IBM DS8000 Storage System

This document provides an overview and analysis of the IBM System Storage DS8000. The DS8000 is IBM's high-end enterprise disk system supporting both mainframe and open system attachment. IBM has continued to enhance the DS8000 with new models since it was first announced on October 12, 2004 and became generally available on December 3, 2004. The DS8880 systems announced in October of 2015 feature multiple models that used Power 8 processors. In January 2017, IBM announced the DS8880F, which is the designation for the all flash models of the DS8880 along with Gen-2 Flash Enclosures. The current models are the DS8900F all flash models which uses Power 9 processors and were announced in September 2019 with periodic enhancements continuing.

The major competitive systems that support both mainframe and open systems attachment include the Dell EMC PowerMax, Hitachi Data Systems VSP 5000, and HPE XP7.

Contents

Highlights

- Two processor complexes with System-p POWER 9 processors
- Models
  - DS8910F All Flash System
  - DS8950F All Flash System
  - DS8980F All Flash System
- PCIe Gen 3 adapters and Flash Enclosure connection
- Up to 4.3 TB of processor memory (cache)
- Up to 16 host adapters (HAs) supporting Fibre Channel and FICON host connectivity
- Easy Tier for dynamic tiering of SSDs and HDDs and the Flash Enclosure
  - API for software developers to control data placement
  - Storage Tier Advisor tool for tiering optimization
- Gen-2 Flash enclosure with dual-active PCIe connected controllers that utilize 2.5" Flash modules and custom flash controllers for accelerated performance. RAID card with custom ASIC, redundant for failover.
- Multi-target replication with PPRC licensed feature
- 19" rack with 40U rack expandable to 46U
- Copy Services Manager – automation for advanced copy services and multi-site migration and disaster recover – both mainframe and open
- Mainframe CKD volume of up to 262,668 cylinders
- Up to 65,280 logical devices
- Thin provisioning of volumes
- Wide striping across disks
- Full disk encrypting devices and solid state
- Space Efficient snapshots with FlashCopy SE
- Remote replication with both synchronous and asynchronous protocol
- Encryption of data with key management by System Key Lifecycle Manager
- Management GUI – common interface from XIV GUI
- Support for VMware VAAI and SRM
- OpenStack Cinder driver
- Copy Services Manager software (formerly TPC-R)
  - Multi-site migration and disaster recovery (MF and Open)
- Transparent Cloud Tiering to cloud object storage for ECKD volumes, managed by DFSMShsm
- zHyperLink for low latency connection from system z
- Compression of data transferred over TCPIP links
- Encryption of data at rest and data in flight meeting SP800-131A requirements

Overview of System

The DS8000 series are storage systems that support attachment to mainframe and open system hosts. The DS8900F systems are the latest generation that use POWER 9 technology. The DS8000 consists of a storage unit and one or two Management Consoles. A storage complex is the storage unit and the Management Consoles that manage it.

The DS8000 storage system currently consists of the DS8910F, the DS8950F models with various choices of base and expansion models to address specific configuration needs from a capacity, connectivity, and performance perspective.

Figure 1: IBM DS8900F System

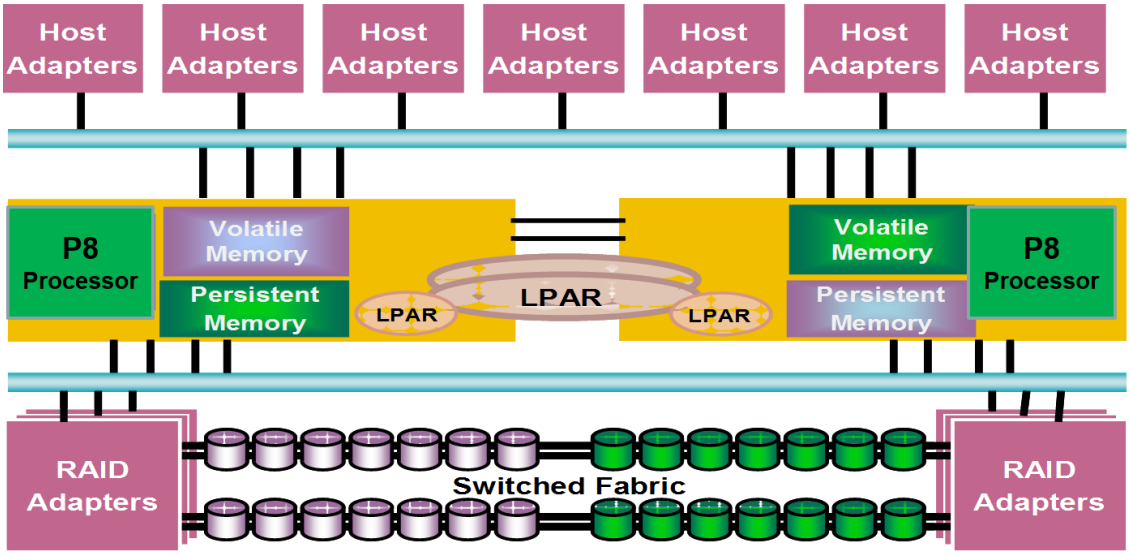

Figure 2: IBM DS8000 Overall Architecture (source: IBM)

DS8900F

The DS8900F series consists of the DS8950F and the DS8910F all flash systems which were announced in September 2019. The DS8900 models use Power 9 processors, unlike the previous models which utilized Power 8 technology. The DS8910F is available both as a rack mountable solution as well as a single rack. The DS8900 systems focus on providing high performance all flash storage while offering features such as transparent cloud tiering, 4 site replication, IBM Storage Insights, and malware protection with Safeguarded Copy.

Figure 3: IBM DS8900F (source: IBM)

Model Overview

	DS8910F	DS8950F	DS8980F
Processor	**POWER9**	**POWER9**	**Power9**
Processor Configuration	16 cores	20 to 40 cores	44 cores
Cache	192GB to 512 GB	512GB to 3.4TB	4.3 TB
Host Adapter Interfaces	16 and 32 Gb/s FC	16 and 32 Gb/s FC	16 and 32 Gb/s FC
Host Adapters	4 to 64	4 to 128	4 to 128
Host Ports	4 to 64	4 to 128	4 to 128
Device Interface	SAS-2	SAS-2	SAS-2
Number Devices, Flash Modules (Cards)	192	384	384
Max Capacity – flash	2.95 PB	5.9 PB	5.9 PB
DeviceTypes	800 GB, 1.6 TB, 1.9 TB, 3.2 TB, 3.84 TB, 7.68 TB and 15.36 TB Flash Cards	800 GB, 1.6 TB, 1.9 TB, 3.2 TB, 3.84 TB, 7.68 TB and 15.36 TB Flash Cards	800 GB, 1.6 TB, 1.9 TB, 3.2 TB, 3.84 TB, 7.68 TB and 15.36 TB Flash Cards

Table 1: DS8900F Model Overview

Hardware Management Console

The IBM System Storage Hardware Management Console (S-HMC) is the center piece for configuration, copy services management, and maintenance activities for the DS8000. It is a dedicated workstation that is installed in the base frame, and consists of a workstation processor, keyboard, monitor, modem, and Ethernet cables. The S-HMC automatically monitors the state of the system and notifies the customer as well as IBM when service is required. It can also be connected to a network to enable centralized management using the IBM System Storage DS Command-Line Interface or the storage management software that uses the IBM System Storage DS Open API.

The architecture is designed for one S-HMC to support up to eight DS8000 systems. IBM recommends having a second S-HMC to provide continuous availability. A second S-HMC can be installed in the base frame for redundancy.

Hardware Architecture

The basic hardware configuration of the DS8900 is illustrated in the following figure.

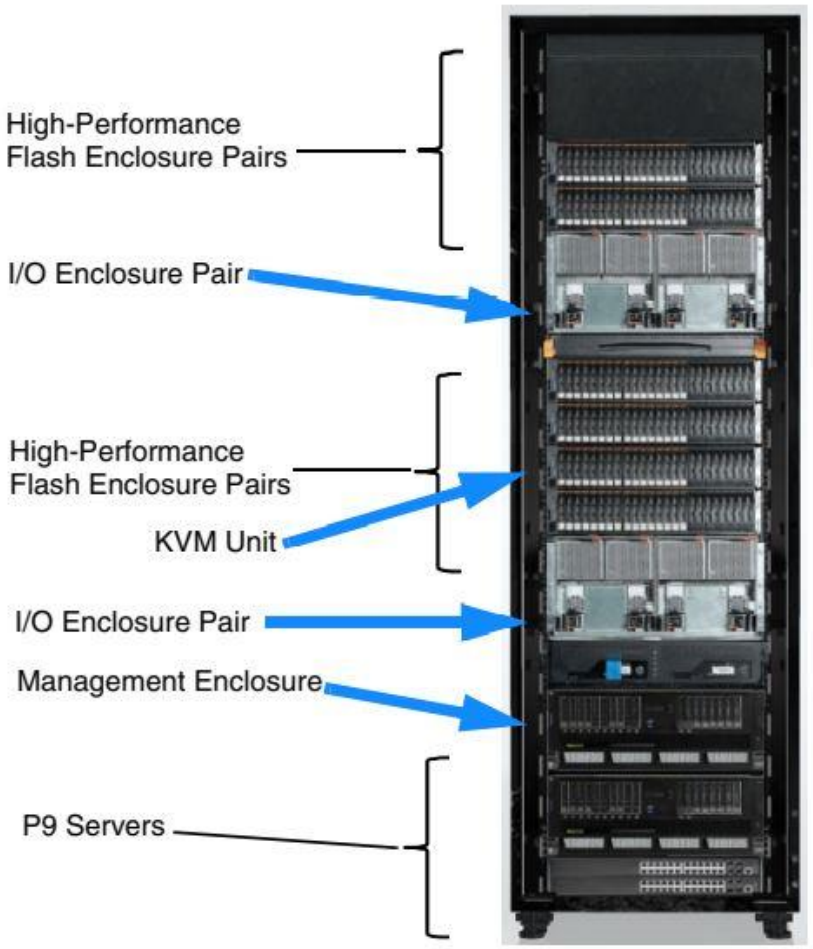

Figure 4: DS8950F Physical Package

The DS8000 consists of two processor complexes, and each of the processor complexes has access to multiple host adapters for host connectivity. A DS8000 can have up to sixteen four-port SAS-2 adapters to connect devices. Flash enclosures are connected directly to the PCIe Gen 3 bus. Each PCIe Gen 3 device adapter connects the complex to two separate SAS-2 disk enclosures with up to 24 2.5" device. A DS8000 system consists of a base frame and, depending upon the model, one to three expansion frames.

