Josh Krischer & Associates GmbH

Enterprise Servers, Storage and Business Continuity

White Paper

Storage is Still Not a Commodity: an Updated Comparison of High End Storage Subsystems

Table of Contents

August 2009

Josh Krischer

The information contained herein has been obtained from sources believed to be reliable. Josh Krischer & Associates GmbH disclaims all warranties as to the accuracy, completeness or adequacy of such information. Josh Krischer & Associates GmbH shall have no liability for errors, omissions or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve its intended results. The opinions expressed herein are subject to change without notice. All product names used and mentioned herein are the trademarks of their respective owners.

Storage is Still Not a Commodity: an Updated Comparison of High End Storage Subsystems

Josh Krischer

Josh Krischer is an expert IT advisor with 39 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 he covered high-end storage-subsystems 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.

Two things annoy me when I hear people talk about storage: one is referring to storage subsystems as a commodity, and the other quoting or asking about price per gigabyte. The design of high-end storage subsystems and their functionalities is much more complicated than designing a server; therefore, lowering the discussion of these storage subsystems to a commodity level demonstrates a severe lack of knowledge. Today, only three companies continue to develop such systems, EMC, Hitachi, and IBM:

- EMC offers the Symmetrix DMX family (Direct Matrix) and the newly-announced Virtual Matrix Architecture (V-Max)
- Hitachi Data Systems offers the Hitachi Universal Storage Platform V (USP V). This product is sold under a distribution agreement with Sun Microsystems (Sun StorageTek 9990V) and via an OEM agreement with HP (HP StorageWorks XP24000**)**
- IBM offers the TotalStorage DS8000.

The fifth high-end subsystem, the SVA (Storage Virtual Array) from Sun/StorageTek, has negligible market share and will not be further developed.

Each one of the vendors listed above supports different models of their respective products. However, in order to simplify the comparisons, this research report will focus on the “top-of-the-line” models only.

Basic Designs

The four different products sport four different designs:

EMC’s DMX design is based on a “matrix” of connections between the mirrored cache, which is the heart of the design, the Channel Directors (CDs) as the front-end and the Disk Directors (DDs) on the back-end.

EMC’s Virtual Matrix Architecture, announced on April 14, 2009, is a loosely-coupled cluster of industry-standard modules.

Hitachi Data Systems USP V (announced on May 14th, 2007 and technically equivalent to HP’s XP24000 and the Sun StorageTek 9990V) is based on a massively parallel crossbar switch architecture (called the Hitachi Universal Star Network V), separate data cache (with mirrored write blocks only), and a separate metadata and control memory (with some of the content being mirrored), channel host Front-end Directors (FeDs) and Back-end Directors (BeDs). The central point of this design is the Application Specific Integrated Circuit (ASIC) of the non-blocking crossbar switch architecture technology with nanosecond port-to-port latency, which has embedded logic for checking, routing, and managing data, and has been designed specifically for the USP V series. It was developed collaboratively by engineers from Hitachi's supercomputer, semiconductor, networking, and data storage research divisions. Unlike other vendors, Hitachi has access to researchers and intellectual property from multiple IT disciplines and is not limited to storage only. Hitachi relies on proven cross-pollination research & development techniques that enable it to repurpose IP from one division to another and innovate from within as opposed to relying on off-the-shelf components and third party manufacturers. For example, Hitachi designed the Universal Virtualization Layer for the USP back in 1998—6 years before the product was actually introduced to market.

IBM's DS 8300 structure is based on a 4-way clustered p5+ server (p570) with persistent memory sometimes (for historical reasons) called Non- Volatile Storage (NVS).

Short History

EMC Symmetrix DMX (see Figure 1) was announced in February 2003 as the follow-on to the seven generations of shared-bus structure Symmetrix models starting from 1990. These models, starting with the Symmetrix 4800 and ending with the Symmetrix 8830, were the first high-end storage subsystems to support SCSI, Fiber Channel, Remote Copy (SRDF), and Point-in-Time copies (TimeFinder), but, on the other hand, lacked basic functions such as second copy of write data in cache or RAID-5 support. The major hardware enhancements from model to model were faster processors, larger cache and faster buses, all of which contributed to increased bandwidth, which improved performance and enabled larger scalability. In 1999 EMC's long-term road map was to continue with the shared-bus structure, increasing the number and speed of the buses. However, EMC was late to discover the limits of the shared-bus structure, which resulted in the bandwidth of the last model Symmetrix 8830 being significantly lower than the competitive Hitachi Lightning 9980 V or IBM ESS Turbo subsystems.

In June 2002 EMC acquired storage startup Cereva Networks, Inc. Founded in 1998, Cereva received funding of US$160 million to design and build the Cereva 5000, a fault-tolerant array based on a multi-protocol switch. EMC purchased Cereva's intellectual property for less than $10 million and also hired around 20 former Cereva engineers. EMC leveraged Cereva's intellectual assets in bringing the Symmetrix DMX to market in early 2003. The DMX architecture is in fact an extension of the original Symmetrix, has very similar front and back-end structures, but the "matrix" connections replaced the previous shared buses. The following DMX model was the DMX-2 which was announced in February 2004, and the DMX-3 was announced in July 2005 and shipped in August of the same year. On 16th July 2007 EMC announced the DMX-4 series, mainly enhancing the performance but not adding any new functionality. In addition to more effective microprogram and security enhancements (a contribution of the RSA security division), EMC implemented 4 Gb/s FC connectivity both for host front-end and for a new point-to-point switched back-end. These moderate enhancements were mainly targeted at reducing the gap with the competition, so changing the model nomenclature from DMX-3 to DMX-4 seems slightly exaggerated. On January 14th, 2008, EMC announced the availability of Solid State Disks (SSDs), Virtual provisioning (EMC's version of thin provisioning) and 1 TB SATA-type disks. In comparison to competitive products by Hitachi and IBM, DMX has changed very little in the last four years.

