  0.1  0.000    0   0.00  0.00
                                                          ...
0107    B101              65      0.084   0.00  0.00  0.0   3.6
 00000016                 34      0.083   0.00  0.00  0.0   4.7
                          99      0.082   0.00  0.00  0.0   1.3
                          *       0.250   0.00  0.00  0.0   3.2    0.000      0.00  0.8  0.000    0
                                                          ...
011A    CF01              65      0.072   0.00  0.00  0.0   3.8
 00000016                 34      0.073   0.00  0.00  0.0   4.9
                          99      0.071   0.00  0.00  0.0   1.2
                          *       0.217   0.00  0.00  0.0   3.4    0.000      0.00  0.7  0.000    0
0106    B001              65      0.084   0.00  0.00  0.0   2.1
 00000017                 34      0.082   0.00  0.00  0.0   4.1
                          99      0.083   0.00  0.00  0.0   1.2
                          *       0.250   0.00  0.00  0.0   2.5    0.000      0.00  0.8  0.000    0
```

Figure 5-8 Snippet from I/O Queuing Activity

5.2 Multiple allegiances

If any IBM Z host image (server or LPAR) performs an I/O request to a device address for which the storage disk system is processing an I/O that is from another IBM Z host image, the storage disk system sends back a device busy indication. This process delays the new request and adds to the overall response time of the I/O. This delay is shown in the Device Busy Delay (`AVG DB DLY`) column in the RMF DASD Activity Report. Device Busy Delay is part of the Pend time.

With multiple allegiances, the requests are accepted by the DS8000 storage system and all requests are processed in parallel, unless a conflict occurs when writing to the same data portion of the CKD logical volume, as shown in Figure 5-9.

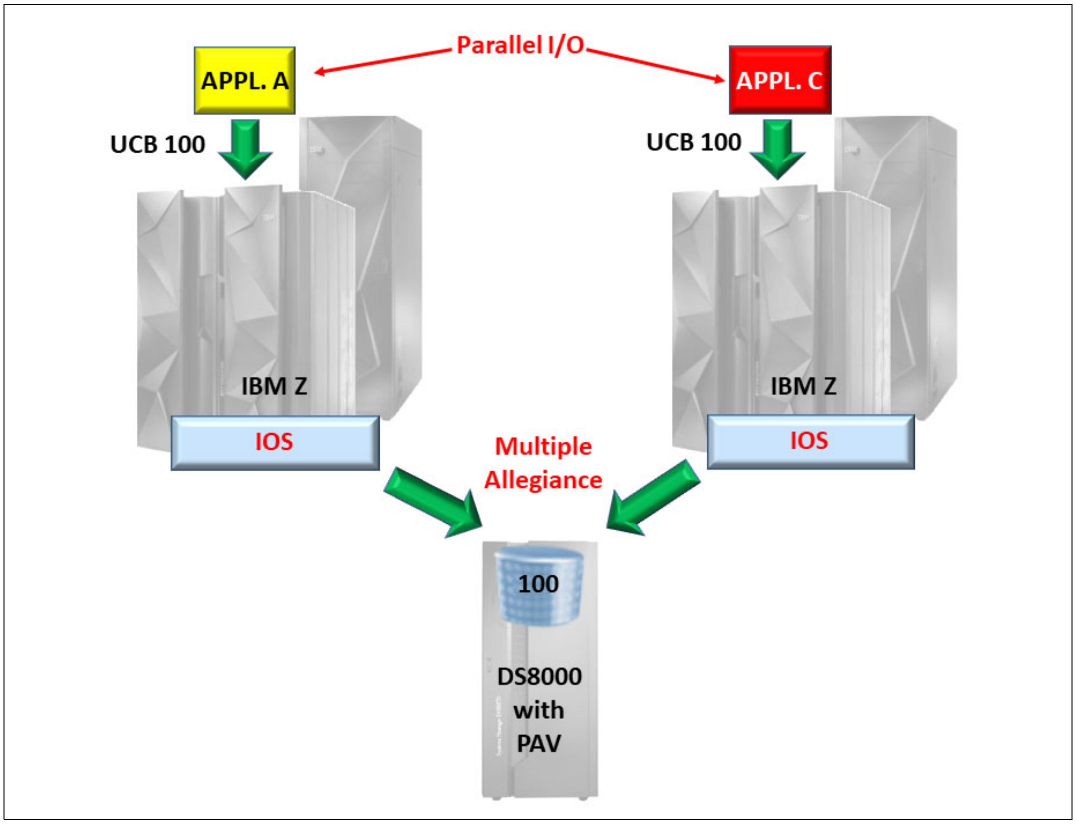


Figure 5-9 Parallel I/O capability with multiple allegiances

The DS8000 storage system accepts multiple I/O requests from different hosts to the same device address, which increases parallelism and reduces the effect on channels. In older storage disk systems, a device has an implicit allegiance (that is, a relationship) that was created in the CU between the device and a channel path group when the device accepts an I/O operation. This allegiance causes the CU to ensure access (no busy status presented) to the device for the remainder of the channel program over the set of paths that are associated with the allegiance.

Good application software access patterns can improve global parallelism by avoiding reserves, limiting the extent scope to a minimum, and setting a suitable file mask (for example, if no write is intended).

In systems without multiple allegiances (except for the first I/O request), all requests to a shared volume are rejected, and the I/Os are queued in the IBM Z channel subsystem. The requests are listed in Device Busy Delay and PEND time in the RMF DASD Activity reports.

Multiple allegiances allow multiple I/Os to a single volume to be serviced concurrently. However, a device busy condition can still happen. This condition occurs when an active I/O is writing a specific data portion on the volume and another I/O request is received and attempts to read or write to that same data. To ensure data integrity, those subsequent I/Os receive a busy condition until that previous I/O is finished with the write operation.

Multiple allegiances provide significant benefits for environments that are running a sysplex or IBM Z servers that are sharing access to data volumes. Multiple allegiances and PAV can operate together to handle multiple requests from multiple hosts.

5.3 Modified Indirect Data Access Word facility

The Modified Indirect Data Access Word (MIDAW) facility was designed to improve Fibre Channel connection (FICON) performance, especially when accessing Db2 on z/OS. This facility offers a method of gathering data into and scattering data from fragmented storage locations during an I/O operation.

The MIDAW facility achieves superior performance for various workloads by improving the throughput and efficiency of the channel subsystem. Although the usage of MIDAWs does not cause the bits to move any faster across the FICON link, they reduce the number of frames and sequences flowing across the link, which makes the channel more efficient.

Because MIDAWs are used only by the Media Manager, and MIDAWs benefit only small record sizes, only specific certain types of data sets are beneficiaries. Some examples of data sets that are accessed through Media Manager are Virtual Storage Access Method (VSAM) data sets (including all linear data sets [LDSs]), Extended Format data sets, and PDSEs. The most benefit occurs with Extended Format data sets that feature small block sizes. Because Db2 depends on Extended Format data sets to stripe the logs or to enable data sets to be larger than 4 GB, Db2 is a major beneficiary.

The DS8000 storage system provides MIDAW support. The MIDAW facility is enabled on z/OS by default. To verify whether the MIDAW facility is enabled, run the following command:

```
DISPLAY IOS,MIDAW
```

Figure 5-10 shows the output of the command.

```
D IOS,MIDAW
IOS097I 19.53.16 MIDAW FACILITY 666
MIDAW FACILITY IS ENABLED
```

Figure 5-10 D IOS,MIDAW output

If the facility is unavailable, update the **IECIOSxx** member with `MIDAW=YES`.

5.4 A caching algorithm that is optimized for IBM Z

One main differentiator among available enterprise storage systems today is the internal cache and its algorithm. Cache size and its utilization efficiency are important factors to consider when sizing the storage to meet a client's performance requirements.

With the DS8000 storage system and its powerful IBM POWER® processors, managing a large cache with small cache slots of 4 KB is possible. Disk systems generally divide a cache into fixed size slots. A slot (sometimes known as *segment* or *cache page*) is used to hold contiguous data, so randomly read or written blocks are assigned different slots.

The more random the I/Os and the smaller the block sizes, the more cache that is wasted because of large slot sizes. Therefore, the DS8000 small cache slots are the main contributing factor regarding the efficient cache utilization.

The small cache slots are served by sophisticated caching algorithms, which are another significant advantage of the DS8000 storage system from a performance perspective. These algorithms, along with the small cache slot size, optimize cache hits and cache utilization. Cache hits are also optimized for different workloads, such as sequential workloads and transaction-oriented random workloads, which can be active at the same time. Therefore, the DS8900 storage system provides excellent I/O response times.

The following caching algorithms are used in DS8000 storage systems:

- Sequential Prefetching in Adaptive Replacement Cache (SARC)

  The SARC is a self-tuning and self-optimizing solution for a wide range of workloads with a varying mix of sequential and random I/O streams. SARC is inspired by the Adaptive Replacement Cache (ARC) algorithm and inherits many of the ARC algorithm's features.

  SARC attempts to determine the following cache characteristics:

  – When and which data is copied into the cache.
  – Which data is evicted when the cache becomes full.
  – How the algorithm dynamically adapts to different workloads.

  The decision to copy data into the DS8000 cache can be triggered by the following policies:

  – Demand paging

    Eight disk blocks (a 4 K cache page) are brought in only on a cache miss. Demand paging is always active for all volumes and ensures that I/O patterns with some locality discover at least recently used data in the cache.

  – Prefetching

    Data is copied into the cache speculatively, even before it is requested. To prefetch, a prediction of likely data accesses is needed. Because effective, sophisticated prediction schemes need an extensive history of page accesses (which is not feasible in real systems), SARC uses prefetching for sequential workloads.

    Sequential access patterns naturally arise in many IBM Z workloads, such as database scans, copy, backup, and recovery. The goal of sequential prefetching is to detect sequential access and prefetch the likely cache data to minimize cache misses.

- Adaptive multi-stream prefetching (AMP)

  AMP is an autonomic, workload-responsive, and self-optimizing prefetching technology that adapts the amount of prefetch and the timing of prefetch on a per-application basis to maximize the performance of the system. The AMP algorithm solves the following problems that plague most other prefetching algorithms:

  – *Prefetch wastage* occurs when prefetched data is evicted from the cache before it can be used.

  – *Cache pollution* occurs when less useful data is prefetched.

  By wisely choosing the prefetching parameters, AMP provides optimal sequential read performance and maximizes the aggregate sequential read throughput of the system. The timing of the prefetches is also continuously adapted for each stream to avoid misses and any cache pollution. SARC and AMP play complementary roles.

- Intelligent Write Caching (IWC)

  IWC is another cache algorithm that is implemented in the DS8000 series. IWC improves performance through better write cache management and a better destaging order of writes. This algorithm is a combination of CLOCK, which is a predominantly read cache algorithm, and CSCAN, which is an efficient write cache algorithm. From this combination, IBM produced a powerful and widely applicable write cache algorithm.

- Adaptive List Prefetch (ALP)

  ALP enables prefetch of a list of nonsequential tracks, which provides improved performance for Db2 workloads.

IBM Z workloads, in particular z/OS applications (such as Db2), are modified to provide various hints to the DS8000 storage system on sequential processing in addition to database I/O operations. Optimization of what data is in the DS8000 cache at any point enables clients to optimize the usage of cache by improving cache hits and I/O response time.

For more information about DS8000 cache algorithms, see *IBM DS8900F Architecture and Implementation: Updated for Release 9.2*, SG24-8456.

5.5 High-Performance FICON for IBM Z

High-Performance FICON for IBM Z (zHPF) is an enhanced FICON protocol and system I/O architecture that results in improvements in response time and throughput. Instead of channel command words (CCWs), transport control words (TCWs) are used. Any I/O that uses the Media Manager, such as Db2, partitioned data set extended (PDSE), VSAM, z/OS File System (zFS), volume table copy (VTOC) Index (Common VTOC Access Facility (CVAF)), catalog basic catalog structure (BCS) / VSAM volume data set (VVDS), or Extended Format SAM, benefit from zHPF.

The FICON data transfer protocol involves several exchanges between the channel and the CU, which can lead to unnecessary I/O impact. With zHPF, the protocol is streamlined and the number of exchanges is reduced, as shown Figure 5-11.

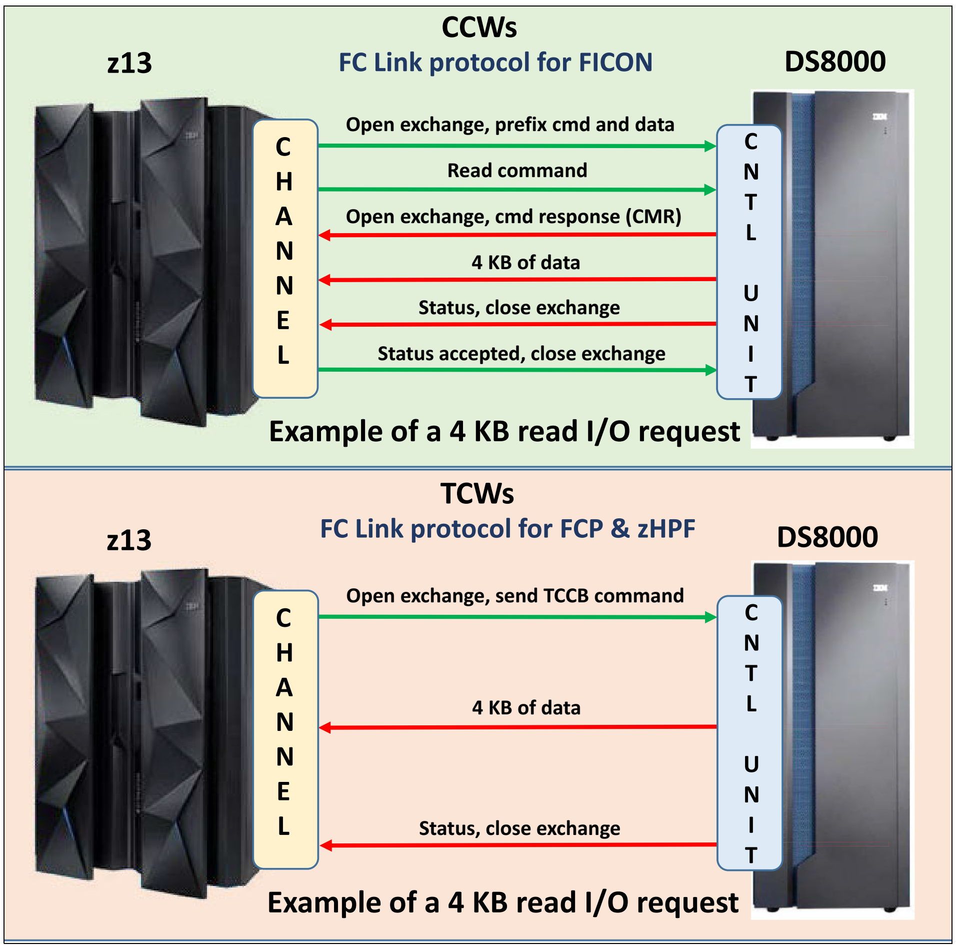


Figure 5-11 FICON and zHPF comparison

zHPF was enhanced following its introduction in 2009, as shown in Figure 5-12. Many access methods were changed in z/OS to support zHPF.