Processor Enclosures

All base frames contain two processor enclosures or complexes. The processor enclosures contain the computer electronic complex (CEC).

DS8910F:

- 2 Power 9 processors with 8 cores each
- Up to 512 GB of system memory
- 16 and 32 Gb/s FC
- 16 to 192 flash drives

DS8950F:

- 2 Power 9 processors with up to 20 cores each
- Up to 3.4 TB memory
- 16 and 32 Gb/s FC
- 16 to 384 flash drives

DS8980F:

- 2 Power 9 processors with up to 22 cores each
- Up to 4.3 TB memory
- 16 and 32 Gb/s FC
- 16 to 384 flash drives

Flash Enclosures

Flash enclosures can be added to the DS8000 to accelerate performance with solid state technology.

Figure 9: IBM High Performance Gen 2 Flash Enclosure

The flash enclosure contains two controllers that can be active at the same time and connect to the Power systems with PCIe Gen 3 interfaces. A dual port PCIe-3 x4 adapter is required per flash controller.

The flash controller contains 1.6 TB, 3.2 TB, 3.8 TB, 7.68 TB or 15.36 TB (Gen2) flash modules (cards) which are packaged as 2.5" pluggable devices and are configured with RAID 5 (Gen 1 only) or RAID 6 protection. For Gen2, RAID 6 protection is the default with the option to configure as RAID 5 or RAID 10.

Encryption of data stored in the flash modules is included with the encryption performed by the flash controllers.

Easy Tier supports tiering to the High Performance Flash Enclosures as a tier 0. SSDs installed in the DS8000 are also treated as tier 0 but the auto-rebalance feature of Easy Tier will select the higher IOPS of the flash modules to optimize usage over SSDs.

I/O Enclosures and Adapters

All base frames contain four I/O enclosures. The I/O enclosures hold the PCIe Gen 3 adapters and provide connectivity between the adapters and the processors. Both device adapters and host adapters are installed in the I/O enclosure. The path for adapter to processor complex communications is the RIO-G loop. PCIe Gen 3 adapters are installed to attach Flash Enclosures.

Expansion Frames

The expansion frame is used for additional disk enclosures to scale beyond the number of drives available in the base frame.

DS8000 Expansion Frame

- Up to 816 additional disks may be attached across two expansion frames.
- Additional device adapters are installed in the first expansion frame in an I/O enclosure.

Device Adapters

Device adapters provide the connection between storage devices and the storage facility images through I/O ports. They are ordered and installed in pairs. Each pair supports two independent paths to all of the disk drives that are supported by that pair. The two paths are connected to two different network fabrics for redundancy. The physical links allow two read operations and two write operations to be performed simultaneously around the fabric. Device adapters are PCIe Gen 3 connected to the processor complex.

Host Adapters (HAs)

The DS8000 can support up to sixteen host adapters (HAs) that are connected with PCIe Gen 3. Each can support up to 128 host ports.

The 4-port FC/FICON adapter card is available with 16 Gb/s per second in four and eight port versions. Each of the ports independently auto-negotiates to the link speed of the host. Each of the ports can also

independently be either Fibre Channel protocol or FICON, but cannot be both simultaneously. The personality of each port is changeable via the DS Storage Management GUI. Each port can support up to 509 host login IDs, with up to 8,000 per DS8000. The host adapters can either a FICON personality or an FC personality, which is controlled by firmware that is loaded. FICON adapters support dynamic routing for reducing pathing through switches.

Disk Drive Sets

Disk drives are available in sets, with a set consisting of sixteen physical drives that have the same capacity and the same RPM. The drives can be configured as RAID-5, RAID-6, RAID-10, or a combination. IBM states that RAID-10 can offer better performance for selected applications. The set options are listed below. The actual effective capacity for each of the RAID mode options is available in the manuals and Redbooks from IBM.

RAID Set Organization						
RAID Type	RAID-10		RAID-5		RAID-6	
Disks	3+3	4+4	6+P	7+P	5+P+Q	6+P+Q

Table 2: Disk Drive RAID Set

1. A 3x3 array consists of three data drives that are mirrored to three copy drives. Two other drives in the set are used as spares.
2. A 4x4 array consists of four data drives that are mirrored to four copy drives.
3. A 6+P array consists of six data drives and one parity drive. One other drive in the set is used as a spare.
4. A 7+P array consists of seven data drives and one parity drive.
5. A 5+P+Q array consists of five data drives and two parity drives.
6. A 6+P+Q array consists of six data drives and two parity drives.

Disk Virtualization (called storage pooling in the industry)

The term disk virtualization is used to describe the process of preparing a bunch of physical disk drives (DDMs) to be something that can be used from an operating system.

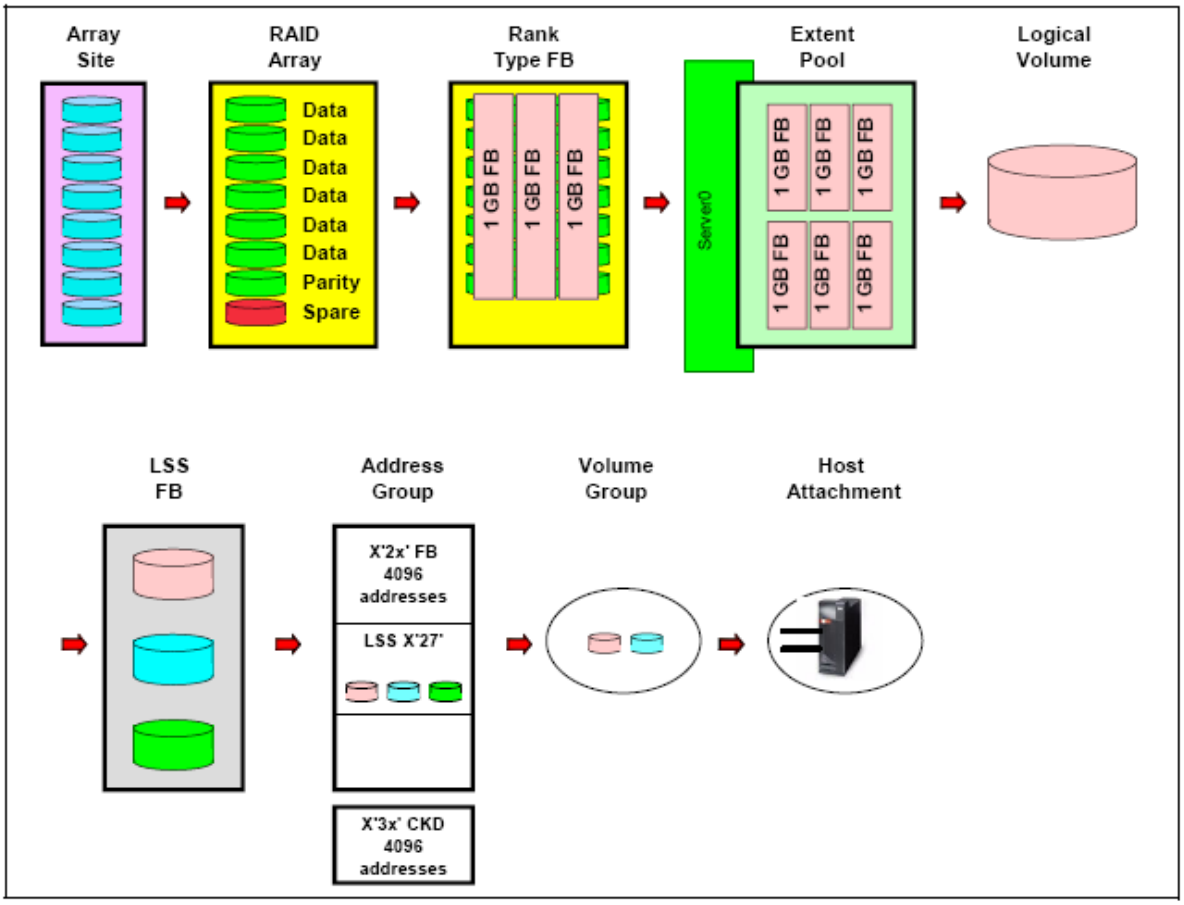

Figure 12: DS8000 Virtualization Hierarchy (source: IBM)

Array Site

An Array Site is a group of eight DDMs. The specific DDMs that make up an array site are pre-determined by the DS8000. All DDMs in an Array Site must be of the same capacity and the same speed (RPM). Array Sites are the building blocks used to define Arrays.