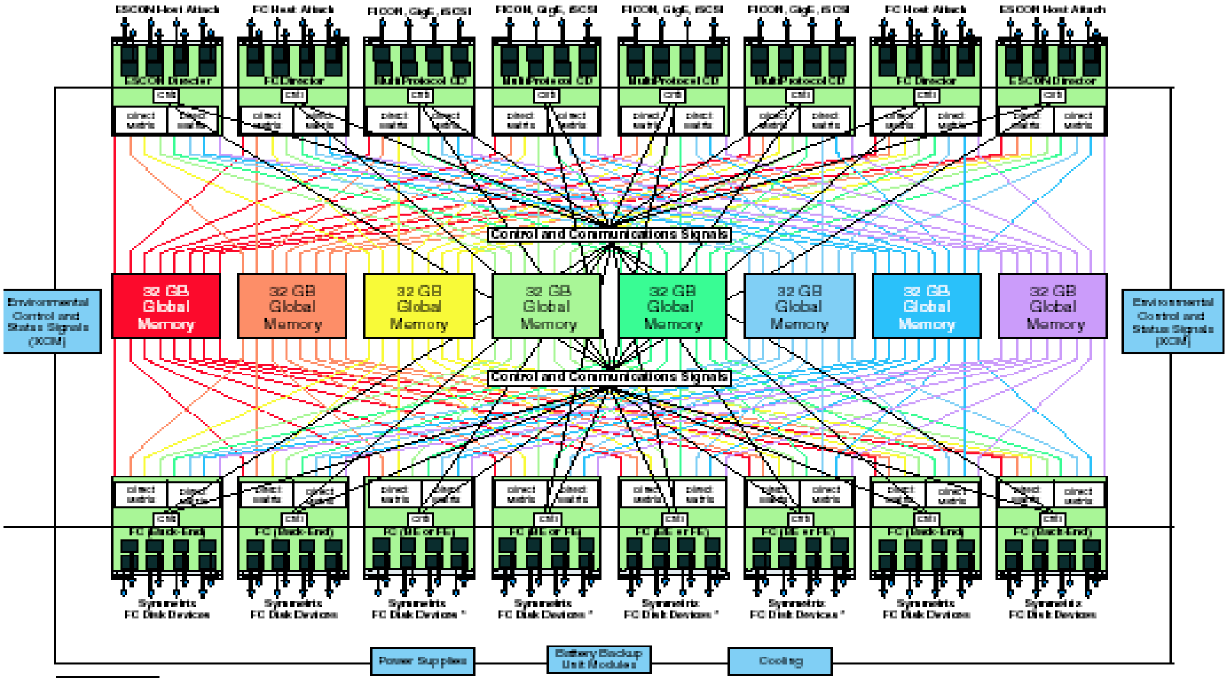

Figure 1: DMX block diagram; source: EMC

On Monday, 14 April 2009, EMC Corporation announced both its new storage architecture - **EMC Virtual Matrix Architecture**™ - and the first storage system based on its design - EMC **Symmetrix V-Max**. To date, EMC has been hesitant to publish detailed technical specifications for the V-Max, so any information contained within this document is based on a best effort assessment of the technology using information that has been made available via the public domain.

The Symmetrix V-Max technology is based on a new system building block, the V-Max Engine. The V-Max engine supports integrated host and disk IO Ports, processors, global memory, up to 360 drives and support for the Enginuity OS. Each Symmetrix V-Max Engine is interconnected using "Virtual Matrix" interface, allowing resource sharing across engines in support of a "scale up and scale out" design. The Symmetrix V-Max supports the latest Fibre Channel, Enterprise Flash and SATA II drives as well as Fibre Channel, FICON, Gigabit Ethernet, and iSCSI host connectivity. ESCON Channels are not supported. The Symmetrix V-Max is available in two configuration models: Symmetrix V-Max SE (supports one V-Max engine) scales from 48-360 drives; and Symmetrix V-Max (supports 1-8 engines) scales from

48-2,400 drives and over 2 petabytes of usable capacity.
In reality the V-Max is a modular, loosely-coupled clustered subsystem with up to 8 modules, or directors and contains industry-standard components. Each V-Max Director or controller contains: two quad-core 2.33 GHz Intel Xeon processors, up to 128 GB of memory, 16 host and 16 drive channel connections, and dual redundant power supplies and cooling fans (see Figure 2). The V-Max Engine physical packaging bears strong resemblance to the CLARiiON CX-4 UltraFlex Dual-controller, though EMC states the technologies are radically different. EMC best practices may recommend each V-Max Engine to have dedicated spare drives.
A “Virtual Matrix” interfaces between the modules, which really 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. EMC states the Virtual Matrix on the Symmetrix V-Max can provide bandwidth capabilities of 192GB/s. However, these values are based on the published theoretical maximums and are not an accurate conveyance of actual sustainable data transfer capabilities. To date, EMC has provided no additional performance data related to inter-engine communication latency and/or specifics related to how the Enginuity 5874 Operating System will exploit the new RapidIO Virtual Matrix.

