WHITE PAPER

IBM DS8870 in the mainframe market

An Analyst Product and Market Review

By Josh Krischer,
Josh Krischer & Associates GmbH
May 2013

Table of Contents

IBM DS8870 in the mainframe market

An Analyst Product Review

Josh Krischer is an expert IT advisor with more than 43 years of experience in high-end computing, storage, disaster recovery, and data center consolidation. Currently working as an independent analyst at Krischer & Associates GmbH, he was formerly a Research Vice President at Gartner, covering enterprise servers and storage from 1998 until 2007. During his career at Gartner responsible for high-end storage systems and spoke on this topic at a multitude of worldwide IT events, including Gartner conferences and symposia, industry and educational conferences, and major vendor events.

Executive Summary

Much is being published on various aspects of storage systems, their features, connectivity to servers, etc. Never the less very little is written on the storage platform on which the vast majority of world's crucial information is stored. Ninety-six of the world's top 100 financial institutions, most of the largest industry corporations, government organizations and many other entities, use IBM System z as their major IT platform. This paper is about storage for IBM System z.
Three vendors manufacture high-end enterprise storage systems that connects to System z: EMC with VMAX, Hitachi Data Systems with VSP (also OEMed by *Hewlett-Packard* Company), and the IBM with the DS8000 series. IBM is the architecture owner, and the two other companies provide System z compatible products. All the systems above are multi-platform products; however this paper will concentrate on features support, synergy, and compatibility with System z.

Storage Requirements and Challenges

Despite the weak world economy, the market for storage systems is not showing any indications of slowing down. It shrunk a little in 2009, but fully recovered in 2010 and 2011. According to an IDC report, in 2012 external disk storage capacity rose 27 percent over the previous year to reach more than 20 Exabytes, while total external storage revenue rose 4.7 percent to $24.7 billion.

Availability and Business Continuity remain on top of the list of requirements. The nonstop global economy, fierce competition, and new levels of service requirements raise the requirements for business continuity. In the last decade, midrange storage users requested advanced functionality and flexibility, which allow them to increase utilization, decrease storage management efforts, and deploy disaster recovery schemes.

Scalability is a must. A storage system should be able to scale seamlessly in capacity, connectivity or performance, without decreasing service levels. It should support multi-tier storage media including Solid State Drives (SSDs), performance Hard Disk Drives (HDDs) in different capacities (Fibre Channel or SAS), and large capacity near-line Serial-

Attached SCSI (SAS) or Serial Advanced Technology Attachment (SATA)] disks. In simple words, scalability should support tiered storage “in a box.”

Performance has two aspects: the first is the throughput measured in number of I/Os per second (IOPS), and the second is the response time measured in milliseconds for hard drives. Performance should meet Service Level Agreement (SLA) requirements, regardless of the used capacity and the workload. Erratic performance levels irritate users more than a slightly slower but constant response times.

Advanced Functionality is required for better resource exploitation and to address storage management. The average organization’s storage capacity grows by 20-40 percent per year while the size of the storage management staff generally remains the same. Data centers which supported terabytes at the end of the previous decade today support petabytes. The only way to cope with this capacity explosion is sophisticated functionality, advanced automation, and user-friendly management tools and interfaces.

Storage Efficiency in usage and energy consumption is equally important. It can be achieved with thin provisioning, tiered storage, automated data placement, deduplication, compression, small-form-factor HDDs, SSDs and virtualization. Efficient storage systems provide better storage utilization, which translates to lower capacity and lower capital and operational expenditure (CAPEX and OPEX), reduced floor space requirements, lower energy consumption, and improved administrators’ efficiency.

The three-level processing structure and the powerful POWER7 processors make the DS8870 the performance leader. The synergy between the DS8870 and System z brings additional performance improvements. The full redundancy design, embedded encryption, powerful data mirroring techniques and other high reliability design consideration ensure business continuity and data protection. Building the storage hierarchy with automated data tiering, small factor disk, near-line capacity disks SSDs and a user-friendly GUI contribute to storage efficiency.

IBM's System Storage DS8870 series is a stable multiplatform high-end storage system able to support current and future user requirements. It uses state-of-the-art technology to achieve top performance levels and supports superior local and remote data mirroring techniques to ensure business continuity. The DS8870 offers unique features to create synergy with System z such as Performance –DB2 –List Prefetch Optimizer with zHPF. The DS8870 has all the required characteristics to be positioned on the top of enterprise storage systems for all platforms and in particular for System z.

IBM DS8870 Structure

Processors

IBM’s System Storage DS8000 was announced in October 2004. This disk system uses a storage architecture based on standard components such as the IBM POWER processors also used in IBM Power Systems (formerly pSeries) (see Fig.1). The DS8000 was a follow-on to the Enterprise Storage Server (ESS) – codenamed "Shark", which was announced in July 1999. The DS8000 series uses three levels of processing: PowerPC and application-specific integrated circuits (ASICs) in the front and the backend adapters

and POWER7 as the main controller in a server-symmetrical multi-processing (SMP) design. The latest model of the IBM System Storage DS8870 series uses dual multi-core processors SMPs. The POWER7 chips are (3.55 GHz) 2, 4, 8 and 16 core processors supporting simultaneous multi-threading (SMT). One of the enhancements of the POWER7 processor is the Simultaneous Multi-Threading (SMT4) mode. SMT4 enables four instruction threads to run simultaneously in each POWER7 processor core to maximize its throughput.