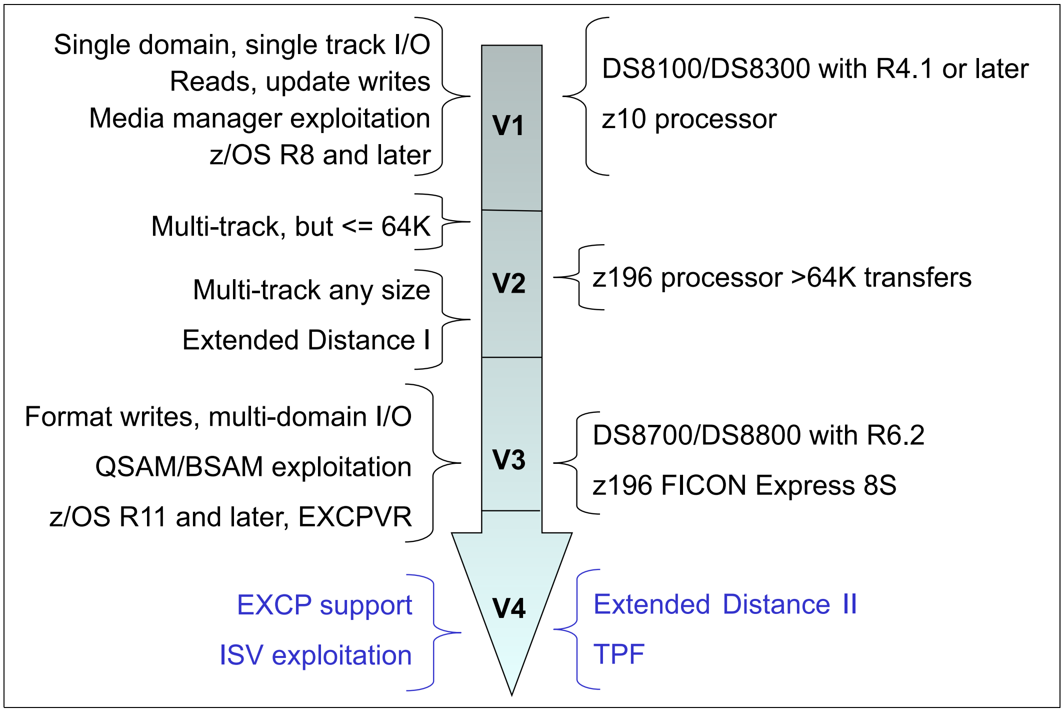


Figure 5-12 The evolution of zHPF

Although the original zHPF implementation supported the new TCWs only for I/O that did not span more than a track, the DS8000 storage system also supports TCW for I/O operations on multiple tracks. zHPF is also supported for Db2 list prefetch, format writes, and sequential access methods. With the latest zHPF version, a typical z/OS workload has 90% or more of all I/Os converted to the zHPF protocol, which improves the channel utilization efficiency.

In situations where zHPF is the exclusive access that is used, it can improve FICON I/O throughput on a single DS8000 port by 250 - 280%. Realistic workloads with a mix of data set transfer sizes can see over a 90% increase in FICON I/Os that use zHPF. These numbers can vary based on the workload and were seen in real client environments, but are not a rule. They are used as a guideline and can have different metrics in your environment.

Although clients can see a fast completion of I/Os as a result of implementing zHPF, the real benefit is the use of fewer channels to support disk volumes or increasing the number of disk volumes that is supported by channels.

In addition, the changes in architecture offer end-to-end system enhancements to improve reliability, availability, and serviceability (RAS).

The IBM Z generations since IBM z10 support zHPF.

zHPF is not apparent to applications. However, z/OS configuration changes are required. The hardware configuration definition (HCD) must have channel-path identifier (CHPID) type FC defined for all the CHPIDs that are defined to the 2107 CU, which also supports zHPF.

The installation of the Licensed Feature Key for the zHPF feature was required for the DS8870 storage system. With later DS8000 models, all zSynergy features are bundled and come together in the zSynergy Services (zsS) license bundle. After these zHPF prerequisites are met, the FICON port definitions in the DS8000 storage system accept FICON and zHPF protocols. No other port settings are required on the DS8000 storage system.

For z/OS, after the PTFs are installed in the LPAR, you must set `ZHPF=YES` in **IECIOSxx** in `SYS1.PARMLIB` or issue the **SETIOS ZHPF=YES** command (`ZHPF=NO` is the default setting), as shown in Figure 5-13.

```
D IOS,ZHPF
IOS630I 12.01.56 ZHPF FACILITY 043
HIGH PERFORMANCE FICON FACILITY IS ENABLED
```

Figure 5-13 D IOS,ZHPF output result

To use zHPF for the queried sequential access method (QSAM), basic sequential access method (BSAM), and basic partitioned access method (BPAM), you might need to enable zHPF. It can be dynamically enabled by using **SETSMS** or by using the entry `SAM_USE_HPF(YES | NO)` in **IGDSMSxx**. The default is `YES`. Figure 5-14 shows the output.

```
D SMS,OPTIONS
IGD002I 12:04:33 DISPLAY SMS 046
...
ACDS LEVEL = z/OS V2.4
SMS PARMLIB MEMBER NAME = IGDSMS00
INTERVAL = 5           DINTERVAL = 150
SMF_TIME = YES          CACHETIME = 3600
...
OAMTASK =            PDSE_SYSEVENT_DONTSWAP = NO
DB2SSID =                 SAM_USE_HPF = YES
...
```

Figure 5-14 D SMS,OPTIONS output result

5.5.1 DFSORT using IBM Integrated Accelerator for Z Sort

DFSORT can also use zHPF for SORTIN, SORTOUT, and OUTFIL data sets and the new instruction **SORTL**. A new special hardware is available that is called IBM Integrated Accelerator for Z Sort that speeds up frequently used functions by using sort in-memory. One accelerator is used per core, which is standard on IBM z15. For this use, zHPF is required and must be enabled.

Based on IBM internal tests on dedicated environments for IBM Integrated Accelerator for Z Sort, jobs can reduce the elapsed time by up 30% and CPU time by up 40% for data sets with record lengths up to 500 bytes.

For more information about DFSORT using zHPF, see *DFSORT User Guide: IBM Integrated Accelerator for Z Sort (PH03207) on zOS V2R3 and V2R4 PTFs UI90067 and UI90068*.

Check your system for DFSORT APARs PH03207 and PH28183 on z/OS V2R3 and later.

5.5.2 Db2 enhancements with zHPF

All Db2 I/Os, including format writes and list prefetches, are eligible for zHPF. Db2 can benefit from the DS8000 caching algorithm that is called *List Prefetch Optimizer*, which provides even better performance. Although the objective of zHPF list prefetch is to reduce the I/O connect time, the objective of List Prefetch Optimizer is to reduce disconnect time.

Db2 typically performs a maximum of two parallel list prefetch I/Os. With List Prefetch Optimizer, zHPF sends a list of tracks in a `Locate Record` command, which is followed by several read commands, which allows many parallel retrievals from disk. Also, the use of the `Locate Record` command now sends a list of noncontiguous records to prefetch, which increases the cache-hit rate for subsequent read commands.

Performance can vary depending on the hardware configuration (FICON Express cards 8, 16, and 32 Gb) and Db2 version. Because Db2 10 and later fully uses zHPF along with the FICON Express 8S/16S/16S+/16SA/32S on IBM Z and the DS8000 storage system, the following Db2 functions are improved:

- Db2 queries
- Table scans
- Index-to-data access, especially when the index cluster ratio is low
- Index scans, especially when the index is disorganized
- Reads of fragmented large objects (LOBs)
- New extent allocation during inserts
- Db2 `REORG`
- Sequential reads
- Writes to the shadow objects
- Reads from a nonpartitioned index
- Log applies
- Db2 `LOAD` and `REBUILD`
- Db2 Incremental `COPY`
- `RECOVER` and `RESTORE`
- Db2 `RUNSTATS` table sampling

Db2 for z/OS and DS8000 zHPF synergy provides significant throughput gains in many areas, which result in reduced transaction response time and batch windows.

For more information about Db2 and List Prefetch Optimizer, see *DB2 for z/OS and List Prefetch Optimizer*, REDP-4862.

> **Note:** By making all Db2 workload zHPF capable, users can benefit from a reduced batch window for I/O-intensive workloads and maximize resource utilization.
>
> For more information, see Db2 Utilities for z/OS APAR PH28183 on your system.

Db2 Castout Accelerator

In Db2, a *castout* refers to the process of writing pages from the group buffer pool to disk. Db2 writes long chains that typically contain multiple locate record domains. Traditionally, each I/O in the chain is synchronized individually, but Db2 requires that only the updates are written in order.

With DS8000, the Media Manager is enhanced to signal to the DS8000 storage system that a single locate record domain exists, even though multiple embedded locate records exist. The entire I/O chain is treated as though this domain is a single locate record domain.

This change was implemented by z/OS Media Manager support and APAR OA49684 for z/OS V1.13 and later. The function is not apparent to Db2 and is used by all Db2 releases.

Typical castout write operations can be accelerated with this function by a considerable degree, and especially when in a Metro Mirror (MM) configuration. For more information about the performance results of this function, see *IBM System Storage DS8880 Performance Whitepaper*, WP102605.

5.5.3 Extended Distance and Extended Distance II

Db2 utilities often use significant write I/Os with a large blocksize or large record sizes. Values of 256 KB or even 512 KB per I/O are not unusual. These kinds of heavy write I/Os can be handled well by using traditional FICON channel I/O, even over a long distance of up to 100 km (62.13 miles).

As initially introduced, zHPF Extended Distance (ED) does not provide results that are as good for larger writes over greater distances than conventional FICON I/O. This situation is corrected by Extended Distance II (ED II) support that were available since IBM z13® and DS8000 firmware releases 7.5 and later.

Such a configuration with MM over a greater distance and with HyperSwap active is shown in Figure 5-15. After a HyperSwap swap and switching to the distant volumes occurs, the host I/Os service time can degrade and negatively affect the Db2 subsystem.

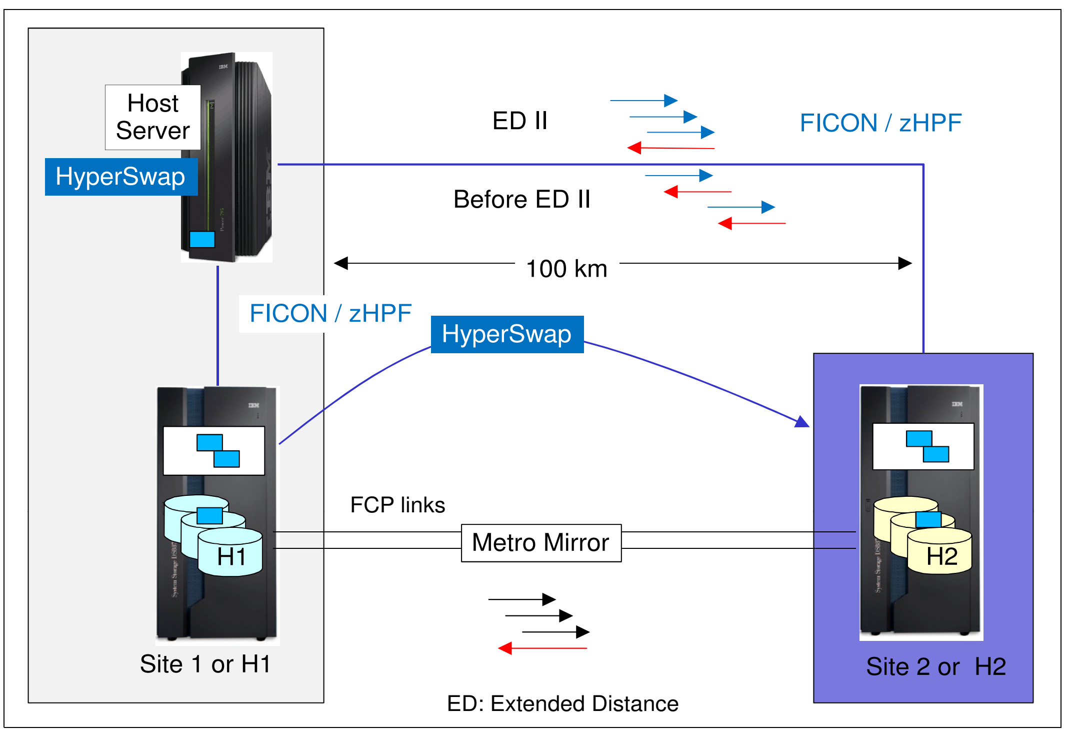


Figure 5-15 Extended Distance II support between IBM Z and the DS8000 storage system

ED II is an enhancement to zHPF in z13 and later. ED II addresses these situations of zHPF for large write I/Os over an extended distance.

Similar considerations are applied to MM over Fibre Channel Protocol (FCP). To allow for increased distances of up to 300 km (186.4 miles), MM and its variations introduced the concept of *predeposit writes*, which reduce the number of round trips of standard FCP I/Os to a single round trip.

Although zHPF ED uses the concept of predeposit writes, the benefits are limited to writes that are less than 64 KB. zHPF ED II goes beyond the capabilities of the FCP by allowing the channel to burst up to the whole write data length (DL) for an operation.

zHPF ED II improves large write I/O performance over longer distances between IBM Z and the DS8000 storage system with firmware releases 7.5 and later, which reinforces the synergy between IBM Z and the DS8000.

5.5.4 Enhanced FICON Information Unit pacing

The Information Unit (IU) pacing credit is the maximum number of IUs that a channel sends over a FICON exchange until it receives a command-response (CMR) IU, which allows the next send.

The FICON standard default for IU pacing credits is 16. At extended distances and greater speed links, this limitation causes relative latency for programs.

Enhanced IU pacing, available with DS8000 storage system Release 8.2 and later, uses the persistent pacing feature to change the operating pacing credit to 64. The latency is reduced by nearly four times with the following channel programs:

- CCW chains that contain more than 16 commands
- Db2 log writes in IBM Geographically Dispersed Parallel Sysplex (GDPS) multi-site workload environments

Enhanced IU pacing also benefits the FICON write performance of large writes at long distances and the IBM z/OS Global Mirror (IBM zGM) (Extended Remote Copy (XRC)) initial copy performance. This feature applies only to 16 G Fibre Channel host bus adapters.

IBM z/OS Global Mirror copy times

This component also benefits from persistent IU pacing because be more CCWs cab exist in the chain.

On the System Data Mover (SDM), the **TracksPerRead** parameter is *not* limited to 15, although it is documented as such. If the necessary resources to use the enhanced IU pacing are in place, a better value for **TracksPerRead** is 61. Those improvements rely on the network, if it can keep up. Results can vary depending on your network.

For more information, see this IBM Documentation web page. The DS8900F models are the latest generation of DS8000 to support z/OS Global Mirror.

5.6 Easy Tier

DS8000 Easy Tier technology enhances performance and balances workloads across different disk tiers. It automatically and without disruption to applications enables the system to relocate data (at the extent level) across up to three storage tiers.

Having the correct data on the correct tier to provide a remarkable (performance and cost) quality of service (QoS) to a customer's data is the main Easy Tier goal. Among IBM Z customers, 10% of T0 drives (flash and solid-state drives [SSDs]) that are managing 90% of the I/Os is common when considering environments that have predominantly random I/Os.

To move extents, some free space or free extents must be available in an extent pool.

Easy Tier I/O workload monitoring collects I/O statistics from the physical backend, which means the back-end I/O activity on the DS8000 rank level. The system monitors the stage and destage activities of each extent that is allocated to a logical volume in the extent pool. It also calculates a temperature (a measure based on I/O activity) metric for each extent, which is also referred to as a *heat map*.