Arrays

An Array is created from one Array Site. The formation of an Array means defining it for a specific RAID type (RAID-5, RAID-6, or RAID-10).

RAID-5

A RAID-5 array is built on one array site with either seven or eight disks. A seven-disk array uses one disk for parity so it is referred to as a 6+P+S (where P stands for parity and S stands for spare). An eight-disk array also uses one disk for parity, but has no spares. It is referred to as a 7+P array.

RAID-6

A RAID-6 array is built on one array site with either seven or eight disks. A seven-disk array uses one disk for parity so it is referred to as a 5+P+Q+S (where P and Q stand for parity and S stands for spare). An eight-disk array also uses two disks for parity, but has no spares. It is referred to as a 6+P+Q array.

RAID-10

RAID-10 combines the features of RAID-0 and RAID-1. RAID-0 optimizes performance by striping data across multiple disk drives at a time. RAID-1 provides disk mirroring by duplicating data between two disk drives.

In the DS8000, RAID-10 is implemented with either six or eight DDMs. If spares exist on the array site, then six DDMs are used to make a three-disk RAID-0 array, which is then mirrored. If no spares exist, then eight DDMs are used make a four-disk RIAD-0 array which is then mirrored.

Spares

The DS8000 microcode determines which sites will contain spares. Normally, the first four array sites will each contribute one spare to the DA pair, with two spares on each loop. Therefore, each DA pair will have access to four spares. A minimum of one spare is created until the following conditions are met:

- Minimum of 4 spares per DA pair
- Minimum of 4 spares of the largest capacity array site on the DA pair
- Minimum of 2 spares of capacity and RPM greater than or equal to the fastest array site of any given capacity on the DA pair

The spares in the DS8000 are “floating spares”. That is when a DDM fails the data it contained is rebuilt onto a spare, then when the disk is replaced, the replacement disk becomes the spare. However, the DS8000 microcode may choose to allow the hot spare to remain where it has been moved or it may choose to migrate the spare to a more optimum position. This is due to balancing the spares across the DA pairs, the loops, and the enclosures.

Arrays Across Loops

Each array consists of eight DDMs. Four DDMs are from the front enclosure in an enclosure pair and four from the rear enclosure. Since the front enclosures are on one switched loop, and the rear enclosures are on a second switched loop, the array is split across two loops. IBM states that the benefit is improved performance.

Ranks

Another logical construct in the DS8000 disk virtualization scheme is a Rank. In the current DS8000 implementation a Rank is built using just one Array.

The process of forming a Rank does the following.

- The Array is formatted for either FB (Fixed Block for open systems) or CKD (Count-Key-Data for System-z) data.
- The capacity of the Array is subdivided into equal sized partitions called Extents.

Extents

The available space on each Rank is divided into Extents, which are the building blocks for logical volumes. An extent is striped across all disks of an array and its size depends on the Extent type (FB or CKD). An Extent for a FB Rank (LUN) has an extent size of 1GB. The extent size of a CKD Rank is 1,113 cylinders (size of 3390 Model 1).

It is possible to define a LUN of less than 1GB, or a CKD volume of less than 1,113 cylinders. However, one extent is the minimum physical allocation unit when a FB or CKD volume is created. Therefore if the defined capacity is not an integral multiple of the capacity of one extent, the unused capacity in the last extent is wasted.

Extent Pools

An Extent Pool is the logical construct to aggregate the Extents from a set of Ranks to form a domain for extent allocation to a logical volume. One or more Ranks with the same extent type can be assigned to an Extent Pool. There can be as many Extent Pools as there are Ranks, but one Rank can be assigned to only one Extent Pool.

The minimum number of Extent Pools is one. However, IBM states that the user would normally want at least two, one assigned to server 0 and one assigned to server 1 so that both servers are active. IBM also states that in an environment where FB and CKD is to be defined, the customer might want to define four Extent Pools, one FB and one CKD pool per server. Additional Extent Pools may be desirable to segregate Ranks with different DDM types.

Dynamic Extent Pool Merge

The Easy Tier advanced feature has a manual mode capability that allows one extent pool to be merged into another extent pool while the logical volumes in both extent pools remain available. Merging of extent pools would be done for consolidation of smaller extent pools to allow a wide distribution of data across disks (increase the striping width) for increased performance and for using a mix of storage technologies (SSD and HDD) within a pool after each has been created independently.

Logical Volumes

A DS8000 can support up to 65,280 fixed block (FB) and/or count-key-data (CKD) logical volumes

Fixed Block Volumes (LUNs)

A logical volume composed of fixed block extents is called a LUN. It is composed of one or more 1 GB extents from one FB extent pool. A LUN cannot span multiple extent pools, but a LUN can have extents from different Ranks within the same extent pool. The maximum LUN size supported by the DS8000 is 16 TB.

LUNs can be allocated in binary GB (2^{30} bytes), decimal GB (10^9 bytes), or 512 or 520 byte blocks. However the physical capacity that is allocated for a LUN is always a multiple of 1GB. Therefore, it is suggested that LUN sizes always have a multiple of a gigabyte. For example, consider a LUN size of 25.5 GB. 26 GB would actually be physically allocated, resulting in 0.5 GB of unusable physical storage.

CKD Volumes

A System-z CKD volume is composed of one or more extents from one CKD extent pool. CKD extents are the size of a 3390 Model 1, which has 1,113 cylinders. The user does not define the number of 3390 Model 1 extents, but the number of cylinders they want for the volume.

The DS8000 will support volume sizes of up to 262,668 cylinders (about 223 GB). However, if the number of cylinders is not an integral multiple of 1,113 cylinders, then some space in the last allocated extent is wasted.

As with FB volumes, a CKD volume cannot span multiple extent pools, but a volume can have extents from different Ranks in the same extent pool.

System-i LUNs

There are some special considerations for System-I (formerly iSeries) LUNs, which are also composed of fixed block 1GB extents. LUNs created on a DS8000 are always RAID protected (RAID-5, RAID-6. or RAID-10). However, the user might want to deceive OS/400 and tell it that the LUN is not RAID protected. This will result in OS/400 doing its own mirroring. System-i LUNs can have the attribute of “unprotected”, which will result in the DS8000 lying to the host and telling it that the LUN is not RAID protected.

System-i LUNs also expose a 520 byte block to the host. Eight of these bytes are used by the operating system, and the other 512 bytes are like other SCSI LUNs.

Space Efficient Volumes (Thin Provisioning)

The implementation of the DS8000 (beginning with the DS8700) of thin provisioning is called Space Efficient volumes. Beginning with the DS8880, the storage pooling and disk virtualization functions were changed to optimize thin provisiong and by default, all volumes will be thinly provisioned with a single volume type. Array group creation and sizes were not changed, however, to maintain the test processes for the DS8000.

With thin provisioning (Space Efficient volumes) space is allocated only when data is actually written to the volume. The granularity of allocation (the amount of space that gets allocated when required) depends on the amount of data and the allocation unit size.

Space reclamation is supported with an RPQ for Symantec Storage Foundation. VxFS will send SCSI Unmap commands to the DS8000 when files are deleted so the space can be reclaimed for the volumes to remain thin. Only AIX and Linux operating systems are supported for this reclamation feature currently.

Wide Striping

Wide striping in the DS8000 (beginning with the DS8700) is accomplished by a featured called Storage Pool Striping: Extent Rotation. When a LUN or volume is created, the extents of a volume can be striped across several ranks. The wide striping will allow for more disks to be active for the volume and provide the opportunity for greater performance. Wide striping is the default condition for the DS8880.

Logical Storage Subsystems (LSS)

Another logical construct is the Logical Storage Subsystem (LSS). Each LSS is defined as either FB or CKD and groups logical volumes in groups of up to 256.

Unlike the early ESS, on the DS8000 there is no fixed binding between any Rank and any LSS. The number of LSSs has also changed. The DS8000 supports up to 255 LSSs versus just 16 for the ESS.

Volume Access

The DS8000 provides mechanisms to control host access to LUNs with the concept of Host Attachments and Volume Groups.

Host Attachment

HBAs are defined to the DS8000 in a Host Attachment construct that specifies the HBA's World Wide Port Names (WWPNs). A set of host ports can be associated through a port group attribute, allowing a set of HBAs to be managed collectively.

A given Host Attachment can be associated with only one Volume Group. Each Host Attachment can be associated with a Volume Group to define which LUNs that HBA is allowed to access.

The maximum number of Host Attachments on a DS8000 is 8,192.

Volume Group

A Volume Group is a named construct that defines a set of logical volumes.

When used in conjunction with CKD hosts, there is a default Volume Group that contains all CKD volumes and any CKD host that logs into a FICON I/O port has access to the volumes in this group.

When used in conjunction with open systems hosts, a Host Attachment object that identifies the HBA is linked to a specific Volume Group. The user defines the group by indicating which FB logical volumes are to be placed into the group.