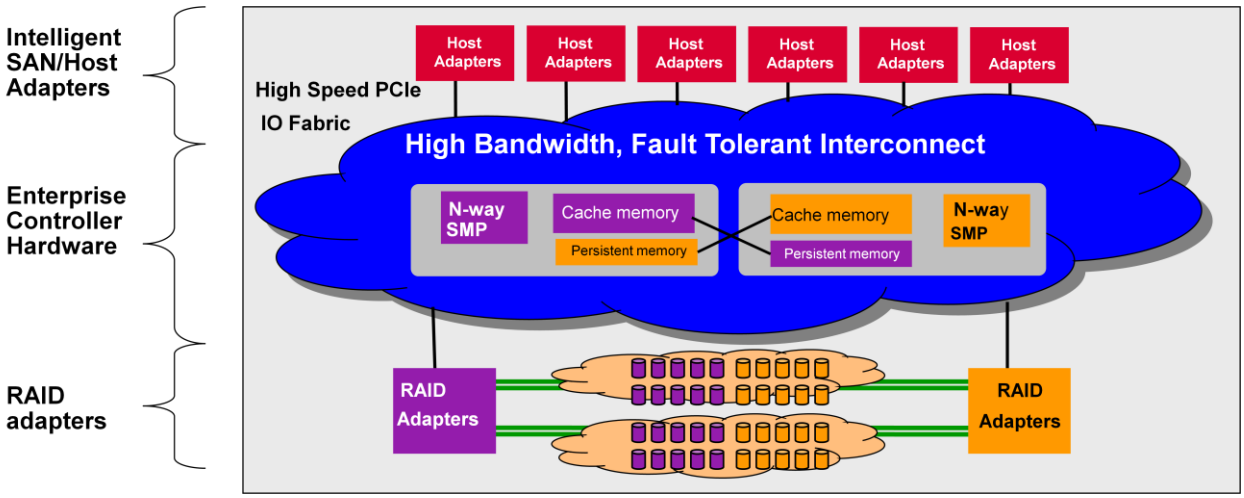

Figure 1. DS8870 High-end Storage System Block Diagram. Source IBM

These multithreading capabilities improve the I/O throughput of the DS8870 storage server. As opposed to common applications processing, where execution of an instruction may depend on a previous result, in the storage control unit the tasks are almost completely independent, which contributes to very high utilization of the multi-core, multi-thread processors.

Cache and Non Volatile Storage

Instead of using a dedicated cache, the DS8000's series cache is allocated as part of the IBM Power server memory. The Power server has a two-level cache (L1 and L2) in addition to its main memory, which creates three levels of memory hierarchy. The tightly clustered SMP, the processor speed, the L1/L2 cache size and speed, and the memory bandwidth deliver better performance in comparison to a dedicated, single-level cache. Each side of the cluster has its own cache and the Non-Volatile Storage (NVS) on the other cluster side to provide better data integrity protection in case of a side failure. During normal operation, the DS8000 series preserves a copy of writes in cluster cache and a redundant copy of writes using the NVS copy in the alternate cluster. In case of extended loss of external power, batteries provide sufficient power to de-stage write data from cache to special disks reserved for that purpose. This “cross connection” protects write data against loss in case of power loss or

“The DS8000 series uses three levels of processing PowerPC and ASICs in the front and the backend adapters and POWER7 as the main control in a server-symmetrical multi-processing SMP design.”

other malfunction, even when one of the cluster sides suffers from a severe malfunction. The effective cache size approximately equals the installed capacity. The DS8000 series uses 4 Kbyte cache slots, which results in improved cache utilization compared to larger cache slots used in competitive high-end disk systems. It improves performance in particular for On-line Transaction Processing (OLTP) of databases, mail servers and most of the Enterprise Resource Programs (ERPs).

Cache Management Algorithms

In July 2007 IBM shipped a new microcode release with a significant improvement to the cache pre-fetching algorithm. The new pre-fetching sequential algorithm is called Adaptive Multi-stream Pre-fetching (AMP). AMP is applied on top of Sequential Prefetching in Adaptive Replacement Cache (SARC) [1] and significantly improves performance for common sequential and batch processing workloads. AMP optimizes cache efficiency by incorporating a self-optimizing pre-fetching algorithm, which is very responsive to changing workloads and delivers up to a two-fold increase in the sequential read capacity of RAID-5 arrays, including for smaller-than-maximum configurations. This technique also further reduces “pollution” of the cache with unnecessary data compared to usual pre-fetch algorithms.

Another feature is the Intelligent Write Caching (IWC), a.k.a. Wise Ordering for Writes. IWC optimizes management of write cache, helping to increase write I/O throughput to disk and reduce contention between read misses and write de-stages, thus enhancing overall I/O performance.

Front-end

The DS8000 supports from 2-to-32 host adapters (installed in pairs) and up to 128 Fiber Channel/IBM FICON host ports (four or eight host ports per adapter). Ports can be configured as 4 Gbps or 8 Gbps Fiber Channel host connections. This option allows connection flexibility and lower cost by saving ports. The ports can independently auto-negotiate to transfer data to match link speed depending on the host HBA. These ports are initially defined as switched point-to-point Fiber Channel protocol (FCP) however, each port can also be set to either FCP or FICON.
The adapter itself is driven by PowerPC processor and function-rich, high-performance, specialty ASICs installed in a PCIe Gen 2 slot.

Back-End

The system can support up to 16 4-port, 8 Gbps Fiber Channel adapters. The connection with the HDDs/SSDs is through 6 Gbps point-to-point switched SAS-2 connection to an 8 Gbps Fiber Channel backbone.

The host adapters include PowerPC processors and ASICs. IBM, which is also one of the biggest semiconductor manufacturers, leveraged this strength by using function-rich/high

[1] Sequential Prefetching in Adaptive Replacement Cache (SARC) algorithm, which is a self-tuning, self-optimizing cache management technique for a wide range of sequential or random workloads

performance ASICs. These ASICs are responsible for managing, monitoring, and rebuilding the RAID arrays. The symmetric dual path structure allows access to the disk enclosures from two independent networks, providing four access paths to each disk drive.

Scalability

The DS8870 [currently] supports from 8 to 1,536 (small-form factor 2.5 inch) or from 8 to 768 (large-form factor 3.5 inch) HDD or SSD disk drive modules (up to 2,304 TB). Six different disk types (in mixed configurations) are supported using RAID 5, RAID 6 and RAID 10 techniques (same disk sizes have to be used in a RAID group):

- 400 GB SSDs
- 146 GB (15k rpm)
- 300 GB (15k rpm)
- 600 GB (10k rpm)
- 900 GB (10k rpm)
- 3 TB (7.2k rpm 3.5") Near-Line disks.