EMC published a lot of marketing material on the V-Max, but very few technical details to support the marketing claims. For example the announcement started with “*EMC Corporation (NYSE: EMC), the world leader in information infrastructure solutions, today unveiled a breakthrough new approach to high-end data storage with an innovative new architecture purpose-built to support virtual data centers,”* but in reality, there is nothing new. Clustered subsystems have been available through 3PAR InServ, LeftHand Networks (now HP) SAN/IQ, NEC HYDRAstore, Sun Fire X4500, and IBM XIV a long time before V-Max. Several storage subsystems are based on industry-standard servers, for example XIV, LeftHand, Compellent etc. EMC also announced Fully Automated Storage Tiering ("FAST")[1] support on the Symmetrix V-Max which automates the movement of data across multiple storage tiers based upon usage. This is similar to Compellent’s Dynamic Block Architecture, which also optimizes data movement and access at the block level with probably better granularity than can be expected from EMC’s FAST and Compellent’s offering is already generally available. EMC further “beefed-up” the announcement with announced support for EMC SRDF/EDP (Extended Data Protection) replication – *“Zero Data Loss Asynchronous Replication”* which in reality is a cascaded replication (synchronous remote copy to “near site” and asynchronous to “out-of-region” site). This solution is comparable to the Hitachi 3 Data Center Universal Replicator cascade pass-through solution which has been available since August 2008.

EMC emphasizes in the announcement that the V-Max is a “*new architecture purpose-built to*

[1] According to EMC, FAST will be available on Symmetrix V-Max systems later this year.

support virtual data centers". In the official press release, the words *virtual* and *virtualization* are repeated 35 times even though the V-Max offers far less virtualization than the Hitachi USP (virtual ports, logical partitioning, and virtual back-end) or IBM's DS8000 with logical partitioning. The major feature supporting EMC's claim is the "EMC Storage Viewer for Infrastructure Client," a tool which provides information about read-only storage mapping, paths employed, and characteristics of the connected storage. This tool may be useful; however, it is only available for VMware's virtual architecture and EMC storage.

The Symmetrix V-Max system is generally available as of this writing; EMC continues to sell the DMX-4, positioning the V-Max mainly for virtual server infrastructures.

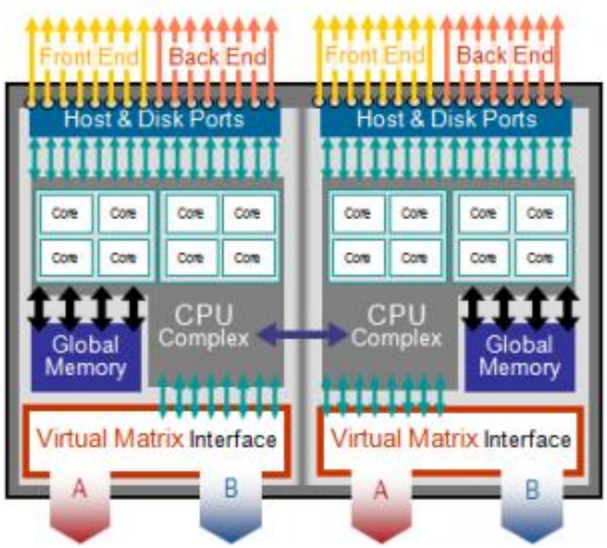

Figure 2: V-Max block diagram; source: Gestalt IT

Hitachi Data Systems Universal Storage Platform V (USP V) (see Figure 3) was announced on May 14th, 2007 and general availability began in June of that year. Over the last twenty years, Hitachi, having the fastest "turn-around" times in the industry, brought to market new control unit designs approximately every four to five years, with a "mid-life" kicker two years after launching the original product. For example; there was the enterprise 7980-3 storage system in 1990, followed in 1995 by the bus-structured 7700, which was the first storage subsystem with separate control memory and separate data cache, as well as the first with full redundancy, without a single point-of-failure, permitting non-disruptive micro-code modifications and allowing "hot" component swapping. Hitachi engineers quickly recognized the limitation of the bus structure, therefore, the follow-on subsystem (Lightning 9900 series) was announced in June 2000. The Lightning 9960 was based on an internal crossbar switch (Hi-Star architecture) with 6,400 MB/sec bandwidth — more than four times the bandwidth of EMC's Symmetrix 8000 at that time and the first system to feature Fibre Channel back end and Fibre Channel disk drives.

"Over the last twenty years, Hitachi, having the fastest "turn-around" times in the industry, brought to market new control unit designs approximately every four to five years, with a "mid-life" kicker two years after launching the original product."