Disk Virtualization Hierarchy Summary

The DS8000 provides mechanisms to control host access to LUNs with the concept of Host Attachments and Volume Groups.

The virtualization hierarchy implemented in the DS8000 was illustrated at the beginning of this section. The sequence of setting up the virtualization is as follows:

1. An Array Site was transformed into an Array, eventually with spares
2. The Array was transformed into a Rank with Extents formatted for FB or CKD
3. The Extents were added to an Extent Pool that determined which storage server would serve the Ranks and aggregated the Extents of all Ranks in the Extent Pool for subsequent allocation to one or more Logical Volumes.
4. Created Logical Volumes with the Extent Pools, assigning them a logical volume number that determined which Logical Subsystem (LSS) they would be associated with and which server would manage them.
5. The LUNs could be assigned to one or more Volume Groups.
6. The host HBAs were configured into a Host Attachment that is associated with a given Volume Group

Evaluator Group Comment: There just seems to be too much complexity here that is exposed to the administrator. The complexity may be due to the issues supporting both mainframe and open system or due to the number of tuning / performance choices. There are too many choices, which may be confusing or be done incorrectly (sub-optimal). Fewer choices and more automatic tuning would be much better.

Reliability, Availability, Serviceability Features

The Reliability, Availability and Serviceability of the DS8000 storage system is addressed by specific features in the system. The system consists of dual controllers (System-p Servers) which are both active and have built-in retry and failure detection. To detect failure and automatically handle failover and recovery, custom software is executing on both of the controllers. In addition, a hardware watchdog timer is used in each controller for monitoring the other controller.

IBM DS8000 systems maintain reliability, availability and serviceability through a variety of mechanisms:

- Redundant controllers with online serviceability
- Dual hard disk drives for embedded software and configuration information
- Mirrored Write cache with battery backup
- System configuration backup and restore capability
- Redundancy of active components including battery backup units, fans and power supplies
- Non-disruptive software (microcode) upgrades to elements in the system
- Dual access to all front-end and back-end ports
- Redundant paths from controllers to storage devices
- Remote call home
- Error notification and logging
- Optional redundant Hardware Management Console
- Disk scrubbing for detection and repair of defects on disk drives
- Floating hot spare disks
- RAID disk protection

The DS8000 provides a Common Information Mode (CIM) agent that supports the Storage Management Initiative specification (SMIs) which will allow it to be managed by compliant Storage Resource Management software.

Microcode Updates

Most of the components in the DS8000 have microcode (firmware or embedded software) that can be updated. This includes the processor complexes, device adapters and host adapters. Each DS8000 server also has an operating system (AIX) and Licensed Internal Code (LIC) that can be updated. The truth is that it will be necessary to make updates to each of these areas. The frequency of change will probably be higher in the early life of the product and then decrease over time. ***The ability to make these changes in a non-disruptive manner is critical.***

The Storage Hardware Management Console (S-HMC) can contain up to six different versions of code. Each server can hold three different versions (previous, active and next).

System Management

Management options include the Storage Hardware Management Console (S-HMC) for configuration and service, TPC, or SSPC. User roles with a different level of access permissions enable more flexibility allocating cluster access and management authorization. Management connectivity is via one of two, onboard Ethernet ports per node.

SMI-S Interface

The DS8000 includes an embedded CIMOM, used to provide an SMI-S interface for IBM and third party storage management.

SNMP

SNMP and alerts may be configured with the management console to provide enterprise alert reporting.

Email and Logging

Email alerts as well as logging capabilities are also available to report error conditions.

The DS8900 models feature IBM Storage Insights, a cloud based management system. Storage Insights allows users to monitor performance, capacity, and system health all from a single view.

DS8000 Software Architecture

Operating System Support

There is a wide range of host Operating Systems supported.
Currently the following are supported:

- IBM Mainframes OS's – z/OS, z/VM, z/VSE, z/TPF
- MS Windows – Server and Hyper-V
- Linux – Red Hat RHEL, SuSE SLES and others
- Oracle Solaris
- HP OpenVMS
- HP Tru64
- HP-UX
- IBM AIX
- VMware – ESX, ESXi
- IBM OS/400 – System i

FC Multi-Path IO

Multi-path I/O options are supported on multiple OS platforms. The primary method for providing I/O failover and multi-path support is to utilize the IBM SDD driver.
These include the following:

- Solaris (utilizing IBM SDD with Solaris MPxIO driver, or VERITAS DMP)
- Windows Server 2003, 2008 and 2012 (utilizing IBM SDD, or IBM Windows SDDSM and Windows MPIO driver)
- Linux (utilizing IBM's SDD driver with Linux DMMP)
- IBM AIX (utilizing the SDD driver)
- HP-UX (utilizing IBM SDD, HP PVLinks, or the Mass storage stack in HP-UX 11i v3 or later)
- VMware (VMware includes multi-path support, no device drivers are needed for the hypervisor)

Software Features

IBM offers a variety of device specific and heterogeneous software features for the DS8000. Functions include point-in-time copy, remote copy, and PAV.

Data Protection

Safeguarded Copy

Cyber Resilience (CR) is supported by the DS8000 with a feature called Safeguarded Copy where an unalterable copy of data is made that is not host accessible for potential logical corruption (from deletion, encryption, selective manipulation). Safeguarded Copy is different from snapshot (FlashCopy):

- Up to 500 recovery copies can be made in a separate storage location (FlashCopy has a maximum of twelve per volume)
- Any recovery copy is available to recover volumes and the recovery is not done to the source volumes.
- The recovery copy is immutable.
- Recovery copies are not visible to hosts (no device addresses (or UCBs if mainframe) are exposed.

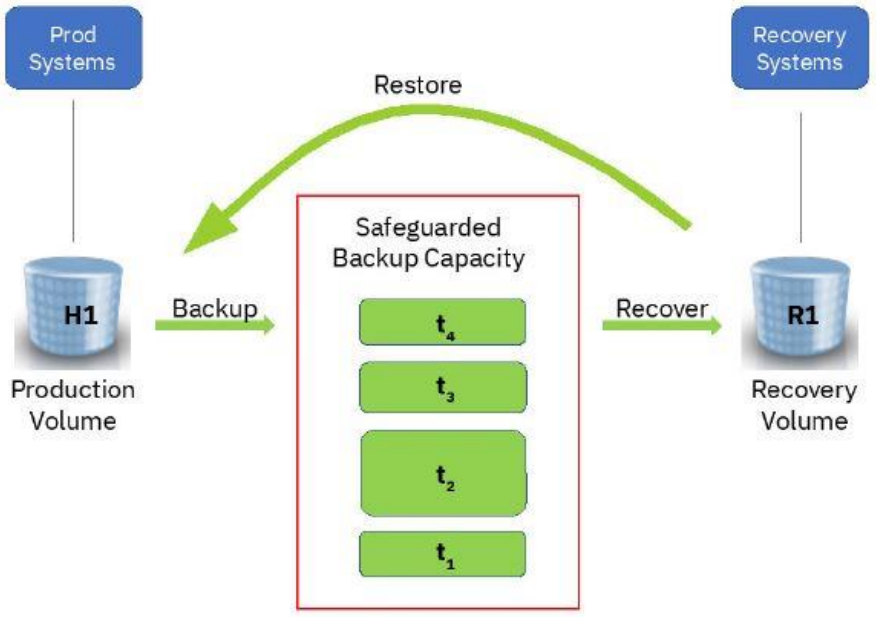

Figure 13: Safeguarded Copy Operation (source: IBM)

IBM Copy Services Manager (CSM) is used by administrators to create, delete, and perform manual recovery of Safeguarded Copies. These copies are called backups in the documentation for CSM. CSM can manage the lifecycle of backups and the retention periods.

Capacity is consumed for the Safeguarded Copies with the following attributes:

- Copies are thin provisioned and allocated as the backups are performed.
- Data is written in 64K blocks (55.3K for CKD).
- Safeguarded Copies are not seen by Easy Tier and are not subject to tiering.
- Consistency Groups are supported.

Safeguarded Copies can be replicated to a second site using standard IBM replication features.

FlashCopy

The FlashCopy feature provides an instant, point in time copy of a source volume to a target volume. The name "FlashCopy" is IBM's term for point in time copy software on various platforms. The implementation of FlashCopy allows point-in-time copies of volumes (LUNs) to be created. The FlashCopy is host independent and can be created for any volume. It is implemented as either a full volume copy or a copy-on-write type, which makes it a space efficient copy. FlashCopy is an extra charge feature. Consistency Groups (where volumes that may be related and must be dealt with as a set to maintain write ordering for dependent writes) may be established for the virtual volumes for which FlashCopy may be used to make a point-in-time copy. This allows the FlashCopy to operate on multiple virtual disks as an atomic operation.

FlashCopy is invoked via software by DFSMSdss, TSO, ICKDSF, z/OS API, DS CLI, and the DS Storage Manager GUI. The user has the option of requesting one or more types of FlashCopy operations.

Full Volume FlashCopy

The most basic type of FlashCopy is "Full Volume", which can be specified as either background copy or nocopy.

Background Copy

As illustrated in the following figure, a background copy operation will result in all of the tracks from the source volume being physically copied to the target volume. Once all of the tracks have been copied, the operation is completed and the relationship between the target and source volume is automatically ended. It can also be explicitly withdrawn by issuing the corresponding commands.