Functionality

General Functionality

The DS8870 supports variety of advanced functions, including thin provisioning, advanced disaster recovery solutions, business continuity solutions, and advanced copy functionalities. All disk drives in the DS8870 storage system support the Full Disk Encryption (FDE) feature. Because encryption is done by the disk drive, it is transparent to host systems, and can be used in for all platforms – z/OS, AIX, Linux, Windows, etc..Also, since each drive has its own encryption engine there is no performance degradation with added capacity. The Storage Pool Striping feature distributes a volume's or LUN's data across many RAID arrays and across many disk drives, which helps maximize performance without special tuning and greatly reduce hot spots in arrays. The IBM EasyTierfeature can automatically move the data and optimize the use of each storage tier, particularly SSD drives. Data areas that are accessed frequently are moved automatically to higher tier disks; for example, to SSDs. Infrequently accessed data areas are moved to lower tiers such as Near-Line SAS drives without manual intervention.

System z functionality and the synergy between DS8870 and the System z

IBM owns the System z mainframe architecture, which, from time to time, allows the company to offer exclusive features for at least a limited period. EMC and Hitachi purchase the specifications for such exclusive features to remain compatible. IBM provides EMC and Hitachi with technical specifications[2] of these features, but not the code itself, which the competing companies have to develop. Thus, for example, in October 2012 EMC Corporation extended its longstanding technology licensing

[2] What is licensed are specifications for interfaces, e.g., I/O commands to invoke a function. That does not include implementation information such as optimizations developed by IBM to support functions or design elements that are included to enable planned future enhancements.

agreement with IBM for EMC storage interoperability with IBM zEnterprise environments to 2017.
Typically, the other vendors offer their own version of the features after 12–24 months, but sometimes this takes much longer, and some features are not supported as of today. See Chart 1 (below) support status of selected System z features. The most important System z features are:

- **High Performance FICON (zHPF) –** zHPF offers high-performance data transfer, reduces the command overhead and thus better utilizes existing bandwidth. It has four functions, each of which requires cooperation between the System z server and the storage system. Selected zHPF functions are: multitrack (allow reading or writing more than two tracks worth of data by a single transport mode operation), extended distance, format writes, QSAM, BSAM, BPAM, and DB2 list prefetch.
- **Parallel Access Volumes (PAV)** and **HyperPAV** allow using multiple devices or aliases to address a single ECKD disk device.
- **I/O Priority Manager s**upports "importance" and "achievement" information provided by z/OS Workload Manager to manage execution priorities. The I/O Priority Manager Quality-of-Service (QoS) function integrates with zWLM and allows automated I/O priority management for mainframe applications. The zWLM monitors every I/O and manages I/O QoS through the Storage System based on zWLM Service Class. zWLM has the ability to provide information to the DS8870 I/O Priority Manager to allocate the required resources based on required goals. Thus, zWLM can now manage host and storage resources end–to-end to optimize the workload execution based on the specified Service Class.
- **Performance – DB2** Specialized cache algorithm can optimize DB2 list prefetch operations by multiple, parallel data fetches. Using simple FICON, the storage system would read a single page per protocol exchange. zHPF reduces the Host-to-Storage System I/O protocol to a single protocol exchange for the full I/O CCW chain. DB2 List Prefetch Optimizer with zHPF, enables the storage system to read all 32 DB2 pages in parallel, transferring all 32 pages back to the host in a single exchange. At any point in time, DB2 has two List Prefetch I/Os outstanding; therefore the DS8870 is always reading 64 DB2 pages in parallel.
- **Performance – IMS** provides enhanced performance for IMS write-ahead data set (WADS).
- **Performance – zDAC** supports optimization to improve performance of z/OS Discovery and AutoConfiguration (zDAC).
- **Volume Management** supports dynamic volume expansion for standard (thick) 3390 volumes, Extended Address Volumes (EAV) – supports 3390 volumes up to 1 TB capacity.

Supported feature	EMC[3]	HDS/HP[4]
HyperPAV	Yes	Yes
High Performance FICON (zHPF basic)	Yes	Yes
zHPF – multitrack	Yes	Yes
zHPF – QSAM, BSAM, BPAM	No	No
zHPF – format writes	No	No
zHPF – DB2 list prefetch cache optimization	No	No
Performance – DB2 (cache algorithm)	No	No
Performance – IMS	No	Yes
Performance – zDAC	No	Yes
Volume Management	No	No
Multiple Readers for z/GM	No	No
Large 3390 Volumes /EAV) 1 TB	No	NO, SOD
zWLM integration for I/O Priority Manager	No	No

Chart 1: EMC VMAX 40K and Hitachi VSP subsystem support matrix

DS8000 series Local and Remote Copy Techniques

The DS8000 storage systems support wide range enterprise-level replication techniques for System z and other platforms to fulfill any requirements. All these techniques support data consistency. It supports:

Local replications:

- FlashCopy and FlashCopy SE
- Remote Pair FlashCopy (Preserve Mirror)

Remote Mirror and Copy:

- Metro Mirror
- Global Copy
- Global Mirror
- Global Copy Metro/Global Mirror
- z/OS Global Mirror
- z/OS Metro/Global Mirror

Local replications

Standard **FlashCopy** is a Point-in-Time (PiT) technique that uses a normal volume as target volume. This target volume must have equal or larger capacity. The space is fully allocated in the storage subsystem.

FlashCopy Space Efficient (SE) is a so-called “snapshot” technique that creates a virtual volume. At the creation, no space is allocated for this volume. Space is allocated just for updated tracks only when the source or target volumes are written.

Remote Pair FlashCopy a.k.a Preserve Mirror transmits the FlashCopy command to the remote site if the target volume is mirrored with Metro Mirror.