5.6.1 Easy Tier application

IBM Z applications can communicate performance requirements for optimal data set placement by providing application performance information (hints) to the Easy Tier application programming (API) interface. The application hint sets the intent, and Easy Tier moves the data set extents to the correct tier.

The Easy Tier architecture for IBM Z applications that provides data placement hints to the Easy Tier API (by using Data Facility Storage Management Subsystem [DFSMS]) is shown in Figure 5-16. As of this writing, this architecture is limited to Db2 for z/OS. Db2 for z/OS uses specific DFSMS functions to communicate directive data placements to Easy Tier.

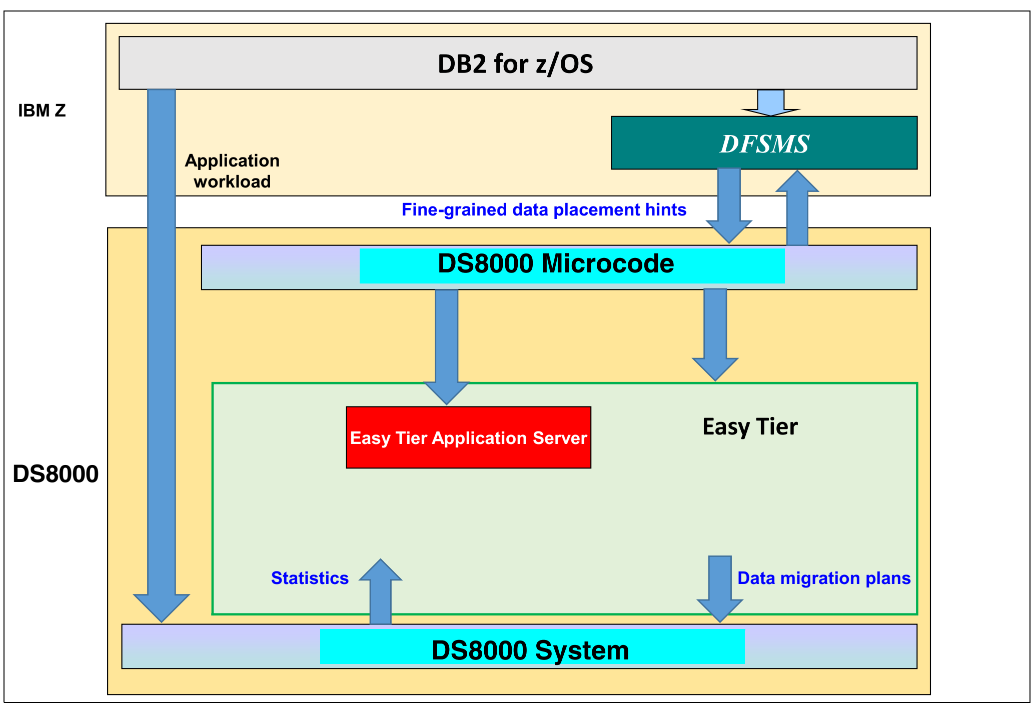


Figure 5-16 Easy Tier Application architecture

The Easy Tier Application software-defined storage (SDS) data placement API allows Db2 to proactively instruct Easy Tier about the intended usage of data sets. This capability removes the requirement for applications and administrators to manage hardware resources directly. The programming must be done only once, and then the application with Easy Tier enforces the policy.

With Db2 V10 or later (with a small programming enhancement [SPE]), Db2 can query and then set the tier location that you want for data sets that use internal DFSMS functions, which interface with the Easy Tier Application API.

5.6.2 Heat Map Transfer

Like host operations, DS8000 Copy Services (MM, Global Mirror [GM], and IBM zGM) are unaware of the extent or volume-level relocations that are performed.

Easy Tier at the primary DS8000 storage system sees a normal workload, and sees only the write workloads at the secondary DS8000 storage system. This situation means that the optimized extent distribution on the primary system can differ considerably from the one on the secondary system if the Heat Map Transfer (HMT) function is not used.

By using Easy Tier HMT, you can export the data placement statistics that are used at an MM, Global Copy, and GM primary site to reapply them at the secondary site. As such, Easy Tier HMT complements the DS8000 Copy Services Remote Mirroring functions. It also provides automatic tiering solution for high availability (HA) and disaster recovery (DR) (HA/DR) environments.

Easy Tier HMT is installed on a separate management server and can work in the following ways:

- Stand-alone server (Windows or Linux)
- Integrated with IBM Copy Services Manager (CSM)
- Integrated with GDPS

In complex, 3-site DR environments with GDPS or CSM management for a Metro/Global Mirror (MGM) configuration, the heat map is propagated to each site, as shown in Figure 5-17. In this cascaded replication configuration, the HMT utility transfers the Easy Tier heat map from H1 to H2 and then from H2 to H3 based on the volume replication relationships.

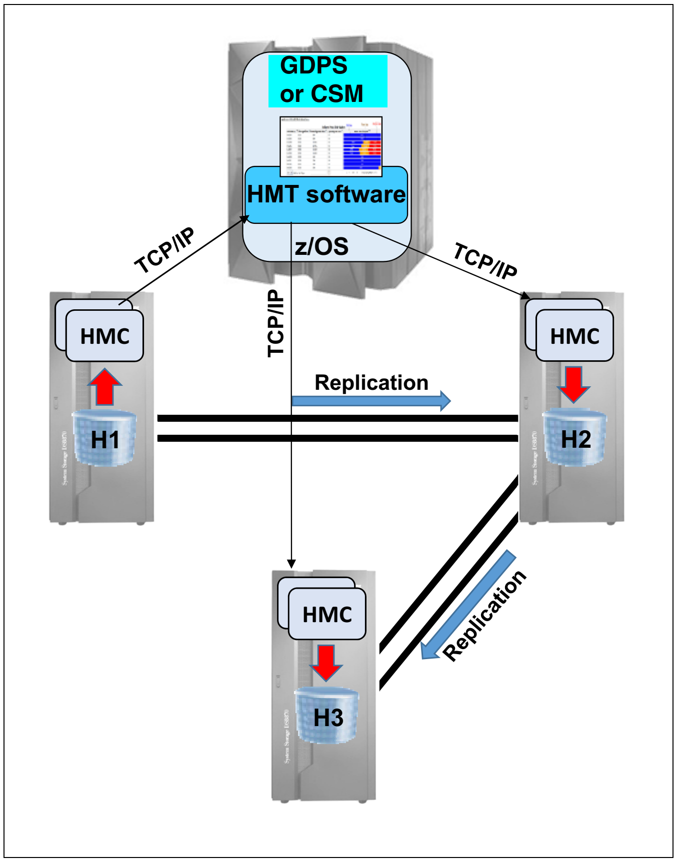


Figure 5-17 Easy Tier Heat Map Transfer support for 3-site MGM configuration

Since Version 3.12, GDPS provides HMT support for GDPS/XRC and GDPS/MzGM (MM and XRC) configurations. Therefore, Easy Tier Heat Map can be transferred to the XRC secondary or FlashCopy target devices.

With DS8000 R7.5 or later, HMT is fully supported for GDPS 3-site and 4-site MGM configurations.

Note: IBM Transparent Data Migration Facility (IBM TDMF) V5.7 and later also can use HMT to place moved data in the same storage tiers where the source data is stored.

For more information about Easy Tier, see the following publications:

- *IBM DS8000 Easy Tier*, REDP-4667
- *DS8870 Easy Tier Application*, REDP-5014
- *IBM DS8870 Easy Tier Heat Map Transfer*, REDP-5015

5.7 I/O priority queuing

The concurrent I/O capability of the DS8000 storage system means that it can run multiple channel programs concurrently if the data that is accessed by one channel program is not altered by another channel program.

5.7.1 Queuing of channel programs

When the channel programs conflict with each other and must be serialized to ensure data consistency, the DS8000 storage system internally queues channel programs. This subsystem I/O queuing capability provides the following significant benefits:

- Compared to the traditional approach of responding with a *device busy* status to an attempt to start a second I/O operation to a device, I/O queuing in the storage disk subsystem eliminates the effect that is associated with posting status indicators and redriving the queued channel programs.
- Contention in a shared environment is minimized. Channel programs that cannot run in parallel are processed in the order in which they are queued. A fast system cannot monopolize access to a volume that also is accessed from a slower system.

5.7.2 Priority queuing

I/Os from separate z/OS system images can be queued in a priority order. The z/OS WLM uses this priority to grant precedence for I/Os from one system. You can activate I/O priority queuing in the WLM Service Definition settings. WLM must run in Goal mode.

When a channel program with a higher priority is received and moved to the front of the queue of channel programs with a lower priority, the priority of the low-priority programs is increased, as shown in Figure 5-18 on page 96.

Important: Do not confuse I/O priority queuing with I/O Priority Manager (IOPM). I/O priority queuing works on a host adapter level and is available at no charge. IOPM was a licensed function of a DS8880 and was discontinued with the DS8900.

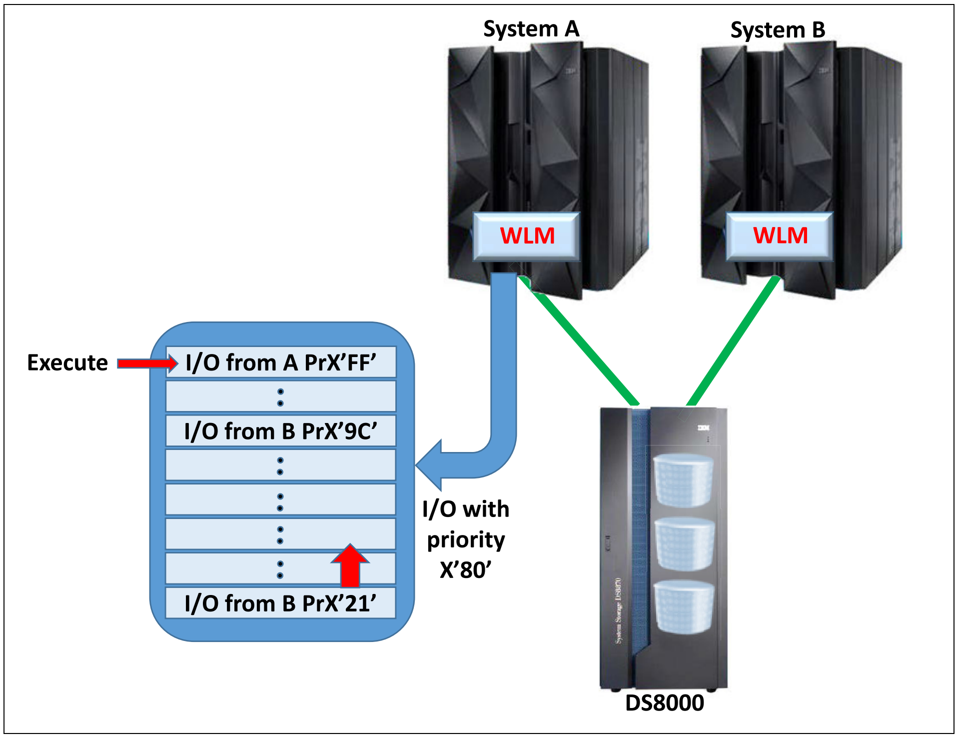


Figure 5-18 I/O priority queuing

5.8 IBM zHyperWrite

IBM zHyperWrite technology is provided by the DS8000 storage system. It is used by DFSMS to help accelerate Db2 and IBM Information Management System (IMS) log writes in MM synchronous data replication environments when GDPS or CSM and HyperSwap technology are used. IBM zHyperWrite is also supported in Multi-Target Metro Mirror (MTMM) environments and can be used in each relationship independently, depending on the state of the relationship.

When an application sends an I/O request to a volume that is in synchronous data replication, the response time is affected by the latency that is caused by the replication management, in addition to the latency because of the distance between source and target disk CUs; that is, 10 microseconds per 1 km (0.62 miles).

Although the latency cannot be avoided (we are limited by the speed of light), an opportunity exists to reduce the latency that is added by managing the synchronous relationship when application usage allows this technique to be used. IBM zHyperWrite combines concurrent DS8000 MM (Peer-to-Peer Remote Copy [PPRC]) synchronous replication and software mirroring through Media Manager (DFSMS) to provide substantial improvements in Db2 and IMS log write latency.

With IBM zHyperWrite, I/O writes to the database logs are replicated synchronously to the secondary volumes by DFSMS. For each write to the primary log volume, DFSMS updates its secondary volume concurrently. Both primary and secondary volumes must be in the MM (PPRC) relationship, but z/OS instructs the DS8000 storage system to avoid the use of MM replication for these specific I/O writes to the logs.

The I/O write to the log with IBM zHyperWrite is completed only when both primary and secondary volumes are updated by DFSMS. All I/O writes that are directed to other Db2 and IMS volumes (table spaces, indexes, and others) are replicated to the secondary volumes by DS8000 MM, as shown in Figure 5-19.

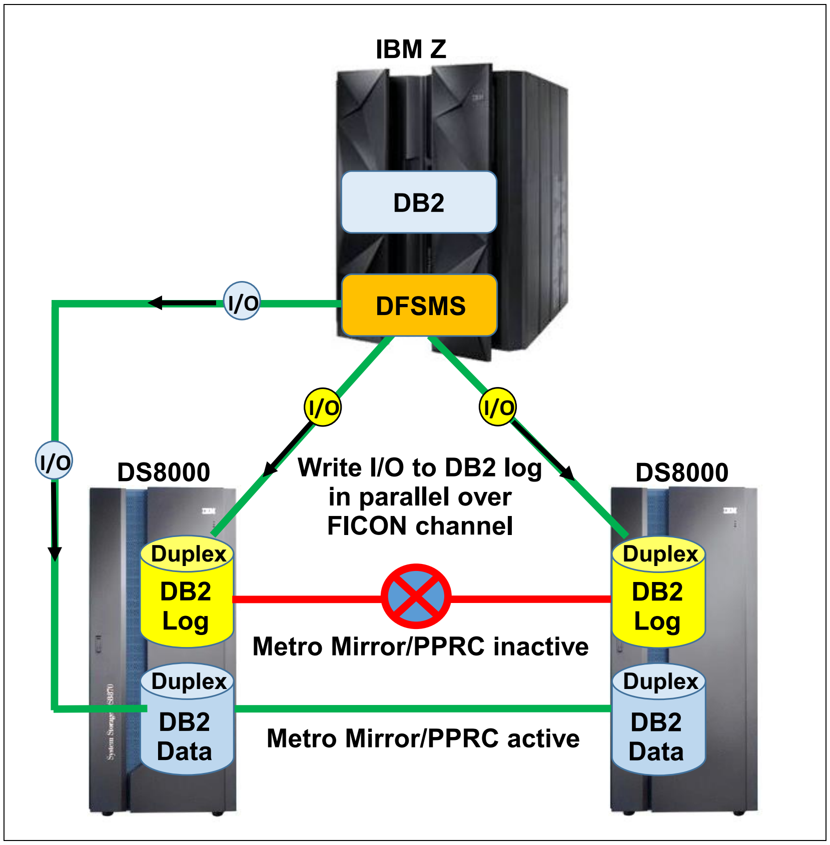


Figure 5-19 IBM zHyperWrite active: Db2 logs updated by DFSMS

The same logic applies if you do not have dedicated volumes for Db2 logs only (that is, a mixture of Db2 logs and other Db2 data sets on the same volume). Only the I/O writes that are flagged as eligible for IBM zHyperWrite (writes to a Db2 log data set) are replicated by DFSMS Media Manager. The other I/O writes are under DS8000 MM control.