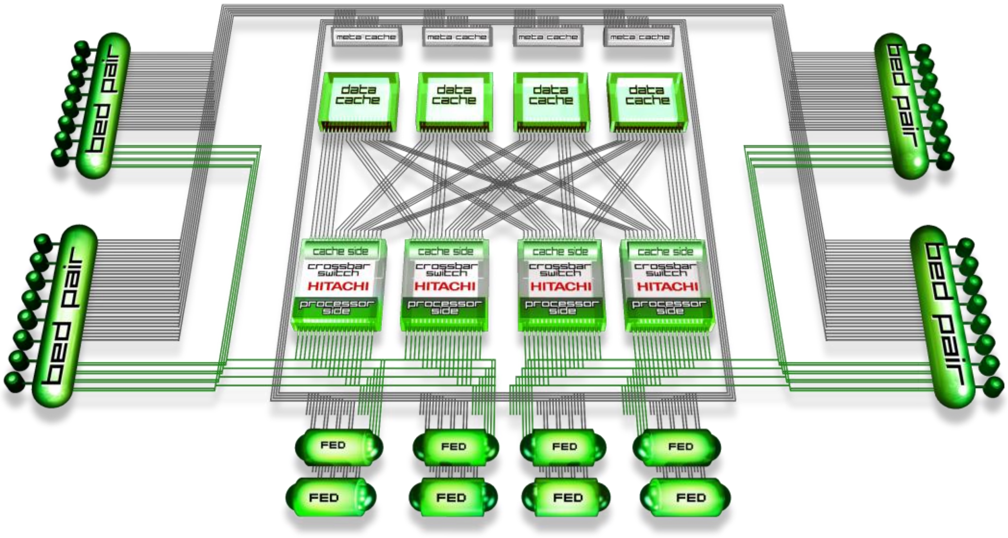

Figure 3: USP-V block diagram; source: Hitachi Data Systems

Two years later Hitachi announced the Hitachi Freedom Storage Lightning 9980V, which was more than a mid-life kicker. In addition to a huge increase of the bandwidth (up to 15,900 MB/sec -10,600 for data and 5,300 for internal control) - an order of magnitude higher compared with the other high-end subsystems available at that time - Hitachi introduced the concept of "Virtual Ports." This embedded virtualization layer provided up to 128 virtual for each of the 32 physical ports. Hitachi partners HP and Sun Microsystems announced the 9980V the same day (as the HP SureStore XP1024 and Sun StorEdge 9980, respectively).

On September 7th 2004, Hitachi Data Systems and its partners introduced the Hitachi TagmaStore Universal Storage Platform, enabling Hitachi to effectively change the high-end storage landscape. In addition to extremely high performance, scalability and resiliency, it featured an integrated virtualization layer able to control third party subsystems and supporting partitioning. Hitachi OEM partner Hewlett-Packard and re-seller Sun Microsystems announced the products as the HP StorageWorks XP12000, and the Sun StorEdge 9990, respectively. In May, 2006 Hitachi Data Systems announced "mid-life" kicker enhancements to the USP, with a 25% performance boost achieving maximum IOPs of 2.5 million through its cache, as well as new security features such as audit logging and extended business continuity and disaster recovery capabilities.

In May 2007 Hitachi Data Systems and its partners announced the all-new control unit USP V (HP XP24000, Sun StorageTek 9990V) which introduced an expanded virtualization layer, thin provisioning, large logical storage pools, performance boosted to 3.5 million peak IOPs (later increased to more than 4 million) through cache, and greatly increased scalability (support of up to 247 Petabytes, up from 32). **IBM TotalStorage DS8000** (see Figure 4) was announced in October, 2004 with first shipments in 2Q05. This is the second version of IBM's disk storage subsystem based on the "Seascape" architecture concept, a Storage Enterprise Architecture, which is based on using standard components such as the IBM System p processors. Earlier Seascape product offerings included the IBM 3466 Network Storage Manager and the Magstar MP 3575 Tape Library DataServer.