[3] To the best of our knowledge. EMC refused to validate this function support list
[4] This column was verified by HDS product management

Remote Mirror and Copy

Metro Mirror, previously known as PPRC, provides synchronous mirroring of logical volumes between two DS8000s. DS8870 supports up to 300 km separation between mirrored systems; however, due to cost of fiber links and the line latency (1 msec/100 km), it is seldom used for distances exceeding 50 km. In this technique the I/O operation is completed only after the primary subsystem receives acknowledgement that the update has been stored on the secondary.

Global Copy, previously known as Peer-to-Peer Remote Copy eXtended Distance (PPRC-XD), transfers data to the secondary without making the primary wait for copy completion before a host write operation is acknowledged. This technique does not ensure data consistency; therefore it is not suitable for disaster recovery and is mainly used for data migrations.

Global Mirror is a two-site, long distance, asynchronous, remote copy technique. It "marks" the updates in "track bit tables" and transfers them periodically to the secondary site. Typically the remote site lags three-to-five seconds behind the local site depending on connection bandwidth and workload characteristics, which minimizes the amount of data exposure in the event of an unplanned outage. A consistent copy of the data is automatically maintained and periodically updated using FlashCopy on the storage unit at the remote site. Up to 32 Global Mirror hardware sessions can be supported within the same DS8870.

Metro/Global Mirror is the cascaded combination of the two techniques as a three-site, multi-purpose, replication solution. The local site is connected to the intermediate site by Metro Mirror so the synchronous replication data loss is minimal – Recovery Point Objective (RPO) is close to zero. This provides recovery from local disasters such as loss of power or fire. The intermediate site is connected to the remote site with Global Mirror to support long-distance disaster recovery and protects against regional disasters such as floods and hurricanes. This solution is popular among financial institutions.

z/OS Global Mirror, previously known as eXtended Remote Copy (**XRC**), is asynchronous mirroring for z/OS operating systems. In 2007 IBM announced the **Multiple Readers for IBM System Storage z/OS Global Mirror**. The XRC was one of the first remote copy techniques introduced more than a decade-and-a-half ago. With this technique, data modifications are temporarily stored in a side-file of the cache and retrieved periodically by the host based System Data Mover (SDM). There is only one *reader* per SDM. Hence, if "emptying" of the side-file falls behind the new modification pace, the host performance may suffer, and in the worst case, trigger a suspension of remote copy operation.

Since the initial introduction of z/OS Global Mirror, many changes have been made in the storage landscape, including much larger disk capacities and the ability to execute many more I/O operations per second thanks to the Parallel Access Volumes (PAV) and Multiple Allegiance (MA) features. The multiple readers divide the side-file into multiple "sub side-files" and allow parallelism for the SDM when emptying these sub-side-files.

The users of z/OS Global Mirror with DS8000 will benefit from improved performance and fewer disruptions under heavy write load conditions and, as a result, experience significantly better performance, in particular in busy z/OS environments. EMC announced a compatible feature in 2011 which is similar but not the same.

z/OS Metro/Global Mirror is a three-site configuration. The primary is connected with the secondary site (in metropolitan distance) using Metro Mirror and to tertiary site by z/OS Global Mirror.

RAS and Disaster Recovery features

Reliability, availability, and serviceability (RAS) played important roles in the design of the IBM System Storage DS8870 structure and components. The DS8870 is based on a redundant cluster of POWER7 servers. Also all other components are designed to ensure full redundancy. The technology sharing between the DS8000 storage and IBM Power servers brings manufacturing costs and RAS advantages. The DS8000 series is actually the largest user of IBM POWER processors. By using the same components, IBM leverages economies of scale of larger production volumes. The collected field experience from thousands of servers helps to improve performance and reliability.

Geographically Dispersed Parallel Sysplex (GDPS)

System z supports Parallel Sysplex as a local or remote cluster. Up to 32 local or remote mainframes can participate in a single cluster. System z Parallel Sysplex also works in conjunction with IBM's disaster recovery software, called Geographically Dispersed Parallel Sysplex (GDPS). GDPS enables automated complete site fail-over with no or minimum loss of data. IBM's GDPS for System z is a multi-site application availability solution, with fast recovery time and highly-automated control. It manages application availability in and across sites for both planned maintenance and unplanned situations, such as a site failure or full-blown disaster.

In June 2011 IBM extended the zEnterprise's business resiliency via significant enhancements to GDPS, in particular an active/active configuration (in addition to the active/standby, which is commonly used). The GDPS active/active continuous availability is the next generation of GDPS and a fundamental shift from a failover model to a near-continuous availability model. IBM intends to deliver, over time, additional configurations that comprise GDPS active/active continuous availability – a solution for organizations using two sites separated by unlimited distances, running the same applications, and having the same data with cross-site workload monitoring, data replication, and balancing.

HyperSwap

Another availability function is HyperSwap. This function is probably the most important business-continuity and availability improvement for IBM mainframes and Power servers. While entire site outages are rare, hardware failures are more common.

With the current integrated and complex application environments – assuming a highly-available, data-sharing Parallel Sysplex environment – storage system becomes a single-point-of-failure for the entire Sysplex. HyperSwap, which is used by multiple GDPS solutions, is controlled by GDPS automation and can eliminate an outage caused by planned maintenance or disk failure by reducing the time needed to switch disks between sites to a matter of seconds and allowing the primary site to use the secondary site's disk storage systems.

"Another availability function is HyperSwap. This function is probably the most important business-continuity and availability improvement for IBM mainframes and Power Servers."

Basic HyperSwap between two remote or locally-installed storage systems provides automated fail-over for planned or un-planned outages and can be deployed with z/OS alone, without requiring multi-site GDPS. Like with GDPS, there is no equivalent functionality on any other platform besides System z and Power servers.

As mentioned above EMC and Hitachi purchased the GDPS and HyperSwap specifications. In July 2012 IBM and Hitachi, Ltd. have successfully completed compatibility and interoperability testing of Hitachi Virtual Storage Platform (VSP) series products at code level 70-03-34. The VSP passed majority of the compatibility tests for GDPS and HyperSwap. Some functions such as Open LUN management and Global Copy (aka PPRC/XD) mode copy processing were not supported at the time of the test. Currently there are 44 Hitachi installations supporting GDPS. EMC haven't participated in the qualification program with IBM. They have GDPS/PPRC & GDPS/XRC in a three-site configuration (aka GDPS/MzGM) in their development lab. To best of our knowledge there are 6 GDPS installations using EMC storage.