The following prerequisites must be met for IBM zHyperWrite enablement:

- Primary and secondary volumes with Db2 logs are in MM (PPRC) replication.
- Primary and secondary MM volumes are in the full duplex state.
- HyperSwap is enabled; therefore, CSM or GDPS is used for replication management.

If the Db2 log volume (primary or secondary) is not in the full duplex state, the IBM zHyperWrite I/O to the eligible Db2 log data set fails. When the I/O fails, DFSMS immediately attempts to redrive the failed I/O to the primary volume, but this time without IBM zHyperWrite, which gives full control to the DS8000 storage system to replicate it by using the MM, as shown in Figure 5-20.

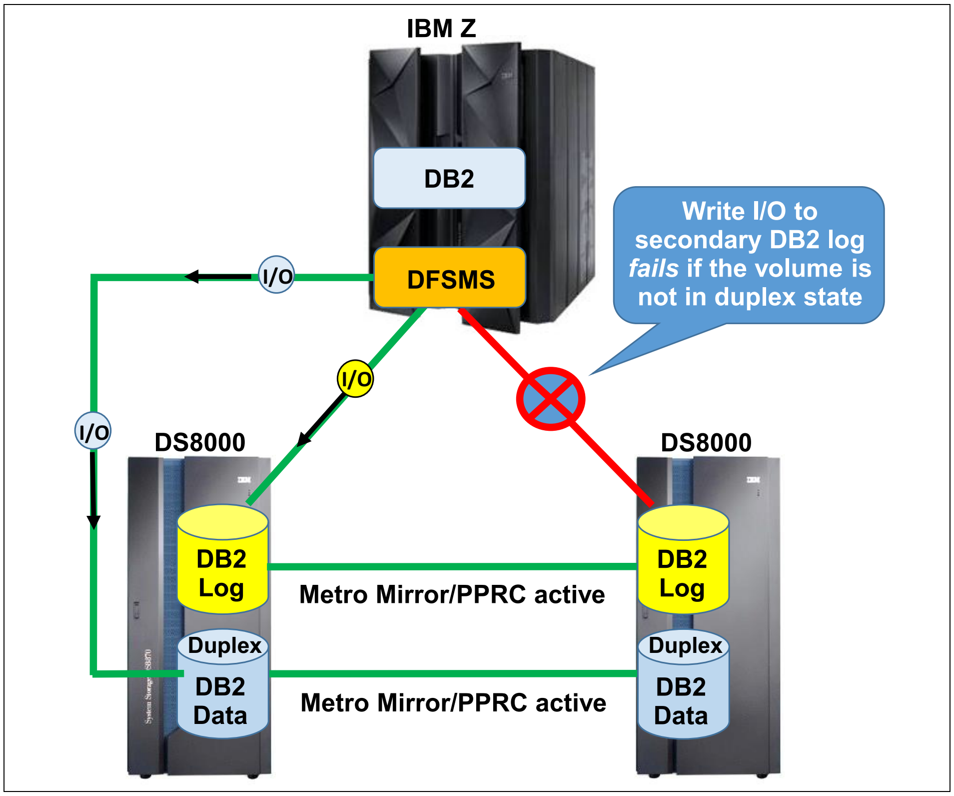


Figure 5-20 IBM zHyperWrite inactive: Db2 logs replicated with DS8000 Metro Mirror

Similarly, if any issues exist with FICON channels to the secondary Db2 log volume, DFSMS gives instruction to the DS8000 storage system to start MM replication.

5.8.1 IBM zHyperWrite installation and enablement

IBM zHyperWrite support is provided on DS8000 storage system R7.4 and later. For more information about PTFs and APARs that are related to IBM zHyperWrite, see this IBM Support web page.

The IBM zHyperWrite function is enabled by default when the required APARs and PTFs are applied and all other IBM zHyperWrite prerequisites are met. The following IBM zHyperWrite statement is included in the **IECIOSxx** member:

```
HYPERWRITE=YES
```

Alternatively, you can enable IBM zHyperWrite by using the following z/OS command:

```
SETIOS HYPERWRITE=YES
```

Example 5-2 shows the z/OS command that is used to verify the IBM zHyperWrite status and determine whether it is enabled or disabled.

Example 5-2 z/OS command for checking the zHyperSwap status

```
D IOS,HYPERWRITE
IOS634I IOS SYSTEM OPTION HYPERWRITE IS ENABLED
```

However, Db2 provides its own control mechanism to enable or disable the use of IBM zHyperWrite. The required Db2 APAR PI25747 introduced a new **ZPARM** keyword parameter (**REMOTE_COPY_SW_ACCEL**) that was added to the **DSN6LOGP** macro; the valid keywords are `ENABLE` and `DISABLE`. By default, this parameter is disabled.

You can verify whether IBM zHyperWrite is enabled by using the Db2 command that is shown in Example 5-3. The output of the **DSNJ370I** message is updated. The `SOFTWARE ACCELERATION` status refers to the IBM zHyperWrite software-controlled mirroring and it can be `ENABLED` or `DISABLED`.

Example 5-3 Db2 command for checking the zHyperSwap status

```
-DISPLAY LOG
DSNJ370I csect-name LOG DISPLAY
CURRENT COPY1 LOG = dsname1 IS pct % FULL
CURRENT COPY2 LOG = dsname2 IS pct % FULL
H/W RBA = hw-rba ,
H/O RBA = ho-rba
FULL LOGS TO OFFLOAD = nn OF mm ,
OFFLOAD TASK IS status
SOFTWARE ACCELERATION IS ENABLED
```

With High-Performance FICON ED, the IBM zHyperWrite function provides overall enhanced resilience for workload spikes. Because of the improved Db2 transactional latency and log throughput improvements, more room is available for workload growth and potential cost savings from workload consolidations.

5.8.2 IBM zHyperWrite and IMS 15

IMS 15 introduced the use of DFSMS Media Manager to write data to the write-ahead data set (WADS). This change enabled IMS to use important I/O features, such as zHPF, which increases I/O throughput, and IBM zHyperWrite, which reduces latency time for synchronous replication products.

The use of zHPF and IBM zHyperWrite can be specially helpful for data sets with high write rates, such as WADS, which increases logging speed. You can reduce service times for WADS data sets by up to 50%, depending on your environment.

For more information about WADS support for IBM zHyperWrite, including migration considerations, see this IBM Documentation web page.

5.9 DS8000 Copy Services performance considerations

IBM z/Architecture and its channel subsystem provide features to help mitigate the performance implications of DS8000 data replication, especially the synchronous data replication called MM.

How IBM Z I/O architecture and technology that is combined with the DS8000 storage system helps to reduce the overall response time and addresses the timing of each component is shown in Figure 5-21. Whether this process is done at the operating system, channel subsystem, or the DS8000 storage system level, each active component interacts with the other components to provide the maximum synergy effect that is possible.

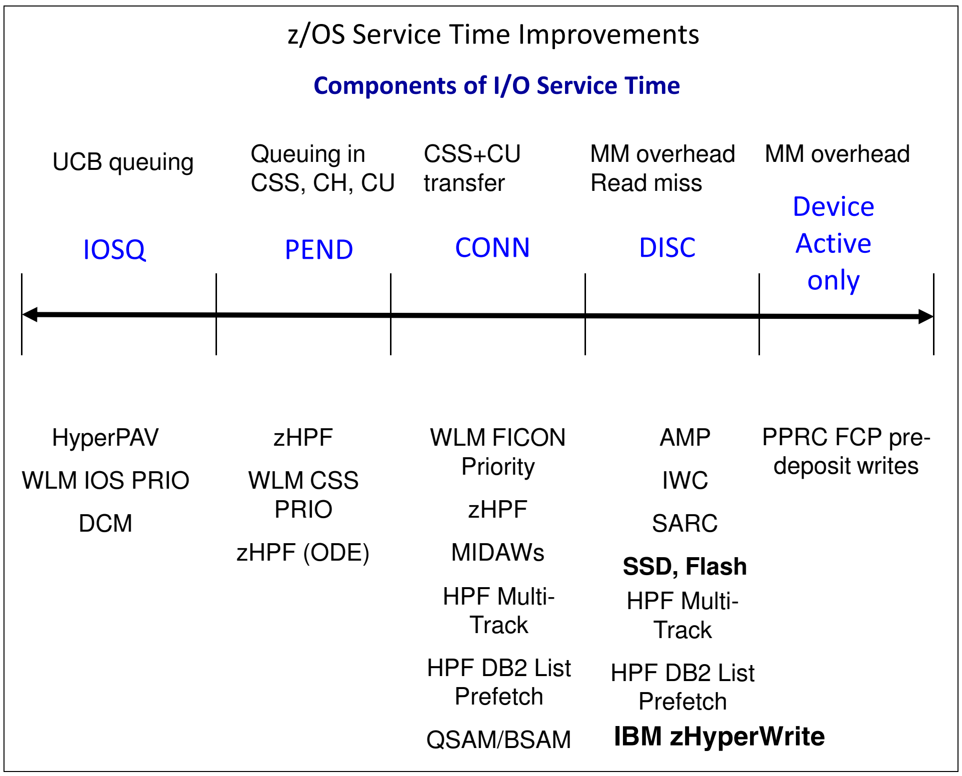


Figure 5-21 How z/OS and the DS8000 storage system address I/O timing components

Some of the basic ideas are also implemented in the DS8000 replication firmware, and include the enhancement to standard FCP. The predeposit writes improve the standard FCP and optimize the number of protocol handshakes during an MM I/O operation. Several I/O requests are sent before a handshake signals to the sending storage system that everything arrived safely at the secondary site.

The number of protocol exchanges that is needed to run the I/O chain is minimized compared to standard FCP. The goal here is to combine several aspects, such as how many I/Os to send before the handshake, the number of resources that are needed for buffers, and how smart error detection and error recovery are to run at the maximal speed and performance. This approach provides good performance with the DS8000 storage system when Copy Services functions of up to 300 km (186.41 miles) distance between sites are used.

This technology is used by DS8000 Copy Services functions and zHPF through ED II. For more information, see 5.5, “High-Performance FICON for IBM Z” on page 86.

The use of PAVs, HyperPAV, and (even better) SuperPAV with DS8000 multiple allegiance and I/O priority queuing functions can help mitigate the potential performance effect through synchronous MM replication.

However, even without those extra performance improvement features, the MM solution as implemented in the DS8000 storage system is one of the most efficient synchronous replication solutions compared to other storage systems. With a single round trip and considering the speed of light in a nonvacuum environment at approximately 200,000 Kps, the synchronous MM effect is only 10 ms per 1 km (0.62 miles) round trip, which means that the synchronous replication effect is about 1 ms when both sites are 100 km (62.1 miles) apart.

Asynchronous MM often has no effect on application write I/Os.

5.10 DS8000 storage system and IBM Z I/O enhancements

Many host I/O performance enhancements were introduced with 16 Gbps FICON, which contributes to simplifying infrastructure, reduces I/O latency for critical applications and elapsed time for I/O batch jobs, and enables even higher standards for RAS. The 32 Gbps FICON, available with IBM z16 and later, again increases host I/O performance significantly.

This section describes the DS8000 storage system and the I/O enhancements that are available with IBM z14 and later:

- FICON Dynamic Routing (FIDR)
- Forward Error Correction (FEC)
- Read Diagnostic Parameters (RDP) for improved fault isolation
- SAN Fabric I/O Priority management
- IBM zHyperLink

For more information about the I/O enhancements that were provided with IBM z13 and above, see *Get More Out of Your IT Infrastructure with IBM z13 I/O Enhancements*, REDP-5134.

5.10.1 FICON Multihop

Planning the FICON connections for mainframe environments can be challenging, especially when your mainframe solution includes multisite requirements for HADR purposes and you are limited to only two cascades switches and one hop. This limitation increases the number of switches that are necessary to achieve your solution.

To overcome this limitation, IBM announced the FICON Multihop, which enables the use of up to four FICON switches or directors and up to three hops between your devices. This feature can reduce the number of switches and the complexity of your network configuration.

A sample FICON configuration for geographically dispersed data centers with noncascaded and cascaded FICON switches is shown in Figure 5-22.

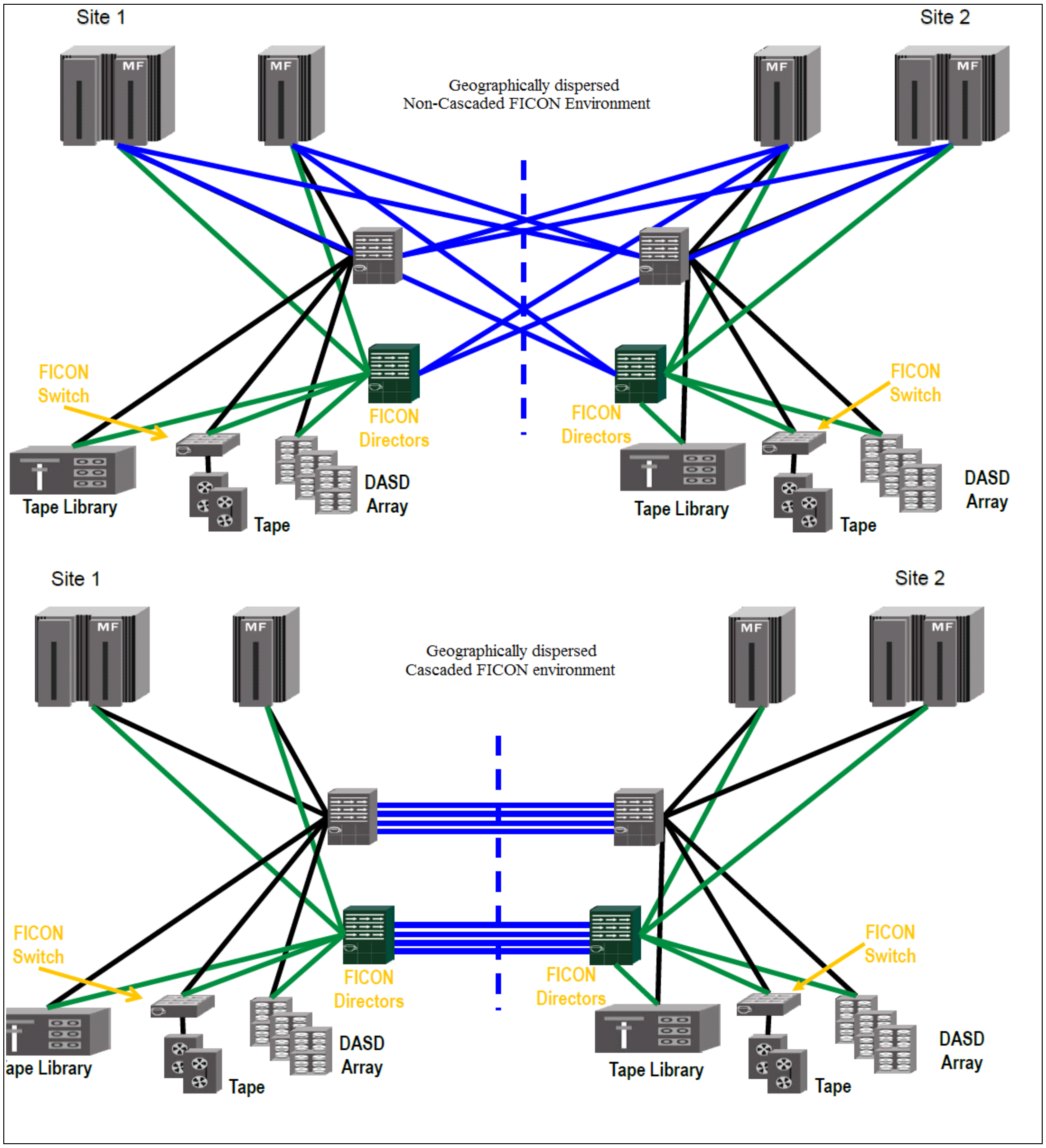


Figure 5-22 Sample FICON configuration for geographically dispersed data centers

When the HCD relationship is defined, you do not define the interswitch link (ISL) connections. All the traffic management between the ISLs is performed by the directors that use the Fabric Shortest Path First (FSPF) protocol (see “Fabric Shortest Path First” on page 103). The HCD assumes that the links are present, and it requires 2-byte addressing that specifies the destination director ID and the port to be used within that director.