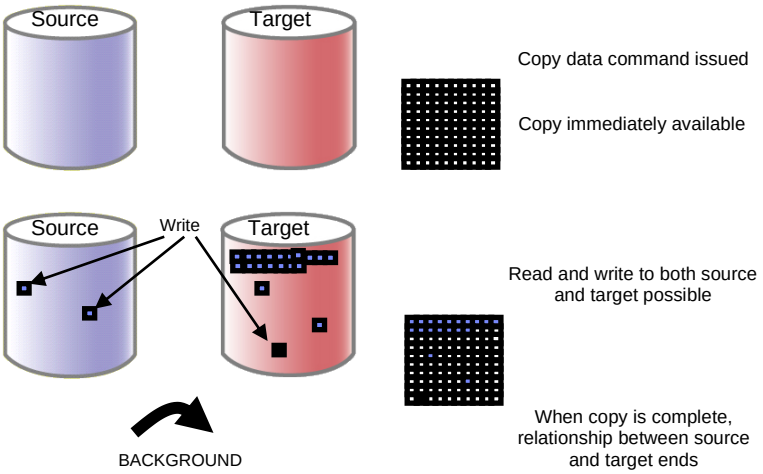

Figure 14: Background Copy Operation (source: IBM)

Background Nocopy

With a background 'nocopy' operation, only those tracks that are updated on the source volume will be physically copied to the target volume. The relationship will continue until it is explicitly withdrawn or until all of the source tracks have been updated.

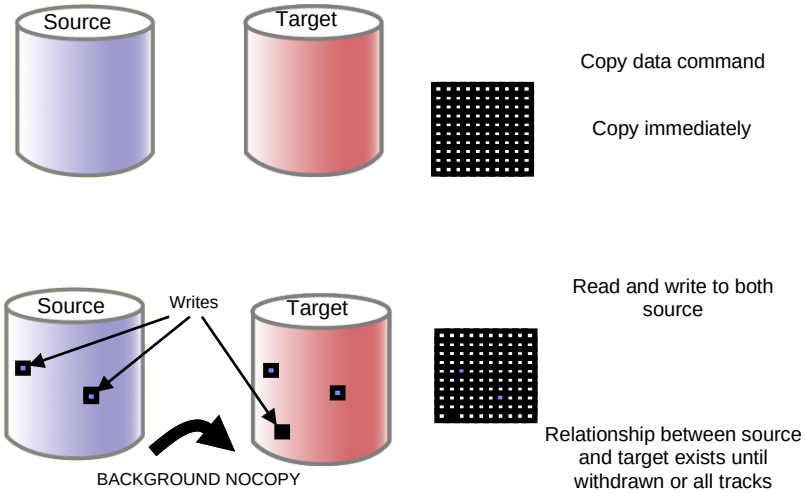

Figure 15: Background Nocopy Operation (source: IBM)

Incremental FlashCopy

Change Recording keeps track of changes made to source and target volumes after establishment of the FlashCopy relationship. The Incremental FlashCopy option provides the ability to only background copy

those tracks that have been changed since the last increment was taken, and to resynchronize a volume or LUN. An incremental relationship is supported on only one target volume/LUN.

Incremental FlashCopy is supported only at the volume/LUN level and is not supported for data set copies.

Persistent FlashCopy

With the Persistent FlashCopy option, the relationship is maintained even after the background copy completes. The relationship must be manually withdrawn.

Data Set FlashCopy

The Data Set FlashCopy function is only supported on z/OS and OS/390 volumes.

Multiple Relationship FlashCopy

Multiple Relationship FlashCopy provides the ability to have up to twelve FlashCopy relationships (targets) to be established to a single source. However, only one of the targets can be defined as incremental FlashCopy. Multiple Relationship is supported for volume and data set level copies.

Consistency Group FlashCopy

FlashCopy Consistency Groups provide a means of achieving consistent copies of dependent data across multiple volumes.

Inband FlashCopy

Inband FlashCopy allows requests to be issued remotely through an existing PPRC link.

Cascading FlashCopy

FlashCopy of Flashcopies or snapshots of snapshots is supported enabling cascading FlashCopies. This allows a target volume or dataset to become a source for another FlashCopy. Up to 12 FlashCopies may be in a cascade.

Reverse Restore

With the Reverse Restore option, the FlashCopy relationship can be reversed. This is accomplished by copying over modified tracks from the target volume to the source volume. The background copy process must complete prior to the reverse restore operation.

Fast Reverse Restore

The Fast Reverse Restore option is used with Global Mirror. It allows the relationship to be reversed without having to wait for completion of the background copy.

Management Options

The following interfaces can be used for FlashCopy management.

- DS Storage Manager (DS SM)
- DS Command Line Interface (DS CLI)
- Total Productivity Center for Replication (TPC for Replication)
- DS Open Application Programming Interface (DS Open API)
- With z/OS the following interfaces can also be used
    - DSFSMSdss utility
    - TSO commands
    - ICKDSF utility
    - ANTRQST application programming interface (API)

Invoking FlashCopy Functions

The following table lists the different FlashCopy functions and the interfaces available to invoke them.

	DS Front-end		zOS Front-end			
	DS CLI	DS SM	ANTRQST	DFSMSdss	ICKSDF	TSO
Full Volume	Yes	Yes	Yes	Yes	Yes	Yes
Multiple Relationship	Yes	Yes	Yes	Yes	Yes	Yes
Incremental	Yes	Yes	Yes	Yes	Yes	Yes
Persistent	Yes	Yes	Yes	Yes	No	Yes
Dataset	No	No	Yes	Yes	No	Yes
Consistency Group	Yes	No	Yes	Yes	Yes	Yes
Reverse Restore	Yes	Yes	Yes	No	Yes	Yes

Table 3: Software Options for Invoking FlashCopy Functions

Performance Considerations

From a performance perspective, IBM has stated the following general recommendations.

- Locate the target volume on the same server as the source volume
- Locate the target volume on a different device adapter (DA) than the source volume
- Locate the target and source volumes on different ranks
- Utilize identical rank geometries (RAID configurations)

Space Efficient FlashCopy - FlashCopy SE

A thinly provisioned volume (called a Space Efficient volume by IBM) only allocates space when data is actually written. FlashCopy SE is a licensed feature that can target space efficient volumes for snapshots (point-in-time copies). A space efficient FlashCopy only uses space for the updates to the source volume, which minimizes the actual space, consumed for the changes to a snapshot of a volume.

For an update to a space efficient volume to occur, the space to write the update must be available when needed since it is dynamically allocated. To ensure the space is available, a repository is created in an extent pool. When data (the updated information for the snapshot) is destaged from cache, the space is allocated from the repository in the extent pool for the volume.

There are many considerations for the usage of FlashCopy SE. IBM recommends that it be used when the retention of the snapshot volume is for a limited amount of time. These considerations need to be investigated before actual usage.

Incremental FlashCopy is not possible with FlashCopy SE. Dataset FlashCopy is not possible with FlashCopy SE.

Multiple Relationship FlashCopy SE

Per IBM: "Standard FlashCopy supports up to 12 relationships and one of these relationships can be incremental. There is always some impact when doing a FlashCopy or any kind of copy within a storage subsystem. A FlashCopy onto a Space Efficient volume has more impact because additional tables have to be maintained. All IBM FlashCopy SE relations are *nocopy* relations; incremental FlashCopy is not possible. Therefore, the practical number of IBM FlashCopy SE relationships from one source volume will be lower than 12. You should test in your own environment how many concurrent IBM FlashCopy SE relationships are acceptable from a performance standpoint." This indicates that more research and testing should be done before implementing FlashCopy SE.

FlashCopy Consistency Groups

As with standard FlashCopy, FlashCopy SE supports Consistency Groups. The Consistency Groups may be used to maintain a relationship between multiple Flash Copies. This option will also ensure write-order consistency among virtual disks within the FlashCopy consistency group.

Replication Options

IBM offers multiple options for replicating data between two or more disk storage systems. They enable the participation in remote mirror and copy options between DS8000 systems.

Metro Mirror (formerly PPRC)

Metro Mirror was previously known as synchronous Peer-to-Peer Remote Copy or PPRC. It provides synchronous replication, or mirroring, of logical volumes between two systems that can be located up to 300 km from each other. IBM states that Metro Mirror is typically used for applications that cannot suffer any data loss in the event of a failure. However, as with any synchronous technology the distance between the primary and secondary systems will determine the impact on application response time.

Forward Error Correction is used for Metro Mirror links to provide better resiliency for connections. Metro Mirror is a hardware based technology that provides synchronous mirroring between two systems that can be located up to 300 km apart. The sequence of operation is illustrated below.