Case study of robust SAP on zEnterprise infrastructure

The Nationwide Building Society[5] is a large financial institution in the U. K. Its previous infrastructure was based on several platforms (some old and some new), which proved difficult to develop and manage. The Society wanted to modernize that architecture, improve operational efficiency, reduce costs, improve resilience, ensure future scalability, and guarantee customers' (internal and external) service satisfaction. After evaluating several options, the bank decided to select SAP on System z with GDPS/MzGM, and the application server on IBM Power servers with AIX. The major factors influencing this decision were the reliability, availability and scalability attributes of these platforms. An example of hardware configuration is shown in Fig. 2.

[5] Nationwide is the world's largest building society as well as the second largest savings provider and a top-three provider of mortgages in the UK. It is also a major provider of current accounts, credit cards, ISAs and personal loans. Nationwide has around 15 million members.

Customers can manage their finances in branch, on the telephone, Internet, and post. The Society has around 15,000 employees. Nationwide's head office is in Swindon with administration centers based in Northampton, Bournemouth and Dunfermline. The Society also has a number of call centers across the U. K.

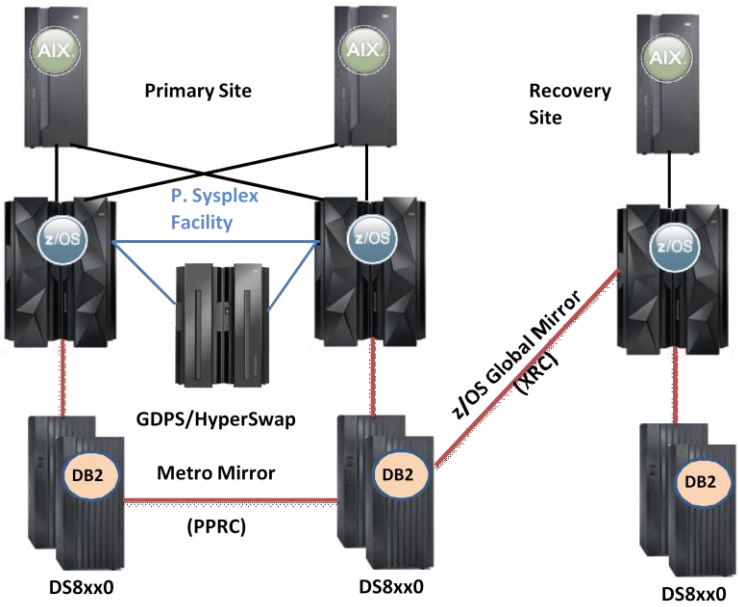

Figure 2: Example of GDPS/MzGM Hardware Infrastructure

DS8870 Competition

EMC VMAX

The first model of VMAX was announced on 14 April 2009. Three years later, EMC announced the VMAX 40K, and the entry model 10K (previously known as VMXe). The original VMAX was renamed to 20K.

The VMAX series is based on standard modules called engines. The V-Max engine is a single level processor module which supports integrated host and disk IO Ports, processors, global memory, up to 360 drives, and support for the Enginuity OS. Each Symmetrix VMAX Engine is interconnected using the Virtual Matrix interface. An engine can be seen in Fig. 3.

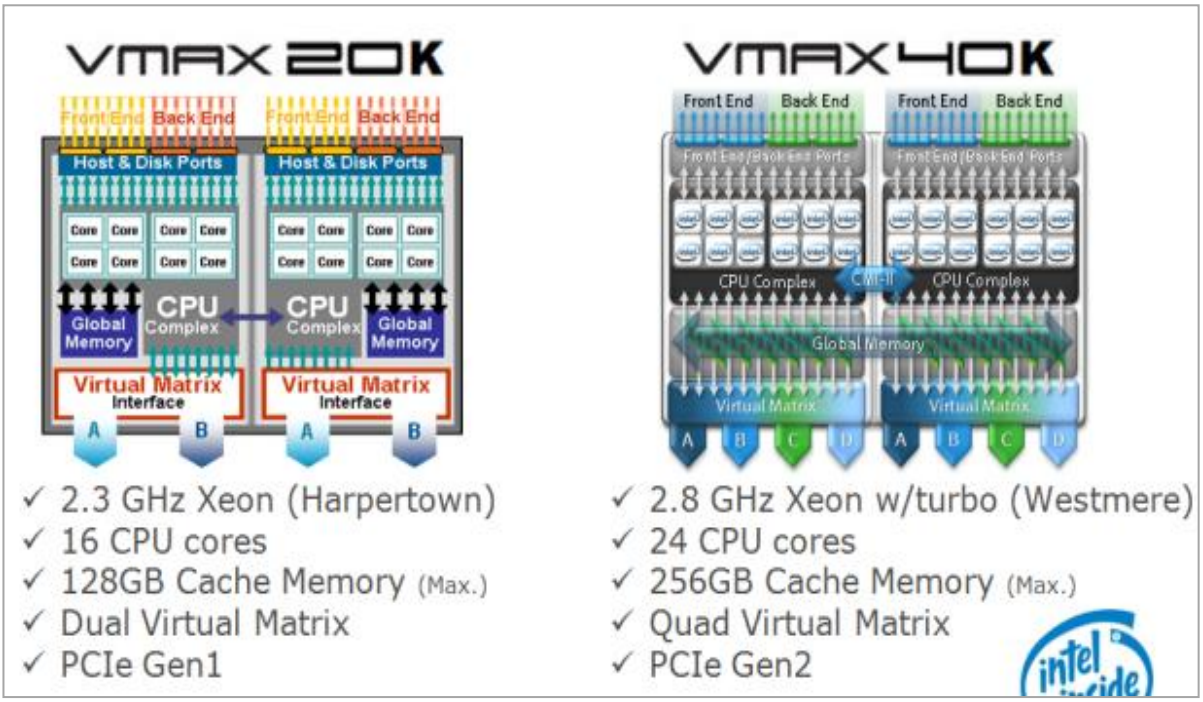

Figure 3 : VMAX 20K, 40K engines

In reality the VMAX is a modular, loosely-coupled clustered subsystem with up to 8 modules, or directors, each with two processors. The VMAX 10K only supports one-to-four engines with up to 128 GB of DRAM per engine (0.5 TB max). The VMAX 20K model does not support 2.5 Inch, small form-factor HDDs.