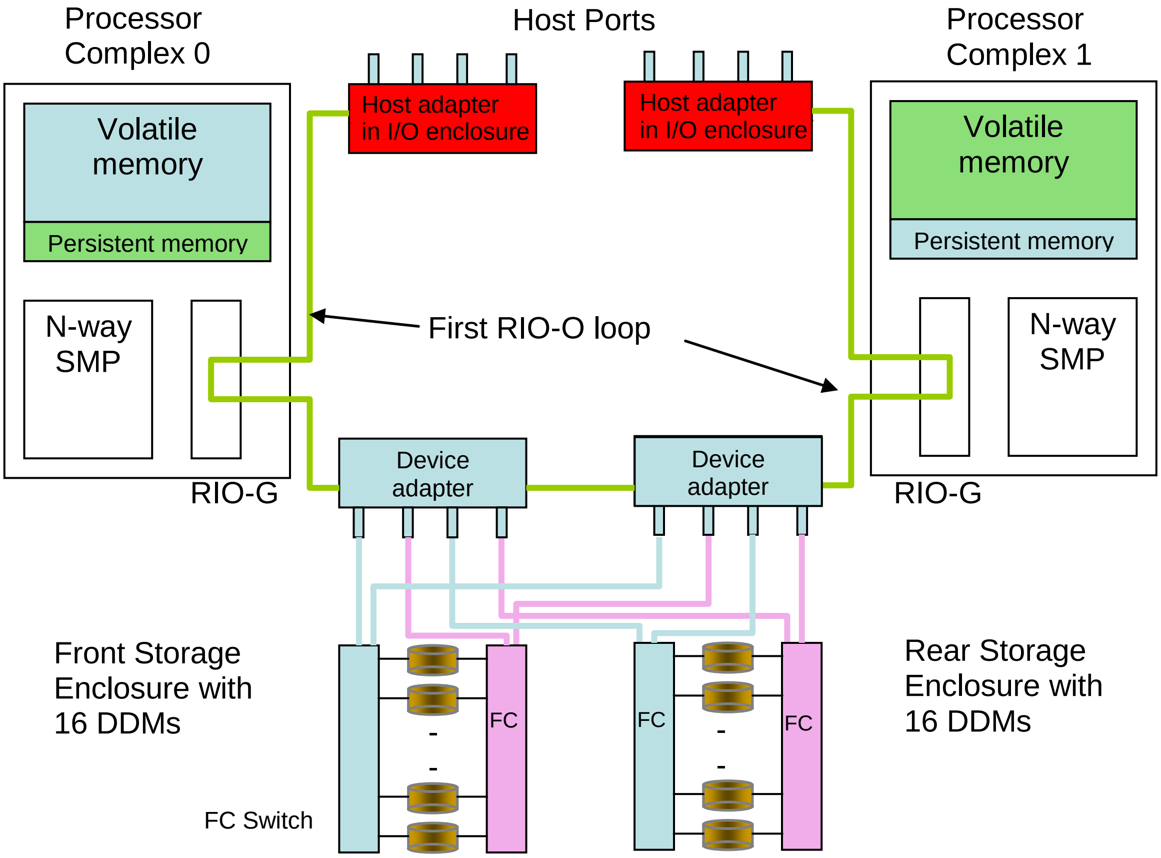

Figure 4: DS8000 block diagram; source: IBM

The DS8000 series is a follow-on of the Enterprise Storage Server (ESS) – code-named "Shark" – which was announced in July 1999 and shipped from September of the same year. According to my estimation, the launch of the Seascape Architecture disk subsystem (initial codename Seastar) was about five years behind the original schedule, which caused huge problems for IBM's storage division. The last of the conventional, monolithic control units was the 3990-6, which had been announced in September 1993. This control unit, which was initially designed for the 3090 traditional disk system, also supported the IBM RAMAC Scalable Array Storage launched in 1994. However, RAMAC was never planned as a strategic storage subsystem and was intended to fill the gap until the arrival of Seascape, and subsequently, 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). IBM 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. StorageTek profited from this deal on a short-term basis, but only at the cost of future earnings. From 1999 StorageTek tried to continue selling the product as the SVA but never regained any material market share. Sun Microsystems acquired StorageTek in 2005 but remained loyal to Hitachi high-end subsystems, thus giving the SVA a *coup de grace* as a storage subsystem but keeping it as a part of the Virtual Storage Management (VSM), a virtual tape subsystem.

The RVA was replaced by IBM's own product, ESS G, followed by the E and F models up until the ESS 800 which was announced in June 2002. Similar to its predecessors, the G, E and F models, the 800 and 800 Turbo were two-node, four-processor clustered SMP (Symmetric Multiprocessing) designs with buses handling data and command movement between subsystems, as well as RAID controller cards offloading RAID functionality from the nodes. The major enhancements to the different ESS models were faster processors (following the System p development) and larger cache sizes. The 800 Turbo had faster clock speeds than the ESS 800 and two additional microprocessors to each SMP for a total of six per node. The ESS remained IBM's high-end storage subsystems until the launch of the DS8000 in October 2005. Less than a year after the announcement, in August 2005, IBM boosted the performance by introducing Turbo models with 2-way (DS8100) and 4-way (DS8300) Power5+ engines. On October 23rd 2007, IBM introduced, among other enhancements, FlashCopy SE ("Space Efficient"); Dynamic Volume Expansion; Storage Pool Striping; and z/OS Global Mirror Multiple Reader. The first two in particular, FlashCopy SE, which is a snapshot-type local replication that consumes significantly less storage capacity compared to full-volume FlashCopy, and Dynamic Volume Expansion, which provides non-disruptive, simplified LUN size management, complemented the existing functionality, making the DS8000 more competitive.

"Since October 2007 IBM has accelerated its development rate, offering enhancements at the fastest pace in the industry."

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 '08, high-performance FICON for System z in October '08, full-disk encryption and a solid-state drive (SSD) option in February '09, and thin provisioning July '09.

How do the four high-end storage subsystems compare?

Cache structure

Cache structure, bandwidth and connectivity play a crucial role in determining maximum throughput, performance, and scalability of high-end cache-centric storage subsystems. These are also the biggest differences among the four major subsystems.

EMC's DMX uses fully-mirrored cache; all data, metadata, control tables 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 incurs a cache overhead: each thin device requires 143 Kbyte plus 8 Kbyte for each GB of reported device size, further reducing the available cache for applications.

Two other design aspects lower the effective cache size: the cache page size and the utilization of SRDF/A. A DMX 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 means there may be a degradation of application performance on the primary Symmetrix.