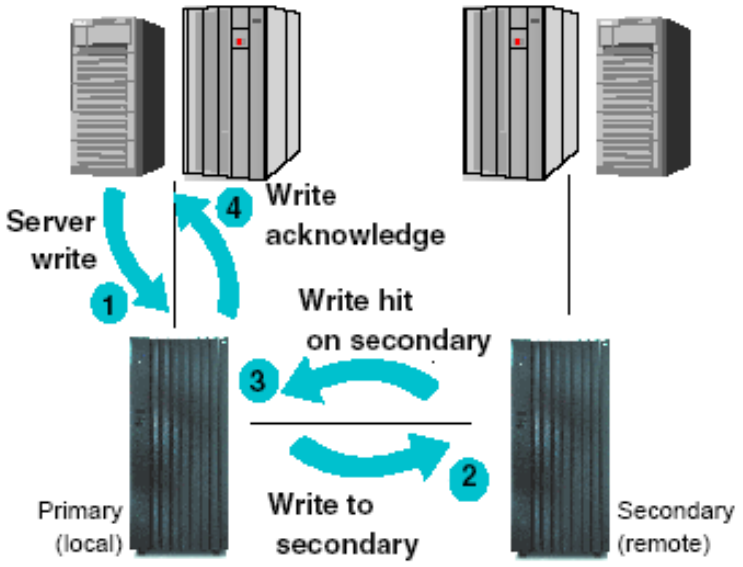

Figure 16: Metro Mirror (source: IBM)

Since it is based upon synchronous operations, the distance between the primary and secondary systems determines the potential impact on application response time.

As of October 2014, multi-target replication is available with PPRC function authorizations (licenses), allowing multi-site mirrors.

The DS8000 supports fibre channel connection only between the primary and secondary systems. It can be direct, through a switch, or through other supported distance solutions (i.e. Dense Wave Division Multiplexor (DWDM).

Metro Mirror Volume States

Volumes that are participating in a Metro Mirror session can be found in any of the following states.

Copy Pending

A volume(s) are in copy pending state after the Metro Mirror relationship has been established, but both the primary and secondary volumes are still out-of-sync. This state can occur immediately after a relationship is initially established, or re-established after being suspended.

Full Duplex

A volume copy pair is in full duplex state when its members are in sync (contain the same exact data). The secondary volume is not accessible when the pair is in full duplex state.

Suspended

The volumes are in suspended state when the primary and secondary systems cannot communicate, or when the pair is manually suspended. During this state a bit map record is maintained of the changed tracks in the primary volume. When the volumes are re-synchronized, only the tracks that have been updated will be copied.

Target Copy Pending

Target copy pending state indicates that the primary volume is unknown or cannot be queried, and secondary state is copy pending.

Target Full-duplex

Target full-duplex state indicates that the primary volume is unknown or cannot be queried, and the secondary state is full duplex.

Target Suspended

Target suspended state indicates that the primary volume is unknown or cannot be queried, and the secondary state is suspended.

Not Remote Copy Pair

Not remote copy pair state indicates that the relationship is not a Metro Mirror pair.

Invalid State

Invalid state indicates that the relationship state is invalid.

Global Copy (formerly PPRC-XD)

Global Copy copies data non-synchronously, and supports distances much greater than is possible with Metro Mirror. The source volume sends a periodic, incremental copy of updated tracks to the target volume. This results in much less impact to the application writes for source volumes and less demand for bandwidth resources. However, Global Copy does not keep the sequence of write operations, and is therefore a "fuzzy" copy. The user can make a consistent copy through synchronization, which is called a go-to-sync operation.

IBM states that Global Copy is appropriate for remote data migration, off-site backups, and transmission of inactive database logs at virtually unlimited distances. It also states that it can be used for application recovery if application I/O can be quiesced and non-zero data loss RPO is acceptable. The operations are illustrated in the figure below.

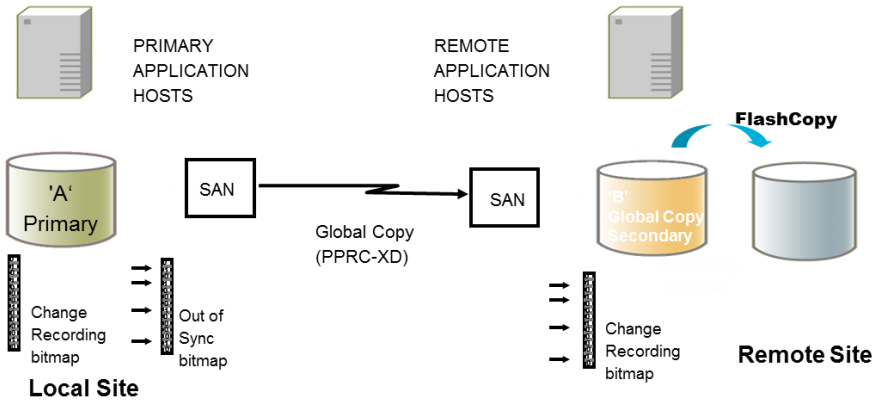

Figure 17: Global Mirror (Source: IBM)

Volume States

The following figure illustrates the basic states and change logic of a volume that is in either a Metro or Global Copy relationship.

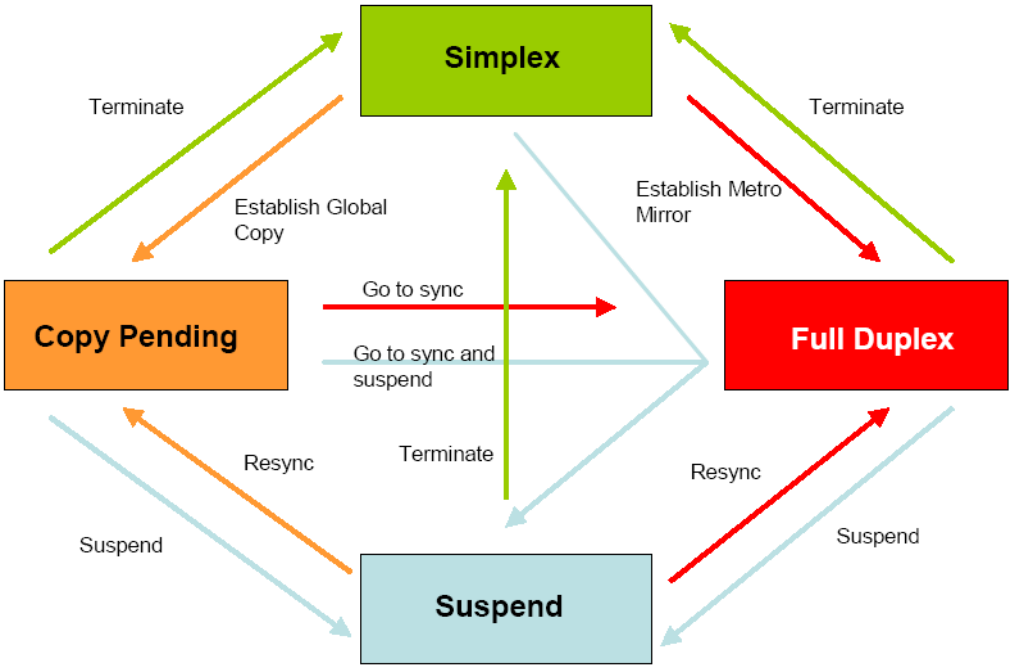

Figure 3: Volume States and Logic Change (Source: IBM)

The following apply to the volume states when the pair is a Global Copy pair.

Simplex
Volume is not in a Global Copy relationship.

Suspended
The writes to the primary are not mirrored onto the secondary volume, and the secondary becomes out-of-sync. A bitmap is kept to record the changed blocks in the primary. When reestablished, only the changed blocks are copied.

Full Duplex
A Global Copy volume never reaches this state.

Copy Pending
Updates to the primary volume are non-synchronously mirrored to the secondary.

Global Mirror (formerly Asynchronous PPRC)

Global Mirror provides a means to have remote copy across two sites over long distances using asynchronous technology. It is based on an integration of Global Copy and FlashCopy. The data written by the host to the storage unit at the local site is asynchronously shadowed to the storage unit at the remote site, and a consistent copy of the data is automatically maintained on the unit at the remote site.

Metro/Global Mirror

Metro/Global Mirror is a 3-site replication solution for both System z and open systems data. Metro Mirror is utilized between the local (site A) and the intermediate (site B), and Global Mirror provides the long distance support between the intermediate and remote site (site C).

The following figure illustrates the processes of Metro/Global Mirror.