A “Virtual Matrix” interfaces between the modules. This in reality is RapidIO, a high-performance, packet-switched, interconnect technology that is a non-proprietary high-bandwidth system packet-switched interconnect intended primarily as an intra-system interface for chip-to-chip and board-to-board communications at Gigabyte-per-second speeds.

This cluster-based structure is far less flexible than the DS8870. For example, it offers less flexibility in upgrades; if more channels are required in a full channel populated system then a new module must be installed with processors and cache, which may be not needed. Cache upgrades are disruptive. Adding modules may increase the contention on the RapidIO interface and cause erratic performance. In contrast to the DS8870 where the channels can be defined FC or FICON, VMAX channels groups are pre-defined.

Cache structure, bandwidth and connectivity play a crucial role in determining maximum throughput, performance, and scalability of high-end cache-centric storage systems. EMC’s VMAX cache is distributed among the modules. EMC has not provided details about cache structure and protection, but it can be assumed that the metadata resides in-cache as it does on the former DMX. The cache is mirrored between the modules; hence, each module probably contains the cache for the physical disks attached to its device adapters, plus the mirrored cache of another module In entry-level VMAX 10K models, with only one module, the cache is mirrored between the two directors similarly to typical mid-range subsystems. The cache coherency between the two images is maintained via the RapidIO interconnection. I/O operations initiated on a host port to a volume, which is cached in another module, may take longer than I/O completed within a module. Such situations may happen often with System z, which allows up to 8 paths to an address, chosen by the Dynamic Channel Subsystem (DCS).

“The effective cache is less than half of the purchased cache because of the mirroring and capacity reserved for the directory and the configuration requirements.”

EMC’s VMAX uses fully-mirrored cache: All data, metadata, control tables, and local and remote replications (BCV, SRDF) are located in cache. All access to data blocks share cache bandwidth with all of these other elements. The effective cache is less than half of the purchased cache because of the mirroring and capacity reserved for the directory and the configuration requirements. Using EMC’s Virtual provisioning store in cache further reduces the available cache for applications.

Two other design aspects decrease the effective cache size: the cache page size and the utilization of SRDF/A (EMC asynchronous remote mirroring). A VMAX uses cache pages of 64 Kbyte – a potential waste of cache capacity in interactive processing, which usually only requires small blocks of 4 or 8 Kbyte. As production writes occur, SRDF/A keeps

modifications in-cache and periodically transfers the information in consistent data groups called *delta sets* to the secondary site. Keeping the updates in-cache requires substantial cache capacity. During periods of heavy write-workloads, the delta sets may reduce the practically-available cache capacity for applications. Having less available cache capacity may degrade application performance on the primary VMAX.

With the upgrade from the original Symmetrix to the Symmetrix DMX and VMAX, EMC did not change the cache structure, which remained static cache mapping, as opposed to the dynamic mapping of Hitachi and IBM. The static-cache design of the DMX, which is similar to the original Symmetrix, requires loading .BIN files[6] for LUN assignments – an uncomfortable process that takes time and can corrupt data if not done properly. This last point is particularly true when the configuration includes Business Copy Volumes (BCVs) or remotely-mirrored Symmetrix Remote Data Facility (SRDF) volumes.
The *Symmetrix VMAX Product Guides* do not include engines in the list of components that can be replaced non-disruptively. If more than 1 HDD (in RAID 5) or two (in RAID 6) of a RAID group are connected to the same engine, in the case of an engine replacement some data will be not accessible. A similar situation will happen in the case of a cache repair or upgrade.

Hitachi's Virtual Storage Platform (VSP)

On September 27th, 2010, Hitachi Data Systems, a full subsidiary of Hitachi Ltd., Japan, announced its new high-end Virtual Storage Platform (VSP) as the follow-on model for its USP-V and VM subsystems.

The Virtual Storage Platform (VSP), unlike VMAX, is a purpose-built storage array. As opposed to previous Hitachi high-end products, there is only one model, which scales out by combining resources in up to two control chassis. Each control chassis is housed in a standard 19" rack, and each controls up to two disk expansion racks. The scalability is achieved by adding boards (in pairs) to the chassis and HDDs or SSDs in storage units or cabinets. See single- control chassis block diagram in Figure 4.

[6] BIN file is used to hold the configuration information for the Symmetrix, DMAX and VMAX. It requires EMC services for the initial installation and for hardware upgrades.