Note: Multihop is supported by the use of traditional static routing methods only. It is *not* supported by FIDR.

Fabric Shortest Path First

The FSPF protocol is the standardized routing protocol for FICON SAN fabrics. FSPF is a link state path selection protocol that directs traffic along the shortest path between the source and destination that is based on the link cost.

FSPF detects link failures; determines the shortest route for traffic; updates the routing table; provides fixed routing paths within a fabric; and maintains correct ordering of frames. FSPF also tracks the state of the links on all switches in the fabric and associates a cost with each link.

The protocol computes paths from a switch to all the other switches in the fabric by adding the cost of all links that are traversed by the path, and chooses the path that minimizes the costs. This collection of the link states (including costs) of all the switches in the fabric constitutes the topology database or link state database.

FSPF is based on a replicated topology database that is present in every switching device in the FICON SAN fabric. Each switching device uses information in this database to compute paths to its peers by using a process that is known as *path selection*. The FSPF protocol provides the mechanisms to create and maintain this replicated topology database.

When the FICON SAN fabric is first initialized, the topology database is created in all operational switches. If a new switching device is added to the fabric or the state of an ISL changes, the topology database is updated in all the fabric's switching devices to reflect the new configuration.

Requirements and support

Several requirements must be met before you can use multihop, including requirements for IBM Z hardware, DASD and storage, SAN hardware, and network and DWDM requirements.

For more information about the requirements, see the *FICON Multihop: Requirements and Configurations WP102704* white paper.

5.10.2 FICON Dynamic Routing

Static and dynamic routing policies are available on SAN fabrics; however, only static routing policies are supported for mainframe FICON users.

The SAN fabric routing policy is responsible for selecting a route for each port in the fabric. Various policy options are available from Brocade and Cisco, as listed in Table 5-1.

Table 5-1 SAN Fabric routing policies

Routing policy	Brocade	Cisco
Static	Port-based routing (PBR)	N/A
Static	Device-based routing (DBR)	Default static routing policy
Dynamic	Exchange-based routing (EBR)	Originator exchange ID routing (OxID)

PBR assigns static ISL routes, based on first come, first served at fabric login (FLOGI) time. The actual ISL that is assigned is selected in a round-robin fashion. Even the ports that never send traffic to the cascaded switch are assigned ISL routes. This situation sometimes results in some ISLs being overloaded while other available ISLs are not used at all.

The routing can change every time that the switch is initialized, which results in unpredictable and nonrepeatable results.

Figure 5-23 shows a situation where the access to the remote disks is routed only through one ISL link. The remaining ISL ports are not used because other channels do not need access to the remote devices. This result occurs because of the static routing policy that assigns ISL routes to each port as it logs in to the fabric.

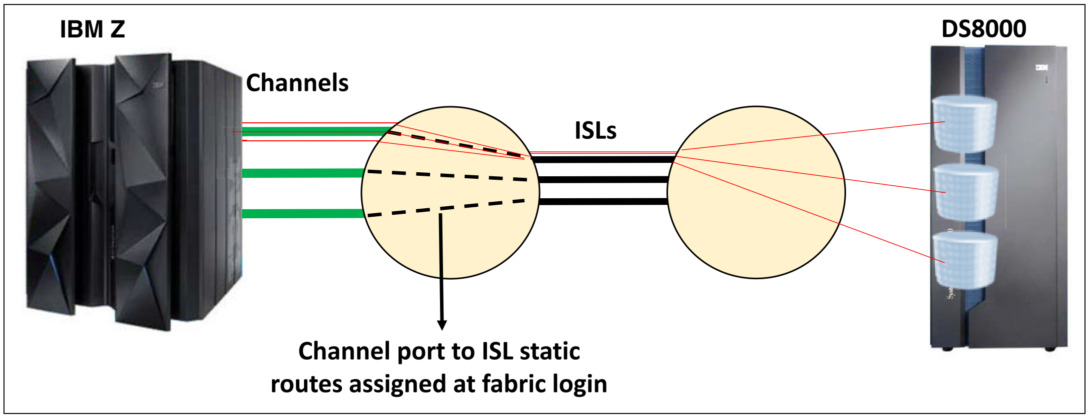


Figure 5-23 Static routing policy: Brocade port-based routing

Brocade DBR and the Cisco default routing policy create a set of static routes that are based on a hash of the source and destination Fibre Channel port addresses. Therefore, every flow in the fabric can take a different path. For Brocade FICON Directors, DBR is more efficient at spreading the work over all the available ISLs than PBR.

Figure 5-24 shows DBR where the ISL route is assigned based on a hash of the source and destination port addresses. This method is much more likely to spread the work across all the available ISLs.

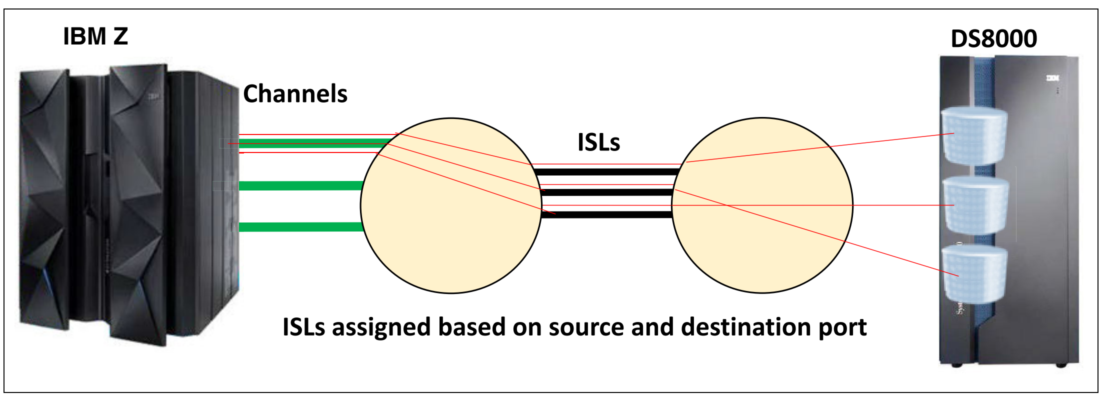


Figure 5-24 Static routing policy: Brocade device-based routing and CISCO default static routing

FICON channels are not restricted to the use of static SAN routing policies for cascading FICON directors. The IBM Z feature that supports dynamic routing in the SAN is called FIDR. It supports the Brocade static SAN routing policies, including PBR and DBR, and the Brocade dynamic routing policy, which is known as EBR.

FIDR also supports Cisco default static routing policies for cascaded SAN switches and the Cisco dynamic routing policy for cascaded SAN switches, which is known as OxID.

Dynamic routing changes the routes between host channels and DS8000 storage systems that are based on the Fibre Channel Exchange ID. Each I/O operation features its own exchange ID. Therefore, all available ISL ports are evenly used because the host I/O traffic is routed evenly across all ISLs.

The FIDR, which is a feature since z13 that you use to set up dynamic routing policies in the SAN, is shown in Figure 5-25.

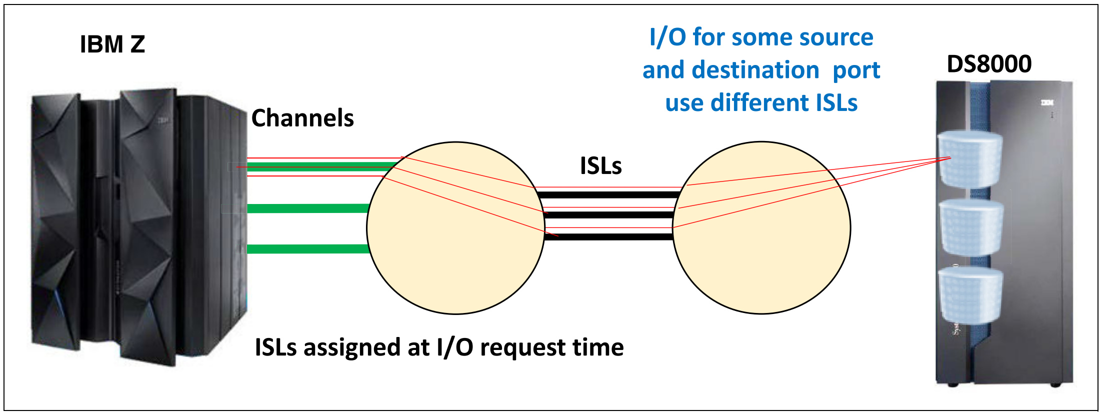


Figure 5-25 Dynamic routing: Brocade exchange-based routing and CISCO OxID routing

> **Note:** FIDR is supported on IBM z13 and later, and Brocade Fabric OS V7.4 and later.

By using the IBM Z support of dynamic routing policy, you can share SAN fabrics between FICON and FCP (PPRC) traffic and reduce overall SAN infrastructure costs. Because the ISLs can be shared between FICON and FCP traffic, extra ISL ports are not needed, and more dark fiber links between sites do not need to be leased.

Shared SAN fabric design provides simplified management, easier problem determination, and improved performance through more efficient and balanced use of ISLs. ISLs can be driven to higher use before incurring queuing delays that might result in longer I/O service times.

> **Tip:** With FIDR, accommodating FCP (PPRC) traffic on the same ISL is possible because you no longer separate virtual fabrics with separate ISLs for FCP and FICON traffic.

This enhancement positions IBM Z together with DS8000 for future innovations in SAN technology.

5.10.3 Forward Error Correction

With 16 or 32 Gbps Fibre Channel link speeds, optical signals become more sensitive to environmental effects. These signals can be degraded because of the poor quality of optical cables (twisting and bending), dust, or faulty optic connectors.

Many clients experienced optical signal sensitivity because of a faulty cabling infrastructure when they migrated from 2 or 4 to 8 Gbps link speeds. This vulnerability increases even more with 16 or 32 Gbps. Standard tools sometimes do not reveal problems before production work is deployed. Therefore, deployment of 32 or 16 Gbps technology requires mechanisms to prevent faulty links from causing I/O errors to occur.

IBM added FEC technology between IBM Z and DS8000 storage systems. This technology captures errors that are generated during the data transmission over marginal communication links. FEC is auto-negotiated by the DS8000 storage system and IBM Z channels by using a standard Transmitter Training Signal (TTS). On the SAN fabric switch, you must enable ports that are connected to IBM Z and DS8000 storage systems to use this standard TTS auto-negotiation.

The Brocade **portcfgfec** command that is used for FEC enablement is shown in Example 5-4.

Example 5-4 Brocade command for FEC enablement

```
switch:admin> portcfgfec --enable -TTS 5
Warning : FEC changes will be disruptive to the traffic
FEC TTS is supported only on F-port.
WARNING: Enabling TTS on E-port, EX-port and D-port will disable the port.
TTS has been enabled.
```

Note: The Brocade **portcfgfec** command includes two parameters (**FEC** and **TTS**) that can be used to enable FEC. The **FEC** parameter is based on the Brocade propriety method. The TTS method is open-standard-based, which is the same as the DS8000 storage system and IBM Z FICON host adapters. The **TTS** parameter must be used to enable end-to-end FEC between an IBM Z host and the DS8000.

By enabling FEC, clients see fewer I/O errors because the errors are corrected by the error correction technology in the optical transmit and receive ports. The end-to-end link should run at the 16 Gbps speed to achieve the maximum latency reduction. Moreover, the full path from the host channel through SAN fabric to the CU should FEC enabled to minimize the risk of I/O errors.

Note: No configuration is needed for direct links between the DS8000 storage system and the IBM Z channel, or direct PPRC links.

Fibre Channel Framing and Signaling 3 (FC-FS-3) and Fibre Channel Physical Interfaces (FC-PI-5) standards (from the T11.org[1]) are being updated to bring FEC to optical SAN fabric for IBM Z and the DS8000 storage system. These standards define the use of 64b/66b encoding, so the overall link efficiency improves to 97% versus 80% with the previous 8b/10b encoding.

FEC encoding operates at the same high efficiency and improves reliability by reducing bit errors. (Errors less than 10,000 times likely to be seen.) Any single bit error or up to 11 consecutive bit errors per 2112 bits can be corrected.

[1] For more information, see http://www.t11.org/index.html.

High-level end-to-end FEC enablement is shown in Figure 5-26.

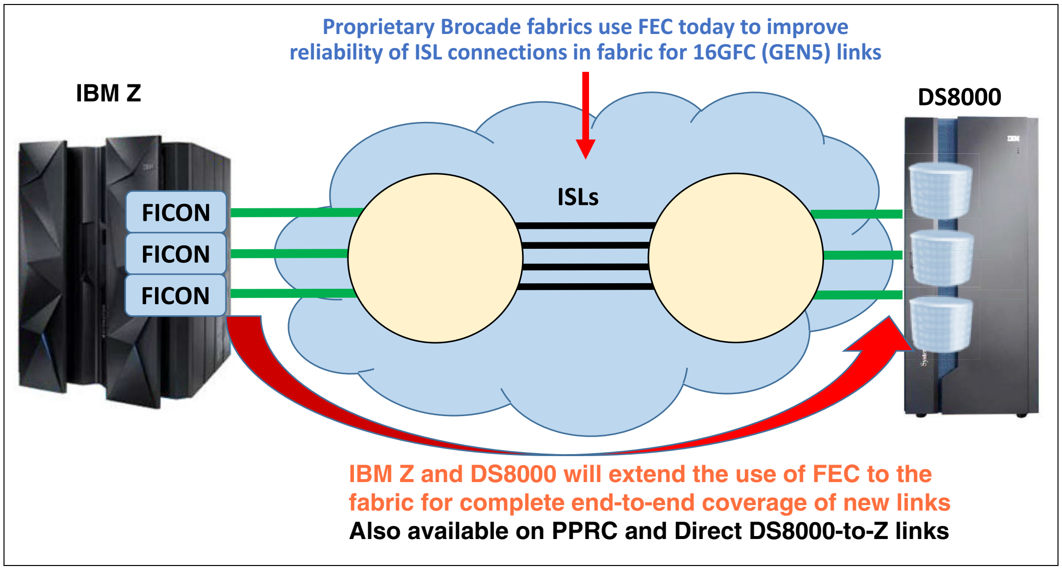


Figure 5-26 Forward Error Correction: IBM Z to DS8000 storage systems

Note: FEC is supported on 16 GFC and faster links (end to end) for IBM z13 and later, and Brocade FOS V7.4 and later.

Also, the IBM c-type switches, or the Cisco MDS 9700 Series of FICON Multilayer Directors, support FEC, including Inter-Switch Links (ISLs) between the directors.