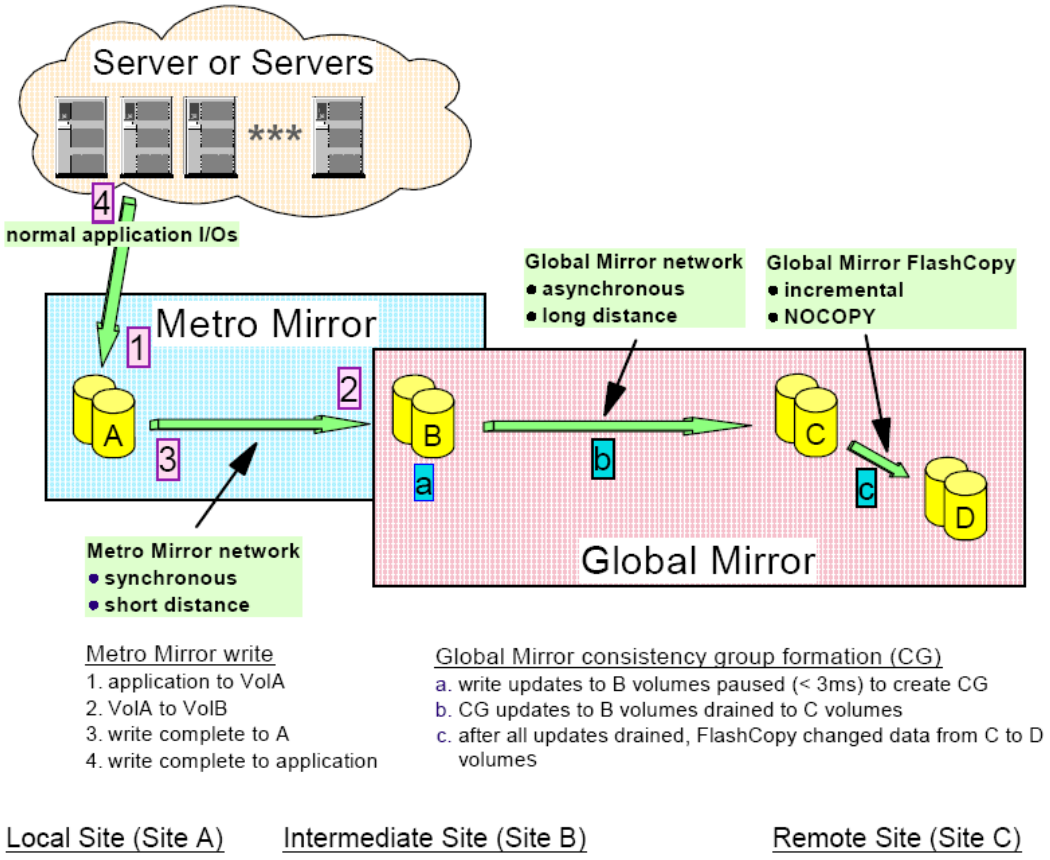

Figure 4: Metro/Global Mirror Processes (Source: IBM)

z/OS Global Mirror (formerly XRC)

z/OS Global mirror, formerly known as XRC, provides asynchronous capabilities and is supported only on System z hosts. It is a software-based implementation and involves a System Data Mover (SDM) that is only found on these servers.

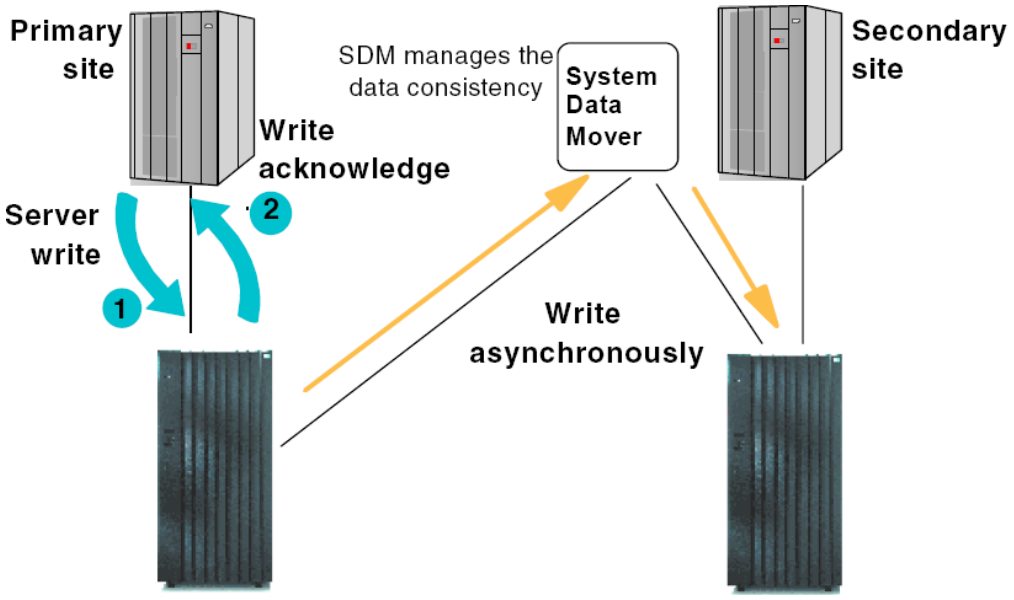

Figure 5: z/OS Global Mirror (Source: IBM)

Two modes of operation are supported: XRC and Migration. In XRC mode it is operating in recovery mode using State, Journal and Control data sets. Secondary volume consistency is guaranteed. In Migration mode, data consistency is maintained within the session but it is not using any Journal or Control data sets, and therefore cannot be used for recovery.

	Metro Mirror	Global Copy	Global Mirror	Metro/Global Copy	z/OS Mirror
Mode of operation					
Asynchronous	No	No	Yes	Yes	No
Synchronous	Yes	No	No	Yes	Yes
Non-Synchronous	No	Yes	No	No	No
Distance	300km	Unlimited	Unlimited	Unlimited	300km
Connecitity					
Fibre Channel	Yes	Yes	Yes	Yes	
FICON	?	?	Yes	?	Yes
Supported Operations					
Migration	Yes	Yes	Yes	Yes	Yes
Recovery	Yes	No	Yes	Yes	Yes
3-site	No	No	No	No	Yes
Consistency Support	Yes	No	Yes	Yes	Yes

Table 4: Remote Replication Summary

Evaluator Group Comment: IBM has noted that the DS8900F family will be the last to support z/OS Global Mirror (formerly XRC). Moving forward new functions for asynchronous replication will be developed only for Global Mirror.

Easy Tier

The Easy Tier function is software that runs on the DS8000 to perform automated tiering of data across internal storage. The Easy Tier software will monitor the storage accesses and move the busiest extents (at the sub-LUN level) to the highest performing tier of storage devices and lesser demanding data to the higher capacity, lower performing storage devices. Easy Tier supports three tiers: the top level tier is flash enclosures or Solid State Devices ant the other two tiers represent different performance (rotation speed primarily) hard disk drives. The size of the tiering element (sub-LUN) is the page size which is equal to the extent size. In the case of the DS8000, the extent size used for Easy Tier is 1GB.

There are two forms of Easy Tier: Manual Mode and Automatic Mode. Manual mode allows the user request dynamic volume relocation to move a volume between tiers and a dynamic extent pool merge to combine extents into a pool. The automatic mode is either on or off and will dynamically manage the capacity allocated to logical volumes in an extent pool for the most appropriate storage by performance. When in automatic mode, the Easy Tier software will monitor the stage/destage activity of each extent allocated to a logical volume and calculate how active (hot) the extent is compared to others. The information is used to generate an extent relocation plan at least once every 24 hours and attempts to place each extent on the appropriate type of storage.

Easy Tier also has a feature to allow administrators and applications to direct data placement on specific storage pools through the Easy Tier Application. Explicit placement requests generate migration operations giving the user controls of data placement for performance reasons.

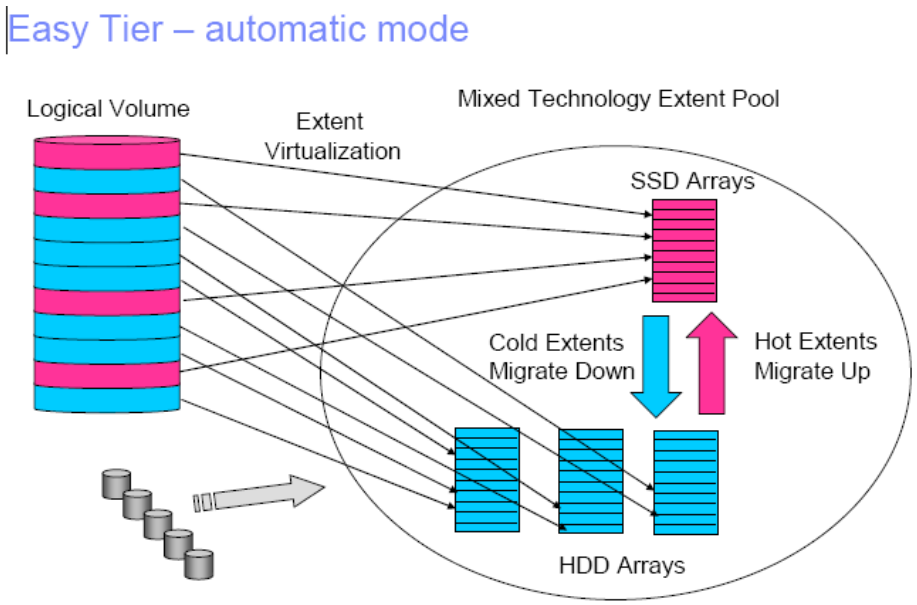

Figure 6: Easy Tier Automatic Mode (Source: IBM)

Storage Tier Advisor Tool (STAT)

The Storage Tier Advisor Tool is software that can be run on a Windows server to analyze the workload running on a DS8000. The software can be downloaded from IBM and requires a DS8000 system with Easy Tier licensed software and the monitor mode activated on the systems using the DS GUI or CLI.

Monitor statistics are gathered every 24 hours and analyzed on the DS8000. The analysis data can be downloaded to the Advisor Tool from which the information can be displayed to see the logical volume extent candidates for automatic migration. The information provided can be useful in determining whether configuration changes such as adding more SSDs may be useful.

The Easy Tier Heat Map can be transferred from a primary site to a secondary site allowing data to be relocated on storage tiers at the second site. With this capability, when a secondary site has to take over because of a failure, the tiering can be in sync with the primary site for performance reasons.