> *"Cache structure, bandwidth and connectivity play a crucial role in determining maximum throughput, performance, and scalability of high-end cache-centric storage subsystems."*

Moving from the original Symmetrix to the Symmetrix DMX, EMC did not change the cache structure, which remained static cache mapping, as opposed to the dynamic mapping of Hitachi and IBM. Because of the static-cache design of the DMX, similar to the original Symmetrix, it requires loading .BIN files 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 DMX cache is built from 2 to 8 cache modules with capacities between 16 and 512 GB. Each Cache module has 8 connections of 1Gbit/s to each of the Channel Directors (CD, host front-end interface) and similar connectivity to the 8 Device Directors (DD, back-end interface), which means that a fully-configured DMX has 128x1 GB/s bi-directional connections. However, this does not mean that the maximum cache bandwidth is 128 GB/s because the DMX cache supports a maximum of only 32 concurrent operations[2] (4 concurrent memory transfers per cache module), which only yield a total theoretical 32 GB/s for data and control traffic. Each of the 1 GBs serial connections is composed of a pair of full-duplex unidirectional serial links – two 250 MB/sec serial transfer links (TX), and two 250 MB/sec serial receive links (RX). The result is 0.5 GB/s per connection in each direction, which, depending on the type of workload, may further reduce the practically-available bandwidth. A modest DMX configuration with two cache modules, two CDs, and two DDs has half of this bandwidth. Because the cache directory is stored in-cache and each access to the cache requires an additional access to fetch the metadata, the effective bandwidth is even lower. Considering all of the above, the maximum achievable bandwidth of the DMX is below 16 GB/s, much lower than the 128 GB/s stated in its documentation.

With **EMC's V-Max,** 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 with the 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 subsystems 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).

Hitachi's USP V has up to 8 data cache modules with a maximum capacity of 512 GB for data and up to 32 GB of separate, dedicated control memory for the metadata (only 28 GB are addressable). Only the "write portion" of the cache is mirrored with a threshold that is automatically adjusted depending on the activity. Hence the effective cache size is reduced by ca. 20% for a typical workload. The dynamic cache structure and the separate control cache allow for dynamic configuration changes in the data cache by changing bits in the control store through a service processor. The Hitachi cache design is the most advanced in the industry and provides any-to-any connectivity between any host port and disk array over individual full duplex 1GB/s paths. Access to storage can also be load-balanced across multiple host ports since they can all view the same cache image. This provides additional resilience since the failure of any one or two components would not be noticed by the end-user. The Hitachi USP V cache bandwidth depends on the number of the cache modules as well. The maximum available bandwidth is 68 GB/s for data and 38 GB/s for the metadata which aggregates to a total of 106 GB/s for the whole cache. The Massively Parallel Universal Star Network Crossbar Switch Architecture supports up to 320 concurrent internal

[2] "The Symmetrix DMX matrix with 128 direct point connections and 32 memory regions provides a data matrix rated to 64 GB/s of internal aggregate bandwidth supporting up to 32 concurrent global memory operations", EMC Symmetrix DMX architecture guide.

cache and control cache operations which, is ten times the maximum number of cache operations of the DMX. The USP V uses 32 Kbyte cache pages.

IBM's DS8000 SMP cluster structure is difficult to compare with the EMC DMX and the Hitachi USP because instead of using a dedicated cache, the DS8000's cache is allocated as part of the System p server memory. The P570 server has two level caches (L1 and L2) in addition to its main memory, which creates three levels of hierarchy. IBM claims that the tightly clustered SMP, the processor speeds, the L1/L2 cache sizes and speeds, and the memory bandwidth deliver better performance in comparison to dedicated, single level caches. The cache size ranges between 32-256 GB. Each side of the cluster has its own cache and persistent memory – still carrying the name Non Volatile Storage (NVS) – of the other side. During normal operation, the DS8000 preserves fast writes using the NVS copy on the other side. This cross-connection protects write data loss in case of a loss of power or other malfunctions. The DS8000 uses 4 Kbyte cache pages, which prevents polluting the cache with unnecessary data during interactive operations.

Front-end connectivity

Front-end connectivity influences the maximum available throughput of storage subsystems. Table 1 shows the maximum connectivity options of the four subsystems.

Port Type	**EMC DMX-4**	**EMC V-Max**	**HDS USP V**	**IBM DS8300**
FC Ports (4 Gb/s)	**64**	**128**	**224**	**128**
Max. FICON Ports	**48**	**64**	**112**	**128**
Max. ESCON Ports	**64**	**-**	**112**	**64**
iSCSI Ports	**48**	**64**	**-**	**-**

Table 1: Front-end connectivity

Please keep in mind that with mixed-channel configurations, the number of maximum ports is lower than in this table. For example, the maximum number of ports of the IBM DS is 128 and can be a combination of 4 port FC/FICON host adapters or two port ESCON adapters.

In 2002 Hitachi launched the 9980V with a new feature called Virtual Storage Ports[3]. In the latest USP V, this virtualization layer provides up to 1,024 virtual ports for each of the 224 physical ports, including LUN 0 for booting. A mode set is specified at the subsystem to set the appropriate server platform and provides separate storage pools for each host. With separate LUN addressing, QoS (access priorities), and LUN security, each storage domain

[3] This is a similar idea to the NPIV (N-Port ID Virtualization) Fiber channel standard, which allows a single Fiber Channel port to appear as multiple distinct ports, providing separate port identification and security zoning within the fabric for each operating system image as if each image had its own unique physical port.

appears as a separate virtual array despite using the same physical port. This ensures safe multi-tenancy as there is no danger of overwriting another server's data. Multiple hosts can safely share a common physical storage system since each host can be assigned its own virtual private storage. Virtual private storage is analogous to virtual private networks in the IP networking world. This embedded virtualization layer is particularly useful for supporting heterogeneous clusters and server virtualization.