5.10.4 Read Diagnostic Parameters for improved fault isolation

One of the significant challenges for clients is problem determination that is caused by faulty connections after hardware is added or upgraded. Even when the problem is detected, identifying the real root cause (that is, which part of the link is fault) can be difficult and time-consuming. Is it because of a damaged cable or connector (SFP transceiver)?

The T11 RDP standard defines a method for SAN fabric management software to retrieve standard counters that describe the optical signal strength (send and receive), error counters, and other critical information for determining the quality of the link. After a link error is detected (such as Interface Control Check: Condition Code 3, reset event, or link incident report), software can use link data that is returned from RDPs to differentiate between errors that are caused by failures in the optics versus failures that are the result of dirty or faulty links.

For example, the cable-connector path (including the cleanliness of optical connectors) is diagnosed by calculating the ratio of RX LightInputPower to TX LightOutputPower. Receivers rarely fail, and the receiver sensitivity does not change. Therefore, an indicator to clean the connector warns when the receiver optical power is too low for good signal reception and the calculated ratio of RX LightInputPower to TX LightOutputPower is too low. If this RX:TX ratio continues to be low, the cable might be broken.

All of this crucial RDP data was available in the DS8000 storage system since R7.5.

A partial output of the host port DS CLI command with the RDP data that is listed at the bottom is shown in Example 5-5. Regarding the `UncorrectedBitErr` and `CorrectedBitErr` entries, nonzero counts are expected. The counter increases during link initialization while the FEC block lock is occurring. After link init, nonzero-corrected bit errors are okay because the FEC is working. Uncorrected bit errors might be an indication of link problems.

Example 5-5 DS8000 showioport command displays the RDP data

```
dscli> showioport -metrics I0300
...
ID                        I0300
...
...
CurrentSpeed (FC)         16 Gbps
%UtilizeCPU (FC)          7 Dedicated
TxPower(RDP)              -2.0 dBm(635.6 uW)
RxPower(RDP)              -2.9 dBm(508.7 uW)
TransceiverTemp(RDP)      49 C
SupplyVolt(RDP)           3347.9 mV
TxBiasCurrent(RDP)        6.502 mA
ConnectorType(RDP)        SFP+
TxType(RDP)               Laser-SW
16GFECStatus(RDP)         Inactive
UncorrectedBlks(RDP)      -
CorrectedBlks(RDP)        -
```

The intended use of the RDP to retrieve programmatically diagnostic parameters to assist you in finding the root cause for problematic links in the SAN environment is shown in Figure 5-27.

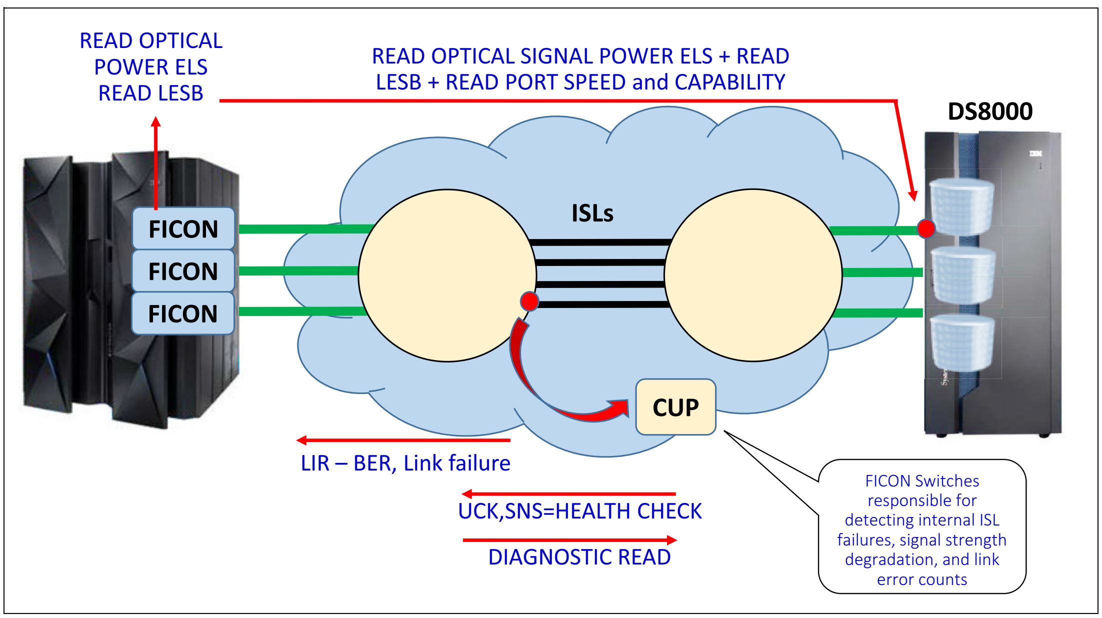


Figure 5-27 Read Diagnostic Parameters that are used to improved fault isolation

New z/OS health checks occur when the end-to-end link speeds are inconsistent and if all paths to a CU have inconsistent link speeds. These checks simplify diagnosing performance problems and reduce the number of useless repair actions.

The IBM Z and LinuxONE Community web page shows an example of how to use the RDP functions and the `LINKINFO` parameter on z/OS level to identify FICON connection problems on this level.

5.10.5 IBM zHyperLink

The business requirements for faster transactions and lower response times for applications drove new technology to reduce the latency that is related to retrieving data from back-end storage, such as Easy Tier and flash storage. Although these solutions help address the time that is required to read the data from your physical media, other parts of your I/O processing can also use valuable amounts of time and affect your latency.

IBM zHyperLink, which was introduced on IBM z14, aims to provide a short distance direct connection of up to 150 meters (492 feet). IBM zHyperLink is a new synchronous I/O paradigm. This paradigm eliminates z/OS dispatcher delays, I/O interrupt processing, and the time that is needed to reload the processor cache that occurs after regaining control of the processor when I/O completes. IBM zHyperLink delivers up to 10 times latency improvement. IBM zHyperLink improves application response time, which cuts I/O-sensitive workload response time in half without significant application changes.

IBM zHyperLink and zHPF

Although IBM zHyperLink represents a substantial enhancement over FICON connections, it does not replace these connections. Instead, IBM zHyperLink works with FICON or zHPF to reduce application latency. The workload that is transferred by zHPF reduces with the implementation of IBM zHyperLink. Not all I/Os are eligible for IBM zHyperLink.

Also, if an IBM zHyperLink I/O is unsuccessful (for example, because of a read cache miss), the I/O is redriven over FICON.

IBM zHyperLink is a PCIe connection and does *not* reduce the physical number of current zHPF connections.

The different latency times for IBM zHyperLink and zHPF are shown in Figure 5-28.

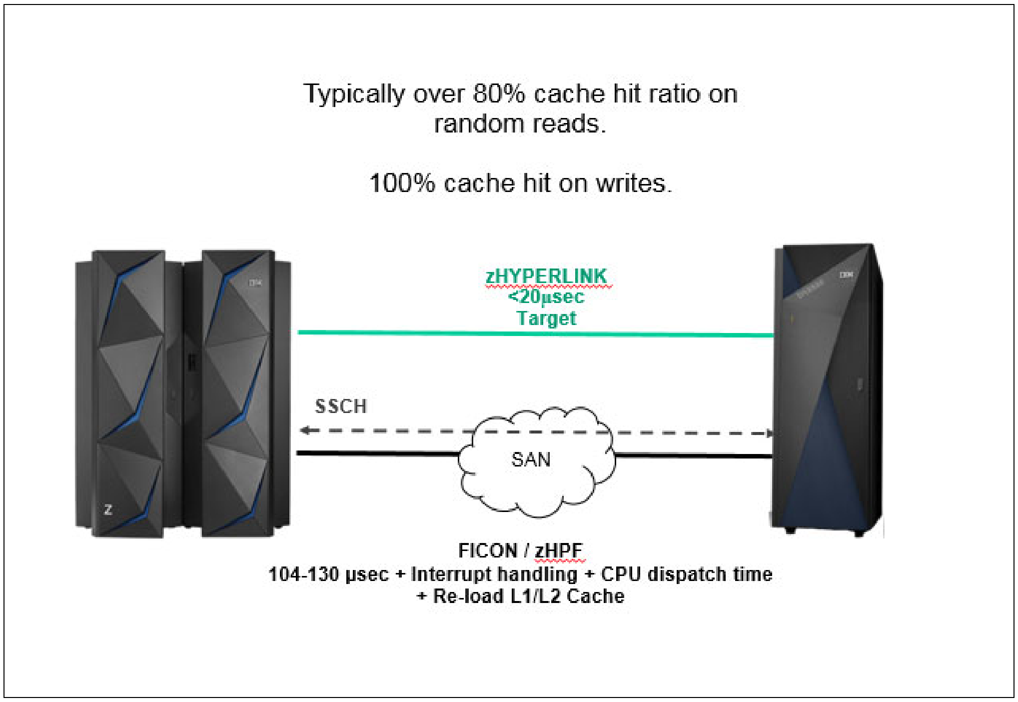


Figure 5-28 IBM zHyperLink and zHPF response times

IBM zHyperLink considerations

A few conditions must be met before you can use IBM zHyperLink. IBM zHyperLink is available only to z14 and later hardware.

The IBM zHyperLink hardware is designed for short distance communication of up to 150 meters (492 feet), so your DS8900F must be no further than this length from your IBM Z hardware. Also, the IBM Z and the DS8000 hardware must have the IBM zHyperLink hardware installed to communicate.

IBM zHyperLink and synchronous I/O

IBM zHyperLink can provide low latency for I/Os, and can deliver an I/O operation in less than 20 microseconds in some scenarios. The time that is spent with interrupt handling, CPU dispatch time, and reload L1/L2 cache can also increase your I/O time (see Figure 5-28).

When an application (in this case, Db2) requests an I/O through IBM zHyperLink, a synchronous I/O operation is requested. Therefore, the task is not dispatched from the CPU while the I/O is performed.

The workflow of an I/O operation that defines whether an IBM zHyperLink synchronous I/O operation is attempted is shown in Figure 5-29.

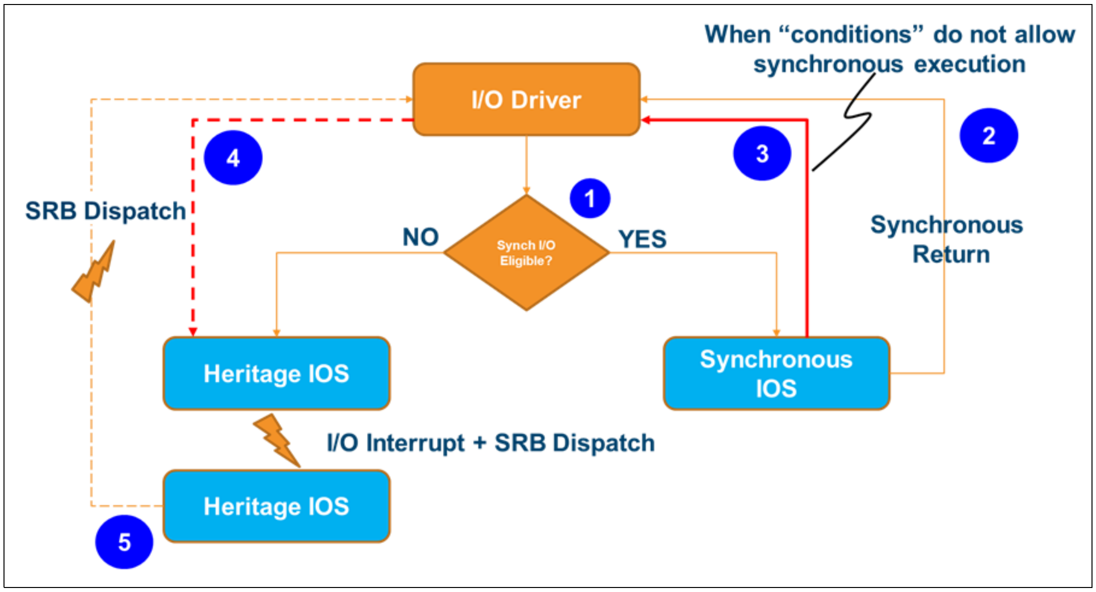


Figure 5-29 I/O operation work flow

As shown in Figure 5-29, the I/O process flow includes the following steps:

1. The I/O drivers request a synchronous execution. The system checks whether the I/O is eligible for synchronous I/O, and if so, it is directed for synchronous IOS through IBM zHyperLink.
2. If the I/O is returned in an acceptable amount of time, the I/O is returned to the I/O driver.
3. If the I/O is not returned in an acceptable amount of time (for example, if the data is not in the DS8000 cache and a prefetch is necessary), the request is returned to the I/O driver.
4. The I/O driver requests the I/O to be performed through the heritage IOS.
5. When the I/O response is received, it is returned to the I/O driver the same way it is done today.

Note: Because the prefetch is requested when the synchronous I/O is attempted, the subsequent heritage I/O can still perform better than a request that did not attempt a synchronous I/O.

Other considerations

You can have up to 16 IBM zHyperLink Express adapters in a z15, for a total of 32 ports. These adapters do not take up fanout slots like short distance couplings do. Each port can be shared among up to 16 LPARs, and have a maximum suggested PCIe Function IDs (PFIDs) per IBM zHyperLink per LPAR.

The DS8900 storage system can have up to 12 IBM zHyperLink connections. The number of connections depends on the model (number of I/O enclosures and cores, as listed in Table 5-2).

Table 5-2 IBM zHyperLink configuration summary

System/Model	Cores per processor	IBM zHyperLink supported	Maximum IBM zHyperLink connections (increments of 2)
DS8910F	8 (993 Model)	Yes	4
	8 (994 Model)	Yes	4
DS8950F	10	Yes	6
	20	Yes	8 (in base frame only) 12 (with two frames)
DS8980F	22	Yes	8 (in base frame only) 12 (with two frames)

IBM zHyperLink enables the following interconnections:

- One IBM Z to one DS8000 storage system by using multiple IBM zHyperLink connection pairs
- Multiple IBM Z mainframes to a single DS8000 storage system
- Multiple DS8000 storage systems to a single IBM Z mainframe

Initially, only Db2 used IBM zHyperLink, for random 4 KB or smaller reads and writes. The scope of IBM zHyperLink use expanded continuously. With APAR OA52941, VSAM read support is provided, and all VSAM record type data sets support IBM zHyperLink. At the time of this writing, only Db2 logs are eligible for write operations. For more information about Db2 active log support, see APARs PH05030 and OA56575.

Note: For writes to an MM DS8000 pair, IBM zHyperWrite is required; that is, z/OS performs IBM zHyperLink operations in parallel to the DS8000 storage systems (for more information, see 5.8, “IBM zHyperWrite” on page 96).

Global Mirror write support

Before GM write support was available for IBM zHyperLink, IBM zHyperLink write adoption often was inhibited because of the distance constraints of IBM zHyperLink.

Starting with DS8000 Release 9.1, the DS8900F supports IBM zHyperLink writes to volumes that are GM sources. The DS8000 GM coordination time when building consistency groups was tuned to interact and work with IBM zHyperLink.

The following specific z/OS software levels, such as z/OS 2.2 with APAR OA56723, are required:

- z/OS 2.4 (HBB77C0) - CA56723 RW20109
- z/OS 2.3 (HBB77B0) - BA56723 RW20109
- z/OS 2.2 (HBB77A0) - KA56723 RW20109

MGM and Multi-Target MM are also supported when MM uses IBM zHyperWrite.