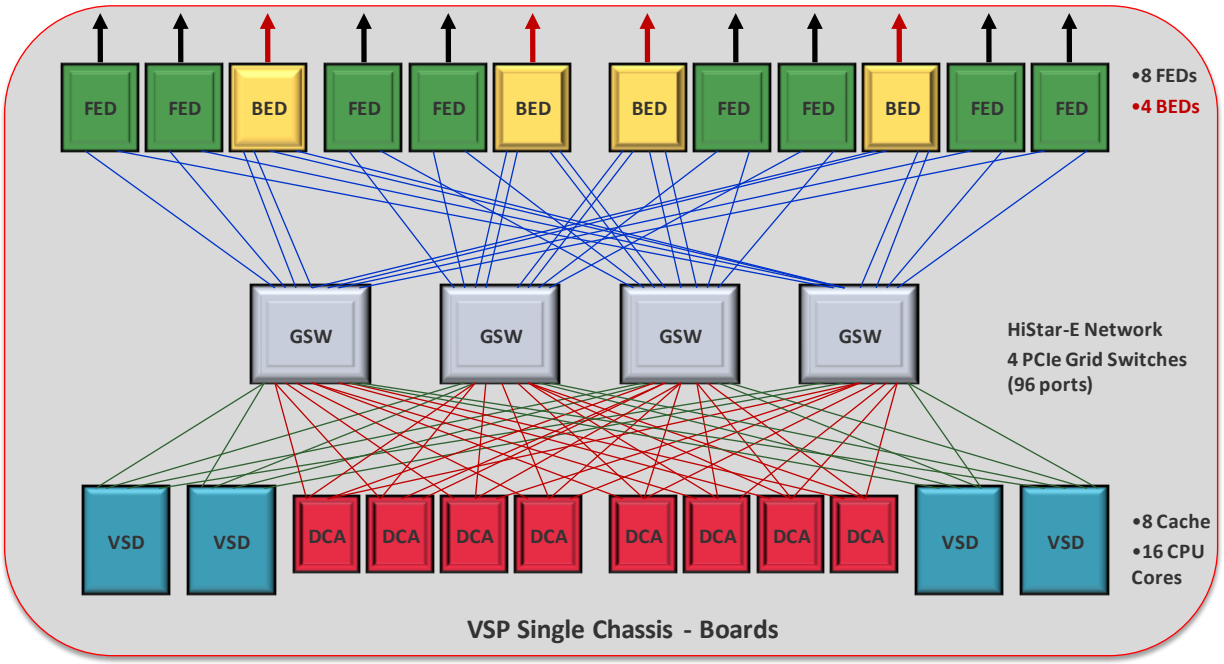

Figure 4 : VSP Single Chassis block diagram/Boards, source: Hitachi Data Systems

FED: **Front-end Director – (host interface)**
BED: **Back-end Director – (disk interface)**
VSD: **Virtual Storage Directors – (central processor boards)**
DCA: **Cache Memory Adapter**
GSW: **Grid Switch**

Similar to the DS8870, the VSP is a tree-level processor structure. The control chassis hosts the Control Unit hardware containing all the control boards in a single rack and one or two drive chassis. The expansion racks can hold up to three drive chassis each. Two control units are cross-connected at the grid switch level across racks and operate as a single integrated array as opposed to a cluster. The Virtual Storage Director (VSD) is a data movement engine, similar to the POWER7 cluster in DS8870. Two or four VSDs are installed per control chassis. Each board contains a quad-core Xeon CPU with 4 GB of local RAM. These boards manage the VSP's operation outside of the data path. The difference between the two and four VSDs configuration is in performance, and it is unrelated to the number of installed FEDs or BEDs. Additional processors and ASICs are located in FEDs and BEDs.

A unique feature of the VSP is the Hitachi Universal Virtualization Layer (introduced with the first version of the USP in 2004). The Hitachi *Universal Volume Manager* software configures, manages, and accesses external volumes in a similar way to how VSP or USP-V internal volumes are handled.

The VSP structure provides full redundancy and ensures non-disruptive repairs and upgrades.

DS8870 Compared to Competition

IBM's System Storage DS8870 series is a stable multiplatform high-end storage system able to support current and future user requirements. It uses state-of-the-art technology to achieve top performance levels and supports superior local and remote data mirroring techniques to ensure business continuity. The DS8870 offers unique features to create synergy with System z such as Performance –DB2 –List Prefetch Optimizer with zHPF.

Technology

As opposed to EMC, Hitachi and IBM are huge technology companies developing the required semiconductor technologies for their products. The corporations exploit synergies between the different divisions to produce state-of-the-art products with "tailored-in" technologies specially designed for these products. It takes several years to design a new storage system, which is why only a company with close relations with disk and semiconductor technology manufacturing can plan for using the latest technologies.

Performance

Transactional performance is measured as response time in milliseconds and as maximum throughput in number of I/O operations per second (IOPS) before the machine enters saturation, causing response time to grow exponentially. Performance of sequential operations is measured in maximum data transfer rates or MB/sec or GB/s.

Storage Performance Council (SPC) benchmarks[7] are modeled on real-world applications, and therefore, help provide customers with meaningful performance results. SPC-1 [8] simulates transactional operation, and SPC-2 [9] simulates sequential access. According to the SPC, from October, 2012, IBM's System Storage DS8870 achieved 451,082.27 IOPS for SPC-1 and 15,423.66 MB/sec for SPC-2. On November 2011 Hitachi Data Systems announced that the USP V achieved 269,506.69 IOPS in the SPC-1 benchmark. Almost a year later, on July 27th, 2012, Hitachi disclosed that the USP V achieved an aggregated average of 13,147.87 MB/sec in SPC-2.

EMC has not published absolute performance figures for its DMX or VMAX subsystems and does not participate in SPC benchmarking.

---

[7] The SPC is a non-profit corporation founded to define, standardize, and promote storage system benchmarks and to disseminate objective, verifiable performance data to the computer industry and its customers. SPC membership is open to all companies, academic institutions and individuals. The SPC created the first industry-standard performance benchmark in 2001, targeted at the needs and concerns of the storage industry and its goal is to serve as a catalyst for performance improvement in storage.

[8] SPC-1 is designed to demonstrate the performance of a storage component product while performing the typical functions of business-critical applications. Those applications are characterized by predominately random I/O operations and require both queries as well as update operations. Examples of those types of applications include OLTP, database operations, and mail server implementations.

[9] The SPC-2 benchmark consists of three distinct workloads designed to demonstrate the performance of a storage system during the execution of business critical applications that require the large-scale, sequential movement of data.