Back-end connectivity

- EMC's DMX-4 supports up to 64 back-end 4Gb/s paths in its full configuration, with 32 active and 32 passive for failover.
- Each of the V-Max modules comes with device adapters that support up to 16 paths of 4 Gb/s each, with 8 active and 8 passive.
- Hitachi's USP V supports up to 64 paths of 4Gb/s each, all active.
- IBM's DS8300 back-end is comprised of 64 paths of 2Gb/s each, with 32 active and 32 passive.

Bandwidth

Array bandwidth is in many cases the most important factor in determining the maximum throughput of a storage subsystem and acceptable performance levels. In cache-centric storage architectures there are several bandwidths to observe:

Max. Bandwidth (GB/s)	**EMC DMX-4**	**EMC V-Max**	**HDS USP V**	**IBM DS8300**
Front-end	256	512	896	256
Back-end	256	512	256	128
Cache	32[4]	24[5]	106	

Out of the three bandwidths, it is the smallest one that determines the maximum available throughput; therefore, despite the fact that the front-end bandwidth of each array is higher than the cache bandwidth, it is the cache bandwidth that dominates. The subsystem cannot send more data to the hosts than it receives from the cache. Also, for designs that use unidirectional cache paths, achieving full usable cache bandwidth means having a 50%-50% read-write workload mix which differs from typical workloads of 80%-20% or 70%-30% .

[4] The effective bandwidth is lower in general.
[5] Virtual Matrix bandwidth according to EMC Symmetrix V-Max SE Storage System maximum specifications

Scalability

All the high-end storage subsystems support industry standard FC, and SATA magnetic disks and SSDs which fit into the same slots. Out of the four products, EMC's V-Max claims to have the largest scalability, supporting up to 2,400 drives (48 to 360 per module). The DMX-4 scales up to 1,920 HDDs. The Hitachi USP V supports up to 1,152 HDDs of internal capacity and 247 Petabytes of internal and externally virtualized storage. The IBM DS8300 supports up to 1,024 drives. The figures above are what I refer to as "PowerPoint" or "brochure" scalabilities, which usually are not achievable under normal utilization. Several factors such as the number of host connections and the different bandwidths influence the practically- installable capacity. For example, the DMX-3/4 with its 64 host ports and limited cache bandwidth can support 1,920 drives only for very-low activity environments; hence, EMC's claim to the "world's largest high-end storage array scalability" is questionable. Lab tests and customers' experience show that the USP V will scale linearly as more disks are added. A USP V with 1,024 disks will deliver almost four times the throughput of one with 256 disks.

Functions and features

EMC reached its position as one of the leading storage companies in the '90s by introducing several features before Hitachi and IBM. The EMC Symmetrix was the first high-end subsystem to support other systems than IBM mainframe, to support point-in-time and remote copies, etc. Today, all four subsystems support these basic functions but differ in more advanced functionalities. The leadership in introducing new storage functions has since passed to Hitachi, which skillfully uses its virtual platform for additional developments. In addition to the Virtual Ports mentioned above, the USP V supports some advanced features such as:

- **Universal Virtualization Layer** (introduced with the first version of the USP in 2004). The virtualization layer is embedded in the processors of the USP V channel adapter cards. These cards function as a normal port for volumes which are resident internally or as a host bus adapter for accessing external storage – which may be from Hitachi Data Systems or third parties. The Hitachi Data Systems Universal Volume Manager software configures, manages, and accesses external volumes in a similar way as if they were USP V internal volumes. Externally-connected storage may use the same functionality as internal storage, which means that data replication software, transparent data movements, dynamic provisioning, intelligent tiering, and other high-end subsystems features can be used in the same way, regardless of whether the data resides on internal or external volumes. The virtualization of heterogeneous storage systems simplifies storage management, enables easier migrations, reduces the complexity of disaster recovery schemes and allows building tiered storage without compromising on functionality. It gives customers the ability to store non-critical data or to archive mainframe data on low-cost SATA systems, for example. On April 22nd, 2009 Hitachi

Data Systems announced the "Switch it On" program, which offers their storage virtualization software free of charge.

- **Hitachi Dynamic Provisioning,** unveiled in May 2007 as part of the USP V announcement, enables two capabilities: high performance wide striping across pools of disks, and "thin provisioning". The thin provisioning enables allocation of virtual storage as needed without having to dedicate physical disk storage up-front. Additional capacity can be allocated without any disruption to mission-critical applications from existing or newly-installed capacity. This feature, in addition to saving investment and running costs (less energy, smaller floor space), also improves performance by striping the data across all the disks in the array. Striping the data among a large number of physical devices practically eliminates "hot spots", which results in almost uniform performance. In November 2007 Hitachi extended Hitachi Dynamic Provisioning support as a service to external 3rd-party attached storage systems. On June 17th, 2009 the company enhanced Dynamic Provisioning with **Zero Page Reclaim**, a feature that returns unused storage blocks back to the storage pool and reclaims storage space and the **Automatic Dynamic Rebalancing** – when physical volumes are added to expand a Dynamic Provisioning pool of storage, existing virtual volumes in the pool are automatically re-striped across these new physical volumes to rebalance the workload. Hitachi is the only high-end class storage vendor that offers automatic rebalancing of the virtual volume pages through active re-striping in order to take advantage of new disks when the pool is expanded.
  EMC followed Hitachi announcing Virtual Provisioning for the DMX-3 and DMX-4. Advantages of Hitachi Dynamic Provisioning include the ability to support remote-pair operations between dynamically-provisioned volumes and static volumes, as well as online volume capacity expansion for Dynamic Provisioning, which is not supported on the DMX. In July 2009 IBM announced thin provisioning for the DS8000 series.
- **Virtual Partition Manager** is subsystem partitioning (introduced in 2004 on the original USP) that allows resources (internal and externally attached) such as capacity, cache, and ports to be dynamically partitioned into "virtual machines" – each with its own virtual serial number (for asset tracking and chargeback purposes). Up to 32 of these virtual machines (logical partitions or LPARs) can be created on a USP V, each separately managed and password-protected, to provide better resource allocation and enhanced protection through isolation between the various partitions. This capability enables users to build different internal service levels, to separate test from production environments, and to reduce the costs for users that previously, for data security reasons, may have required separate storage subsystems.