Full Device Encryption

As optional feature, devices that support encryption of data on the device (called Self Encrypting Devices) can be ordered for the system. With SED drives, any drive removed from the system would have the data encrypted and would be unrecoverable by design. The System Key Lifecycle Manager software is used to manage the keys for the encryption of data on drives. Two of the SKLM servers are required for the key management. SED is inclusive of hard disk drive and solid state devices.

zHyperLink

zHyperLink is a technology to accelerate I/O and replication, which is especially useful in all-flash environments in maintaining the low-latency provided by flash. The technology is the use of a new adapter and physical interface to connect the z Series host directly to the DS8000 I/O bays. With a disttance limit of 150 meters, the I/O can be done in less than 10 microseconds. The DS8000 can use local HyperSwap for business continuity with long distance asynchronous replication.

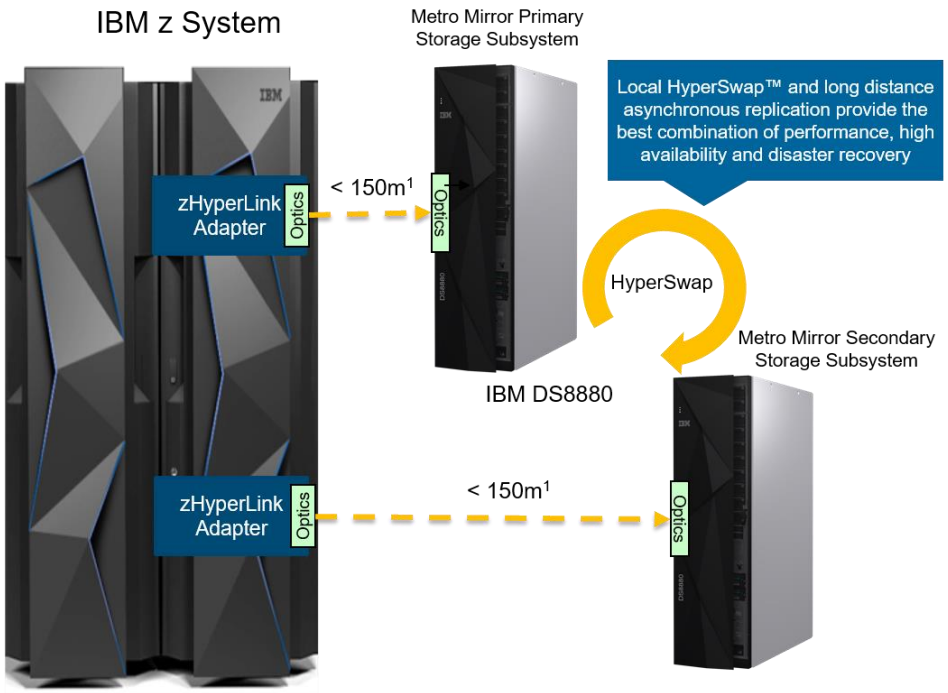

Figure 7: zHyperLink Implementation

Operating Environment License (OEL)

An operating environment license (OEL) feature must be ordered for every DS8000 storage unit, and is a no charge feature. The operating environment model and features establish the extent of IBM authorization for the use of the IBM DS Operating Environment. This operating environment license support function is called the 2244 Model OEL. The OEL licenses the operating environment and is based on the total physical capacity of the storage unit (base model plus any expansion models), and authorizes the customer to use the model configuration at a given capacity level.

Integration with Systems and Software

VMware Support

The IBM DS8000 supports the VMware vStorage APIs for Array Integration (VAAI). VAAI is a set of mechanisms that allow processing for certain data-related services—copying data when creating a new VM, for example—to be offloaded from the ESX host to a storage array. The intent of these APIs is to streamline the functioning of the ESX server and speed-up delivery of storage-supported services.

- **Full copy** — Enables the storage system to make full copies of data within the storage system without having the ESX host read and write the data.

- **Block zeroing** — enables storage systems to zero out a large number of blocks to speed provisioning of virtual machines.
- **Hardware-assisted locking** — The system uses an atomic test and set implementation rather than a SCSI lock/reservation to protect the metadata for VMFS cluster file systems, thereby improving the scalability of large ESX host farms sharing a Datastore.

In addition to the VAAI support, the DS8000 has support for SRM with a Site Recovery Adaptor to communicate between SRM and the DS8000 and implementation of the UNMAP command to release blocks no longer allocated and in use.

A vCenter plug-in is available for reporting information and some management functions.

OpenStack Support

OpenStack is supported by the DS8000 with support for the Cinder driver.

Software Licensing

Beginning with the DS8880, IBM simplified the licensing for DS8000 features. The base function license includes basic features such as thin provisioning, RAID group functions. A separate Z Synergy license adds support for FICON, PAV, and other mainframe specific functions. A Copy Services License provides all the copy services such as FlashCopy and the replication functions.

Copy Services Manager (formerly TPC-R)

Companion software with the DS8000 is Copy Services Manager which provides automation for advanced copy services and multi-site migration and disaster recover – both mainframe and open.

Transparent Cloud Tiering

The DS8000 can automatically move data to or from cloud object storage under control of DFSMShsm. ECKD volumes can be tiered to IBM Cloud Object Storage, IBM Cloud, or TS7700 as objects with transparent movement by the DS8000. The TS7700 can operate as a target with DFSMShsm with transparent cloud tiering. Currently, only simplex devices are supported. Use of Transparent Cloud Tiering reduces the need to use CPU cycles with the DS8880 performing the data mover function. Data integrity is assured with hashcodes and checksums on the data and an audit log of all operations. Data is also encrypted while in flight.

Figure 8: Transparent Cloud Tiering

Performance

Evaluator Group Comment: When properly configured, the DS8000 should be capable of satisfying the performance requirements of the vast majority of environments.

IBM regularly submits products to the Storage Performance Council and publishes results, however no performance data is available for the DS8900F systems at this time.

Real-time monitoring of performance with graphical presentation is available in the GUI.

Evaluator Group believes that the best source currently available for providing comparative performance data is the Storage Performance Council (SPC). IBM is an active member of this organization and has published benchmarks for the DS8000 with Easy Tier and SSDs. The results are available at **http://www.storageperformance.org/results/benchmark_results_spc1*.***

Sequential Pre-fetching in Adaptive Replacement Cache (SARC)

The DS8000 incorporates a self-learning caching algorithm, which IBM refers to as Sequential Pre-fetching in Adaptive Replacement Cache (SARC). IBM states that SARC provides the following.

- Sophisticated, patented algorithms to determine what type of data should be stored in cache based upon the recent access and frequency needs of the hosts
- Pre-fetching, which anticipates data prior to a host request and loads it into cache
- Self-Learning algorithms to adaptively and dynamically learn what data should be stored in cache based upon the frequency needs of the hosts

Adaptive Multi-Stream Pre-fetching (AMP)

Sequential and batch processing has been improved with detection of the nature of the access of data and optimized pre-fetching for sequential data.

Intelligent Write Caching (IWC)

For write performance improvement, the ordering of the writes on destage from cache is managed. IBM has stated that the caching improvement cans double the throughput for random write workloads.

> ***Every vendor claims to have the "best caching algorithms". It will be extremely difficult to prove or disprove any claims in this area.***

Mainframe Performance Options

The DS8000 supports previously introduced performance options for System-z environments. These include FICON connectivity, Parallel Access Volumes (PAV) and Multiple Allegiance (MA), as well as Priority I/O Queuing.

Evaluator Group Comments

The DS8000 systems provide all of the basic requirements of an enterprise level storage system in the areas of capacity, connectivity, availability, and performance. They also address requirements for advanced feature and functions such as point-in-time copy and remote copy.

The DS8000 is a mature system that IBM has continued to add the latest technologies and advanced features in creating new models. The DS8900F models incorporate many new elements to remain competitive in the high-end enterprise space.

As a high-end system that provides both mainframe and open systems connectivity, the DS8000 is one of but a few products that can fulfill the requirements. This is very important for IBM as it continues to market and support System-z with continued hardware and software advances.

Evaluator Group's perception of the strengths and potential concerns of the DS8000 are as follows.

Perceived Strengths

- ***The ability to simultaneously support both mainframe and open connectivity.***
- ***Advanced features for data protection (remote copy, point-in-time copy).***
- ***IBM commitment – examples include PCIe Gen 3 adapters, added support for Solid State Drives, and the implementation of automated tiering – Easy Tier***
- ***Lower cost structure from using "off-the-shelf" components (System-p)***
- ***The addition of the flash enclosure with flash module technology will provide significant performance acceleration for the DS8000 and will extend the potential lifespan of the system. Continued investment as evidenced by the 2nd generation of flash modules increases the value with greater performance and capacity at a lower cost.***
- ***The additions made with the DS8900 provide major performance gains and cost reductions – including improvements in licensing practices. Customers should be assured of IBM's commitment to the DS8900.***

Perceived Potential Concerns

- ***Scaling of performance – having two Power servers for the controllers sets a limit on the number of nodes (controllers) although the performance of the Power technology greatly exceeds single node Intel server-based products.***
- ***The Thin Provisioning implementation on the DS8000 does not provide an automatic means to reclaim capacity that has been deleted or released. Currently an RPQ for Symantec Storage Foundation to reclaim capacity for Linux and AIX but there is no automatic method for other systems. This may cause thin volumes to become fat over time. Running additional software to perform this function is not the best solution.***

More detailed information is available at http://evaluatorgroup.com