Remote Mirroring Techniques

All three vendors have different designs for synchronous and asynchronous remote mirroring. However, IBM techniques seem to have a much more intelligent design in comparison to the VMAX. Some major differences:

Mirroring in asynchronous mode over long distance communication lines is vulnerable to network disruptions. EMC SRDF/A keeps modifications in the cache and periodically transfers the information in consistent data groups called “delta sets” to the secondary site. Keeping the updates in cache requires substantial cache capacity, decreasing cache available to support I/O operations. The interval between successive delta set transfers, which depends on the amount of altered data and available bandwidth, can be set to between five seconds (for Multi-Session Consistency the lower limit is 15 seconds) and the default value of 15 seconds, which means the secondary copy can be a maximum of 30 seconds (two delta sets) behind the primary copy; i.e. the Recovery Point Objective (RPO) can be up to 30 seconds.
IBM’s Global Mirror function, in contrast, does not collect the modifications in cache but uses a “track bit map” technique which consumes almost no cache capacity for temporary storing “delta sets”. Global Mirror almost continuously transmits groups of consistent data to the secondary site as soon as possible after it is written in the primary. The actual “lag time” depends on the amount of data modified and the link bandwidth available. In practice, assuming that sufficient network bandwidth is available, the average lag time is on the order of five seconds – significantly lower than the SRDF/A adjusted time of five to 30 seconds, with 30 seconds being the default.

In a cache overflow situation, the SRDF/A may stop or change to SRDF/S (synchronous mode), which prolongs response time and is impractical for longer distances. An optional Delta Set Extension function allows SRDF/A to offload some data from cache to special defined disk buffers, but implementation raises costs and complexity.
Effective usage of the communications link bandwidth has significant impact on performance as well as the overall cost of a remote copy implementation. Global Mirror is designed to transfer data to the secondary as quickly as possible, which means immediately after receiving it if the link bandwidth is available. SRDF/A transfers the data periodically until the local “delta set” collection time interval expires, even if link bandwidth is available and unused. Because of the wasted link bandwidth between the periodic bursts and the requirements to transfer the data as fast as possible, to avoid cache issues a faster, more expensive link may be required.
. The results are:

- SRDF/A “pollutes” the cache and reduces the effective cache size which may degrade performance main site on heavy write operation
- Potentially larger RPO – larger data loss (up to 60 sec for the default setting)
- Waste of available bandwidth

Market

Only four vendors compete in the IBM mainframe storage market; EMC, HDS, HP (OEMing Hitachi as XP P950) and IBM. From the Nineties of the last century until 2005, the industry believed that IBM had lost this market. The RAMAC was never planned to be a strategic product. Therefore, IBM could not sell well against EMC's Symmetrix and the Hitachi Data Systems enterprise subsystems. This predicament forced IBM to seek an OEM agreement with StorageTek, which was signed in June of 1996. Under this contract, IBM OEMed StorageTek's Iceberg subsystem, which was renamed RAMAC Virtual Array (or RVA).

The sales of the RVA were better than expected, but despite that the agreement was supposed to last until the end of 2000, IBM, after seeing the ESS working in its lab, practically put the brakes on RVA sales in early 1999, while launching the ESS or "Shark" based on the RS/6000 (now Power Systems) technology. There were several models of the ESS; however most of them suffered initially from poor functionality and required many disruptive engineering changes. The last ESS models, the ESS 800 and the ESS 800 Turbo, increased resiliency and stability of the system. The ESS remained IBM's high-end storage systems until the launch of the DS8000 series in October 2004. The first few months of the DS8000 were also plagued with infant deceases. However IBM put enormous efforts into fixing the problems, which are now ancient history.

Since October 2007, IBM has accelerated its development rate, offering enhancements at the fastest pace in the industry. Examples of major enhancements include IBM's introduction of RAID-6 in August 2008, High-performance FICON for System z in October 2008, full-disk encryption and a solid-state drive (SSD) option in February 2009, and thin provisioning July '2009.

In April 2010 IBM announced and delivered the IBM System Storage Easy Tier, which automates data placement within the DS8000 series system for performance and cost optimization. This includes the ability of the system to automatically and non-disruptively relocate data (at the extent level) across drive tiers (or within a tier, to less-loaded ranks), and the ability to manually relocate full volumes. This was the first sub-LUN automated data placement which was so required for optimal SSD and high-performance HDD utilization. In comparison, EMC announced the FAST 2 in April 2009 but delivered it as FAST VT in December 2010. HDS announced and delivered comparable function called Dynamic Tiering in September 2010, with the launch of the VSP.

According to IDC figures, in the years 2011 and 2012, IBM holds 50.1% of the z/OS disk market share. The competition, HDS and EMC, got respectively 25.8% and 20.8% of the market share.

Summary and Conclusions

IBM's System Storage DS8870 series is a stable multiplatform high-end storage system able to support current and future user requirements. It uses state-of-the-art technology to achieve top performance levels and supports superior local and remote data mirroring techniques to ensure business continuity. The DS8870 offers unique features to create synergy with System z such as Performance –DB2 –List Prefetch Optimizer with zHPF. The DS8870 has all the required characteristics to be positioned on the top of enterprise storage systems for all platforms and in particular for System z.

"IBM's System Storage DS8870 series is a stable multiplatform high-end storage system able to support current and future user requirements.

IBM, which 15 years ago lost some of its high-end disk enterprise storage market share, has managed an impressive come-back, delivering technology that should be put on a short list of any high-end storage procurement. In addition to storage products, IBM delivers four server architectures, and thus the benefits of exploiting of the synergy between IBM servers and the DS8870 storage systems should be evaluated as well. System z mainframes and DS8870 systems have many synergies, in particular in Disaster Recovery and Business Continuity but also with DB2, etc.

The operation GUI, ported from XIV, delivers the most user-friendly functionality in the industry and is also shared by the complete portfolio of IBM storage solutions.

Another IBM advantage is IBM Global Technology Services (IGS). IGS is the world-wide largest IT services organization with more than 50 years of experience in almost every vertical industry, with the ability to design anything from a basic data center infrastructure to the most complicated disaster recovery deployments.
Full redundancy, non-disruptive upgrades and maintenance, hot-swappable components, pre-emptive soft error detection and online microcode changes ensure high availability and data integrity. The advanced remote data replication techniques enable any scheme of disaster recovery deployments.

In summary, the DS8870 deserves to be put on a short list when considering a robust, high-performance, advanced storage system for System z or other platforms for the most demanding storage infrastructures.