Enabling IBM zHyperLink

Before you can configure and activate your IBM zHyperLink connection in your host, you must also enable IBM zHyperLink in your DS8000 DASD controller. You activate IBM zHyperLink by logging on as a user with administrator privileges, selecting **Settings** and then, selecting **System**.

In the next window, select the **zHyperLink** tab. Then, select the **I/O Read Enabled** or **I/O Write Enabled** option (select both to enable read/write). After you enable IBM zHyperLink, you must save the changes in your configuration.

A sample IBM zHyperLink enablement window and the corresponding GUI are shown in Figure 5-30.

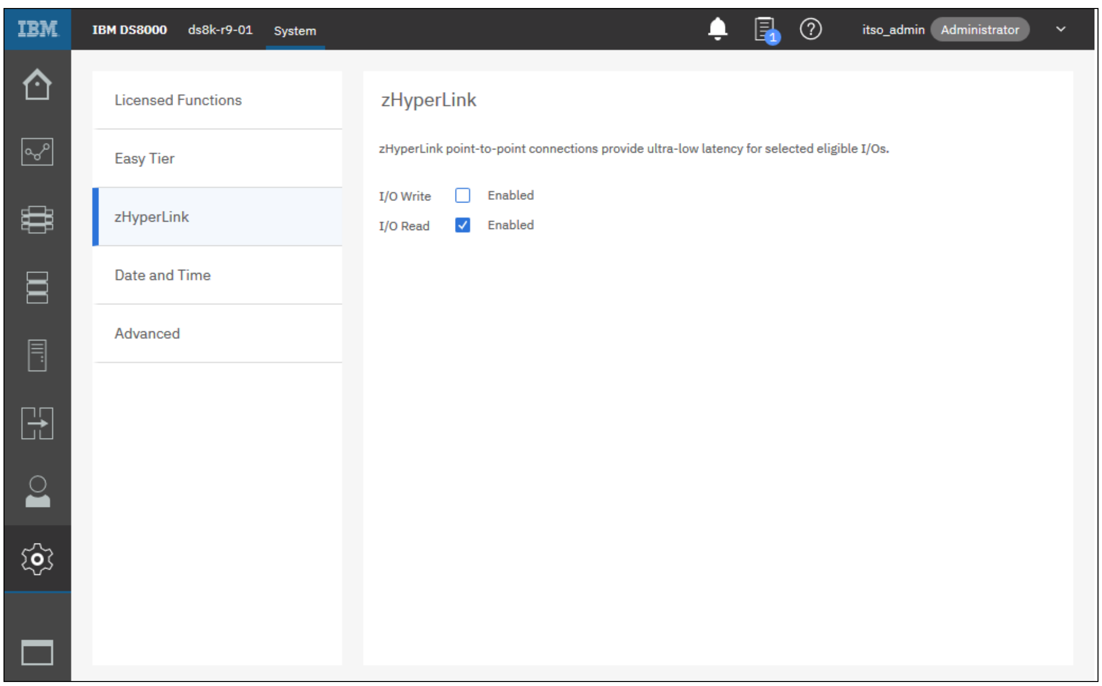


Figure 5-30 IBM zHyperLink enablement window

IBM zHyperLink is supported on native z/OS operating systems only. It is not supported on guest LPARS under z/VM or other operating systems. It was developed for IBM Db2 for z/OS.

You must use the MISSINGFIX report to determine whether any APARs exist that are applicable and not yet installed. The **`REPORT MISSINGFIX`** command checks the zones that are specified on the ZONES operand and determines whether any missing fixes exist that are based on the fix categories of interest.

Run the **`SMP/E REPORT MISSINGFIX ZONES (<your zone names>) FIXCAT (IBM.Function.zHyperLink)`** command and install all the fixes that are listed.

Example 5-6 shows a sample MISSINGFIX JCL.

Example 5-6 Sample SMP/E control cards with the FIXCAT option

```
SET BOUNDARY (GLOBAL) .
REPORT
MISSINGFIX
ZONES (
<your zone names>
)
FIXCAT(
IBM.Function.zHyperLink
) .
```

Also, consider checking `IBM.Device.Server.z14-3906*` in case of a z14 and `IBM.Device.Server.z14-3906.zHighPerformanceFicon` for zHPF.

By default, the IBM zHyperLink functions are disabled in z/OS. To enable z/OS for IBM zHyperLink read/write processing, run the **`SETIOS IBM MVS`** system command after an IPL, or use the **`IECIOSxx`** parmlib to enable IBM zHyperLink processing during IPL, as shown in Example 5-7.

Example 5-7 zHPF, IBM zHyperLink, and HyperWrite parm on the IECIOSxx parmlib

```
ZHPF=YES
ZHYPERLINK,OPER=ALL|READ|WRITE|NONE
HYPERWRITE=YES
```

Note: To activate IBM zHyperLink, the established function IOPM must remain disabled. Therefore (and because they are all-flash systems), the DS8900F models no longer offer IOPM. For DS8880, IOPM must be turned off at the storage system level. Then, this feature is off for all data, even if the data is not eligible for IBM zHyperLink. Therefore, I/O prioritization for IBM IMS data (and other data) might be affected on a DS8880 system.

IBM zHyperLink cables

When the DS8000 configuration is used with DS8900F, only a small selection of longer IBM zHyperLink cables is provided, such as 40 m (131 ft) or 150 m (492 ft). However, IBM zHyperLink installations can benefit from shorter cables; for example, having only a 3 m (9.8 ft) distance between host and DS8000. For more information about available cable lengths, cables, and their sources of supply, see Tables 52 - 54 in *IBM Z Planning for Fiber Optic Links*, GA23-1408.

For more information about the configuration, feature codes, and requirements for IBM zHyperLink on the DS8000 storage system, see the following publications:

- *IBM DS8900F Architecture and Implementation: Updated for Release 9.3*, SG24-8456
- *Getting Started with IBM zHyperLink for z/OS*, REDP-5493

5.11 IBM Fibre Channel Endpoint Security since IBM z15

The IBM Fibre Channel Endpoint Security function reinforces the synergy between DS8000 and IBM Z. IBM Fibre Channel Endpoint Security is designed to protect data that is transferred over Fibre Channel SANs and consists of two components:

- Link authentication
- Encryption of data in flight (EDiF)

Required components for this feature are the DS8900F storage system, which is an IBM z16 or z15 with encryption-capable adapters (FICON Express32S or FICON Express16SA) and IBM GKLM.

Although the 16 GFC adapters of a DS8900F offer the possibility for authentication in the SAN between DS8000 and the IBM Z host, the 32 GFC adapters provide an extra line-rate encryption capability and include support for IBM Fibre Channel Endpoint Security as part of the cybersecurity solutions that are available from IBM.

When the FICON Express 16S+ adapters are used in IBM Z instead of the 16SA or 32S, only authentication is possible. Encryption is not supported in this case.

Figure 5-31 shows possible options for a DS8900F 32-GFC adapter when setting up IBM Fibre Channel Endpoint Security in the DS8000 GUI.

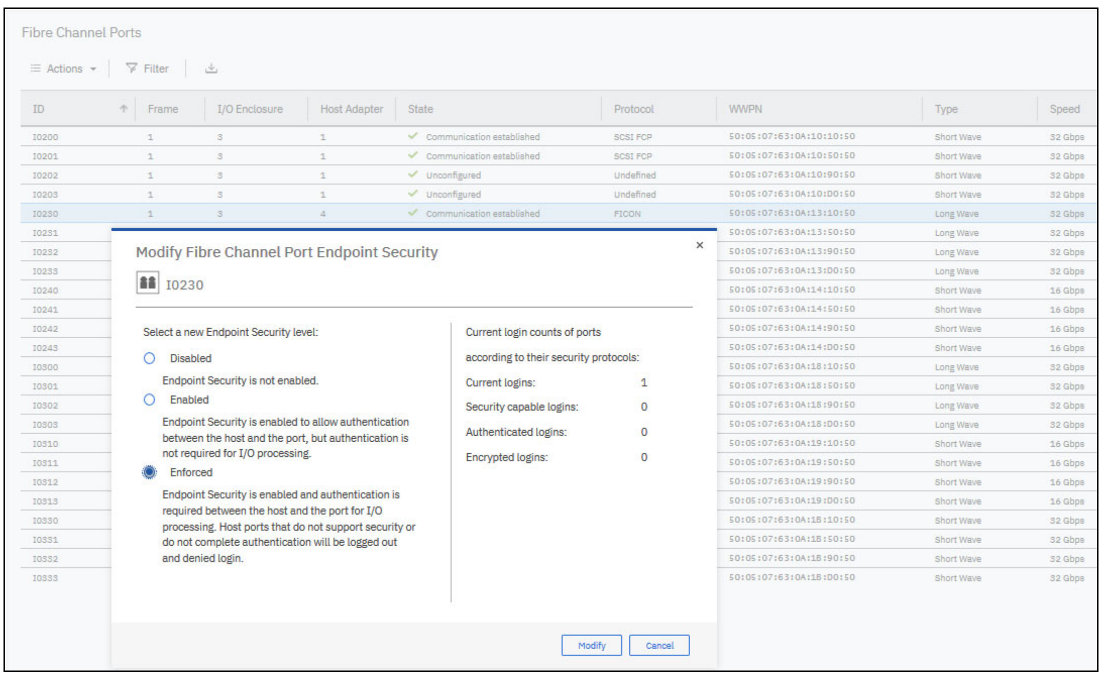


Figure 5-31 DS8900 GUI: Setting up IBM Fibre Channel Endpoint Security

For more information about the IBM Fibre Channel Endpoint Security function and how to set it up, see *IBM Fibre Channel Endpoint Security for IBM DS8900F and IBM Z*, SG24-8455.

5.12 IBM zEnterprise data compression

Another synergy component is the IBM zEnterprise® Data Compression (zEDC) capability and the hardware adapter zEDC Express, which can be used with z/OS V2.1 and later.

zEDC is optimized to be used with large sequential files (but not limited to such files) and improve disk usage with a minimal effect on processor usage.

zEDC can also be used for Data Facility Storage Management Subsystem Data Set Services (DFSMSdss) memory dumps and restores and DFSMS Hierarchical Storage Manager (DFSMShsm) when DFSMSdss is used for data moves.

For more information about zEDC, see *Reduce Storage Occupancy and Increase Operations Efficiency with IBM zEnterprise Data Compression*, SG24-8259.

An example of a zEDC compression result is shown in Figure 5-32.

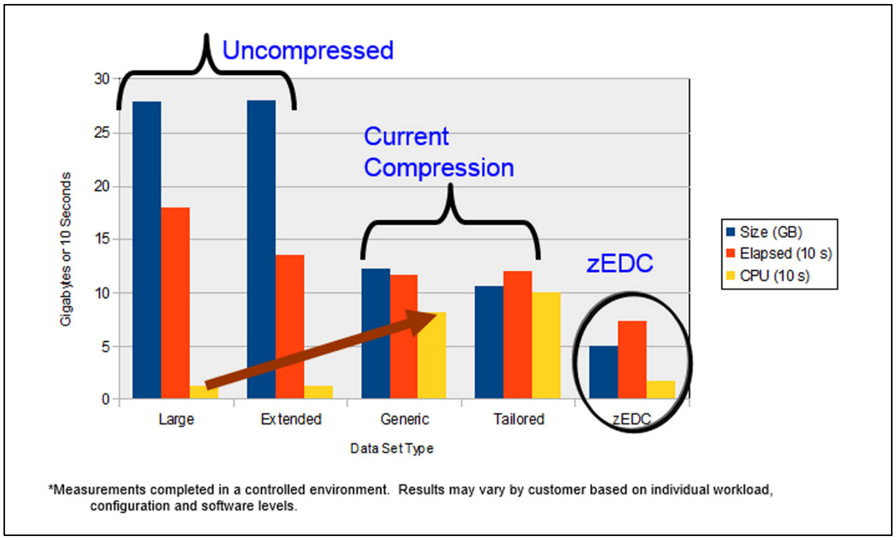


Figure 5-32 zEDC compression numbers

5.13 Transparent Cloud Tiering

Good storage management practices are based on the principle that physical space often is configured in logical pools that can be dynamically reconfigured to increase or decrease the storage capacity that is available for use. This reconfiguration also must be transparent to the user.

The storage cloud architecture introduces a new storage tier that can be used to provide extended storage capacity at a lower cost while making the data available from different locations.

The DS8000 Transparent Cloud Tiering (TCT) function provides a gateway to convert the DS8000 block storage data format for storage on private or public clouds (TCT supports IBM Cloud Object Storage, Amazon S3 API, and OpenStack Swift), or a TS7700 Virtual Tape Library that is configured for object storage.

From a z/OS perspective, CPU usage is reduced (often referred to as a *MIPS reduction*) by using direct data transfer from the DS8000 storage system to IBM Cloud Object Storage. DFSMShsm and DFSMSdss are used to control the migration and retrieval from the cloud, as shown in Figure 5-33, but the data is directly transferred to the cloud.

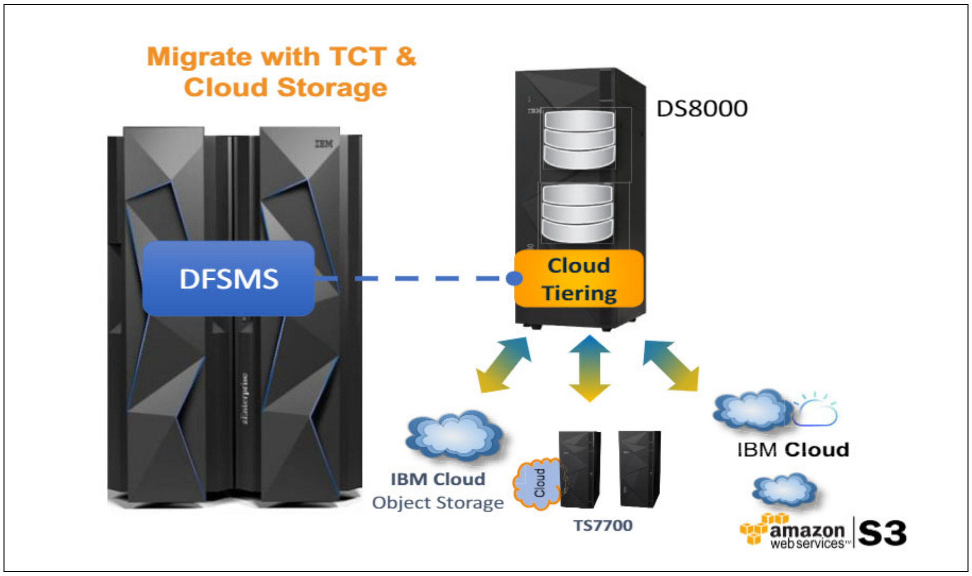


Figure 5-33 Transparent Cloud Tiering data movement

TCT provides support to migrate and recall data sets to volumes that participate in Simplex, FlashCopy, 2-site MM, GM, and MGM relationships.

DS8900F TCT multi-cloud support

Since DS8000 Release 9.2, you can define up to eight cloud connections.

Note: Before DS8000 Release 9.2, the DS8000 supported only a single cloud connection.

With this release, you can back up and archive mainframe data to up to eight clouds (private, public, and TS7770) on a single system. DS8000 Release 9.2 provides extended flexibility and opens the TCT solution for the following new cases:

- Multiple TS7700 Grids: Connect to multiple Grids; for example, to separate production from test Grids environments.
- Public versus private cloud for different types of data; for example, to keep confidential data onsite while other data can move to a public cloud.
- Performance differentiation: Keep more active backups or frequently recalled data on site on TS7700, and move nonreferenced, colder data to public clouds.
- Application separation: Separate HSM and DSS workloads or maintain individual credentials for applications.
- Managed service providers: Provide clients with backup/archive solutions that are tailored to each client that is served by a single DS8000.
- Test environments: Take advantage of having multiple clouds for testing various types of clouds without reconfiguration.

For more information about TCT multi-cloud support, see 4.13, “Transparent Cloud Tiering multi-cloud support” on page 69.

TCT services use up to two 10 Gb (recommended), or two or four 1 Gb Ethernet ports in each of the DS8000 processors. This Ethernet connectivity is also required from the mainframe to the DS8000 or the IBM Cloud Object Storage cloud server, depending on the cloud type that is chosen.

For more information about the requirements and how to configure your system to use TCT, see *IBM DS8000 and Transparent Cloud Tiering (DS8000 Release 9.2)*, SG24-8381.

Encryption and secure data transfer

TCT supports data encryption, but the encryption method depends on the target object storage.

When archiving to the cloud, the data can be encrypted as it is sent to the cloud. The data on the cloud remains encrypted and it is decrypted only when recalled by the DS8000. This method requires an external key manager, such as IBM Security Key Lifecycle Manager.

When data is migrated to a TS7700 that is configured as object storage, TCT uses a secure data transfer that encrypts the in-flight data. The data is decrypted upon arrival at the TS7700 cluster. This method needs no external key manager, but other prerequisites must be met, such as a minimum microcode level for the DS8000 and TS7700, specific features for the TS7700 (#5281), and microcode.

The encryption for the Secure Data Transfer is an AES 256-bit TLS encryption over the Ethernet TCP/IP Grid network, for which hardware acceleration is used. The task is offloaded from the DS8000 central processor complexes (CPCs) to separate IBM POWER9™ crypto-engines.

For more information about all aspects of TCT encryption, see *IBM DS8000 Encryption for Data at Rest, Transparent Cloud Tiering, and Endpoint Security (DS8000 Release 9.2)*, REDP-4500.

TCT compression

Starting with DS8900 Release 9.1, TCT data compression is possible when the TS7700 is used as object storage. Compressing data can make TCT backups faster and more efficient.

Compression is a Lempel-Ziv type and hardware-accelerated. The DS8900 uses the NX842 compression engine (off-chip), which the POWER9 architecture provides without affecting the processor.

DFSMS controls the use of compression and avoids compression if the data set is host compressed or encrypted.

TCT compression requires the DS8900 R9.1 or later microcode, and the following z/OS APARs:

- OA59465
- OA59466
- OA59467
- OA59468
- OA59469
- OA59470
- OA59471

DFSMSdss full volume dump

A full volume dump expands the area of possible TCT use cases to backups of active data. A full volume dump enables storage administrators to dump z/OS volumes to an IBM Cloud Object Storage cloud and recover a volume from an IBM Cloud Object Storage cloud.

TS7700 and traditional object storage are supported. A set of utility commands, `list` and `delete`, are added to help manage full volume dumps.

Figure 5-34 shows a use case.

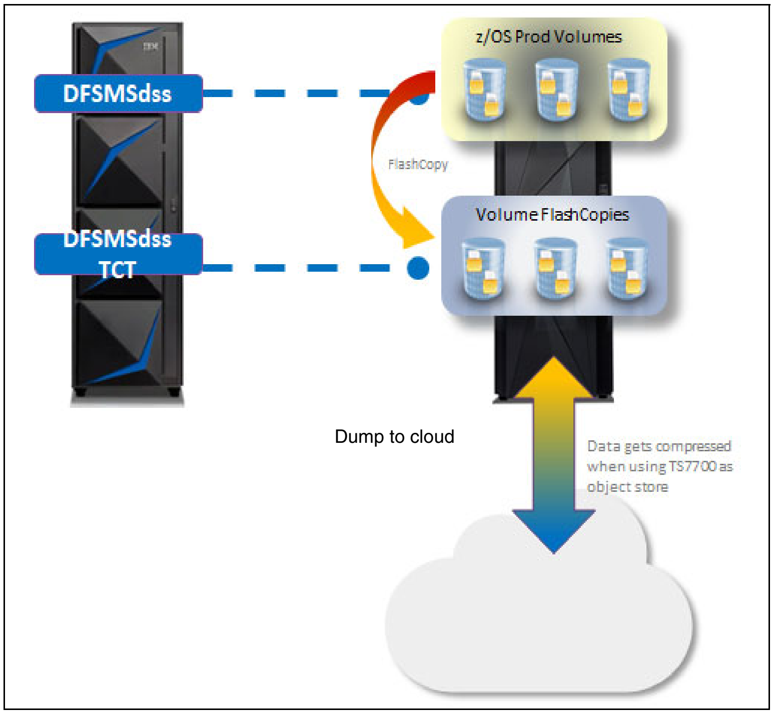


Figure 5-34 TCT full volume dump

To perform a TCT full volume dump, complete the following steps:

1. Create point-in-time volume copies by using DFSMSdss and FlashCopy by specifying `DUMPCONDITIONING`.
2. Perform a DFSMSdss command-based dump to cloud.

The volume dump in the cloud looks as though the dump was directly taken from the z/OS production volumes, and not an intermediate volume. On recovery, no manual recall is needed because data can be recovered directly into the production storage pool and applications restarted.

This native full volume dump and restore are supported for z/OS V2.4 and V2.3 with PTFs for APAR OA57526.

For more information about these TCT functions, see *IBM DS8000 and Transparent Cloud Tiering (DS8000 Release 9.2)*, SG24-8381.

Abbreviations and acronyms

ACS	Automatic Class Selection
ALP	Adaptive List Prefetch
AMG	Alias Management Group
AMP	adaptive multi-stream prefetching
API	application programming interface
ARC	Adaptive Replacement Cache
BCS	basic catalog structure
BDAM	basic direct-access method
BPAM	basic partitioned access method
BPV	Break Point Value
BSAM	basic sequential access method
BSDS	bootstrap data set
CCW	channel command word
CHPID	channel-path identifier
CKD	Count Key Data
CLI	command-line interface
CMR	command-response
CMS	Conversational Monitor System
CPBSG	Copy Pool Backup Storage Group
CPC	central processor complex
CRM	consistent read management
CRS	Consistent Read from Secondary
CSM	Copy Services Manager
CU	control unit
CUIR	control-unit-initiated reconfiguration
CVAF	Common VTOC Access Facility
DAK	Drive Access Key
DASD	direct access storage device
DBR	device-based routing
DC	Data Class
DEK	Drive Encryption Key
DFSMS	Data Facility Storage Management Subsystem
DFSMSdfp	Data Facility Storage Management Subsystem Data Facility Product
DFSMSdss	Data Facility Storage Management Subsystem Data Set Services
DFSMShsm	Data Facility Storage Management Subsystem Hierarchical Storage Manager
DFSMSrmm	Data Facility Storage Management Subsystem Removable Media Management
DK	Data Key
DL	data length
DPR	Dynamic Path Reconnect
DPS	Dynamic Path Selection
DR	disaster recovery
DSCB	data set control block
DVE	dynamic volume expansion
EADSCB	extended attribute DSCB
EAS	extended addressing space
EAV	extended address volume
EBR	Exchange-based routing
ED	Extended Distance
EDiF	encryption of data in flight
EGK	Encrypted Group Key
ENF	Event Notification Facility
ESDS	entry-sequenced data set
ESE	extent space-efficient
FB	fixed block
FC-FS-3	Fibre Channel Framing and Signaling 3
FC-PI-5	Fibre Channel Physical Interfaces
FCP	Fibre Channel Protocol
FEC	Forward Error Correction
FFDC	first-failure data capture
FICON	Fibre Channel connection
FIDR	FICON Dynamic Routing
FLOGI	fabric login
FSPF	Fabric Shortest Path First
GDPS	Geographically Dispersed Parallel Sysplex
GDPS AA	GDPS Continuous Availability
GDPS GM	GDPS Global - GM
GDPS HM	GDPS Metro HyperSwap Manager
GDPS MGM	GDPS Metro Global - GM
GDPS MzGM	GDPS Metro Global - XRC
GDPS XRC	GDPS Global - Extended Remote Copy
GK	Group Key

GM	Global Mirror
HA	high availability
HADR	high availability and disaster recovery
HCD	hardware configuration definition
HDD	hard disk drive
HFS	hierarchical file system
HMC	Hardware Management Console
HMT	Heat Map Transfer
HPFE	High-Performance Flash Enclosure
IBM	International Business Machines Corporation
ICDS	Image Copy Data Set
ICF	Integrated Catalog Facility
ICKDSF	IBM Device Support Facilities
IMS	Information Management System
IMS HP IC	IMS High-Performance Image Copy
IOCDS	I/O configuration data set
IODF	I/O definition file
IOPM	I/O Priority Manager
IOPS	I/Os per second
IOS	Input/Output Supervisor
IOSQ	Input/Output Supervisor queue
ISL	interswitch link
IU	Information Unit
IWC	Intelligent Write Caching
KL	key length
KSDS	key-sequenced data set
LCP	Logical Corruption Protection
LCU	logical control unit
LDAP	Lightweight Directory Access Protocol
LDS	linear data set
LOB	large object
LPAR	logical partition
LSS	logical subsystem
MBps	megabytes per second
MC	Management Class
MCU	multicylinder unit
MGM	Metro/Global Mirror
MIDAW	Modified Indirect Data Access Word
MM	Metro Mirror
MSS	multiple-subchannel set
MTMM	Multi-Target Metro Mirror
MTTR	mean-time-to-recovery
NDSS	nondisruptive statesave
ODD	On-Demand Dump
ODK	Obfuscated Data Key
OxID	originator exchange ID
P/DAS	PPRC dynamic address switching
PAV	parallel access volume
PBR	port-based routing
PDS	partitioned data set
PDSE	partitioned data set extended
PFID	PCIE Function ID
PPRC	Peer-to-Peer Remote Copy
QoS	quality of service
QSAM	queried sequential access method
RAS	reliability, availability, and serviceability
RDP	Read Diagnostic Parameters
RMF	Resource Measurement Facility
RPO	recovery point objective
RRDS	relative record data set
RTO	recovery time objective
SAN	storage area network
SC	Storage Class
SCI	state change interrupt
SDM	System Data Mover
SDS	software-defined storage
SG	Storage Group
SMS	Storage Management Subsystem
SPE	small programming enhancement
SSD	solid-state drive
SSL	Secure Sockets Layer
TCT	Transparent Cloud Tiering
TCW	transport control word
TDMF	IBM Transparent Data Migration Facility
TLS	Transport Layer Security
TSE	track-space efficient
TTS	Transmitter Training Signal
UCB	unit control block
VM	virtual machine
VSAM	Virtual Storage Access Method
VTOC	volume table copy
VVDS	VSAM volume data set
WADS	write-ahead data set
WLM	Workload Manager

XRC	Extended Remote Copy
zDDB	z/OS Distributed Data Backup
zEDC	zEnterprise Data Compression
zFS	z/OS File System
zHPF	High Performance FICON for IBM Z
zsS	zSynergy Services

Related publications

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

IBM Redbooks

The following Redbooks publications provide more information about the topic in this document. Some publications might be available in softcopy only:

- *DB2 for z/OS and List Prefetch Optimizer*, REDP-4862
- *DFSMShsm Fast Replication Technical Guide*, SG24-7069
- *DS8870 Data Migration Techniques*, SG24-8257
- *DS8870 Easy Tier Application*, REDP-5014
- *Get More Out of Your IT Infrastructure with IBM z13 I/O Enhancements*, REDP-5134
- *Getting Started with IBM Z Cyber Vault,* SG24-8511
- *Getting Started with IBM zHyperLink for z/OS*, REDP-5493
- *Getting started with z/OS Container Extensions and Docker*, SG24-8457
- *How does the MIDAW Facility Improve the Performance of FICON Channels Using DB2 and other workloads?*, REDP-4201
- *IBM DS8000 Easy Tier (Updated for DS8000 R9.0)*, REDP-4667
- *IBM DS8000 Encryption for Data at Rest, Transparent Cloud Tiering, and Endpoint Security (DS8000 Release 9.2)*, REDP-4500
- *IBM DS8000 High-Performance Flash Enclosure Gen2 (DS8000 R9.0)*, REDP-5422
- *IBM DS8000 Safeguarded Copy (Updated for DS8000 R9.2)*, REDP-5506
- *IBM DS8000 and Transparent Cloud Tiering (DS8000 Release 9.1)*, SG24-8381
- *IBM DS8870 Easy Tier Heat Map Transfer*, REDP-5015
- *IBM DS8870 Multiple Target Peer-to-Peer Remote Copy*, REDP-5151
- *IBM DS8880 Thin Provisioning (Updated for Release 8.5)*, REDP-5343
- *IBM DS8900F Architecture and Implementation: Updated for Release 9.2*, SG24-8456
- *IBM DS8910F Model 993 Rack-Mounted Storage System Release 9.1*, REDP-5566
- *IBM Fibre Channel Endpoint Security for IBM DS8900F and IBM Z*, SG24-8455
- *IBM GDPS Family: An Introduction to Concepts and Capabilities*, SG24-6374
- *IBM System Storage DS8000: Host Attachment and Interoperability*, SG24-8887
- *IBM z/OS Global Mirror Planning, Operations, and Best Practices*, REDP-4878
- *IBM z/OS V2R2: Storage Management and Utilities*, SG24-8289
- *IBM z16 (3931) Technical Guide*, SG24-8951
- *IBM z16 Technical Introduction*, SG24-8950
- *IBM z16 Configuration Setup*, SG24-8960
- *IBM z15 (8562) Technical Guide*, SG24-8852

- *IBM z15 Technical Introduction*, SG24-8850
- *IBM Z Connectivity Handbook*, SG24-5444
- *IBM Z Functional Matrix*, REDP-5157
- *LDAP Authentication for IBM DS8000 Systems: Updated for DS8000 Release 9.1*, REDP-5460
- *Multiple Subchannel Sets: An Implementation View*, REDP-4387
- *Reduce Storage Occupancy and Increase Operations Efficiency with IBM zEnterprise Data Compression*, SG24-8259
- *System z End-to-End Extended Distance Guide*, SG24-8047

You can search for, view, download, or order these documents and other Redbooks, Redpapers, web docs, draft, and more materials, at the following website:

ibm.com/redbooks

Other publications

The following publications are also relevant as further information sources:

- *IBM IMS High Performance Image Copy for z/OS, Version 4 Release 2, User's Guide*, SC19-2756
- *IBM Z Planning for Fiber Optic Links*, GA23-1408
- *z/OS DFSMS Advanced Copy Services,* SC23-6847
- *z/OS DFSMSdfp Storage Administration*, SC23-6860
- *z/OS DFSMS Using the New Functions*, SC23-6857

Online resources

The following websites are also relevant as further information sources:

- IBM Z product page:

  https://www.ibm.com/it-infrastructure/z
- IBM storage for mainframe page:

  https://www.ibm.com/storage/mainframe-storage
- IBM Z Community:

  https://www.ibm.com/community/z/

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

REDP-5186-07

ISBN 0738460656

Printed in U.S.A.

Get connected

Redbooks®
ibm.com/redbooks