  IBM's TotalStorage DS8000 Series supports two storage system partitions (50:50, 75:25, 25:75 ratios) as well. Unlike the Hitachi USP V, each of the DS8000 partitions can run separate copies of the DS8000 microcode, which may simplify testing different versions of the microcode.

- **Storage Security Services** include several functions, some of them introduced as early on as the 7700 subsystem in the mid ‘90s. These functions include:
  Controller-based data shredding; Write Once Read Many (WORM) software for tamper-proof data protection (required by most compliance regulations); and Role-Based Access, an Audit Log file which stores a history of all user access operations performed on the system to allow users to trace un-certified access to data and more.
  Similarly, EMC’s Symmetrix Audit Log records major activities such as host-initiated actions, service processor activity and attempts to access data which were blocked by security mechanisms.
- **Encryption -** Hitachi Data Systems offers a hardware-based (storage controller) encryption option for the Universal Storage Platform V that is compatible with both open and mainframe systems. This option encrypts the internal drives using strong encryption (AES-256) without impacting throughput or latency. Hitachi's data-at-rest encryption feature also includes integrated and user-friendly key management functionality. IBM’s DS8000 offers similar encryption in conjunction with Seagate disks. The management is performed by the Tivoli Life Cycle Manager, which is also used for tapes.
  EMC’s PowerPath Encryption uses several RSA (company acquired by EMC in 2006) components. The software, which is host-based, encrypts data on Symmetrix and CLARiiON subsystems. Additional agents must be installed on all participating hosts. The RSA Key Manager for Datacenter centralizes the management of RSA-generated keys for EMC PowerPath Encryption.
- **Hitachi Universal Replicator** provides asynchronous, storage-agnostic data replication for internal and externally-attached storage. Unlike other techniques that employ the cache for these purposes, this advanced technique uses disk to log temporary data before transferring it to the remote site(s), and thus, significantly reduces cache utilization and bandwidth requirements. There are many differences in remote copy techniques between the four high-end subsystems, which could be the subject of another extensive research report.
- On May 27th, 2009 Hitachi Data Systems introduced the **Hitachi High Availability Manager** – a local and remote storage array clustering option which allows non-disruptive failover across USP V storage subsystems. This option enables instant data access recovery at the primary site or at a remote site in case the primary subsystem malfunctions or is unavailable. Unlike the mainframe HyperSwap software, this feature is implemented in hardware and targeted at non-mainframe SAN infrastructures. Hitachi High Availability Manager also provides a simplified non disruptive migration to next generation Hitachi storage systems.

Compatibility and synergy with operating systems and applications

IBM as the architecture owner of System z (z/OS), System p (AIX) and System i (i5/OS) has natural advantages in delivering new features for these platforms and exploiting the synergies between the DS8000 and the operating platforms/applications. To remain compatible, both EMC and Hitachi have purchased some of the specifications for IBM's exclusive features. As part of the agreements, IBM provides the technical specifications of these features, but not the code itself, which has to be developed by EMC and Hitachi. Historically, the other vendors (in particular Hitachi) usually offer their own version of the product within 12-24 months, but sometimes development may take longer.

- **HyperPAV** relieves logical volume size constraints and performance limitations of static PAV (introduced in 1999 along with the IBM 2105 ESS) and WLM-managed aliases. HyperPAV was announced in October of 2006 for operating system versions higher than z/OS 1.6. Alias addresses managed by HyperPAV are assigned by the I/O Supervisor (IOS) according to the request priority of the Workload Manager (WLM). These aliases persist only for the duration of an I/O operation. After completion of the I/O operation, the aliases are returned to a free pool to be used by another high priority I/O request. The dynamic operation of HyperPAV may significantly reduce the number of aliases required to meet the WLM's performance objectives compared to Dynamic PAVs.

- **Multiple Readers for IBM System Storage z/OS Global Mirror.** z/OS Global Mirror a.k.a. XRC was one of the first remote-copy techniques introduced more than a decade ago. With this technique, data modifications are temporarily stored in a side-file of the cache and retrieved periodically by the 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, a suspension of the remote-copy operation may occur. Since the initial introduction of z/OS Global Mirror, many changes have been made in the storage landscape, for example, 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" which allows exploiting parallelism for the SDM when emptying these sub side-files. The users of z/OS Global Mirror with DS 8000 benefit from improved performance and less disruptions under heavy write-load conditions, and as a result, significantly better performance in busy z/OS environments in particular.

- **Geographically Dispersed Parallel Sysplex (GDPS)** is a multi-site application availability solution with fast recovery time and highly-automated control. Many of the world's largest organizations in the financial services and other key industries have deployed GDPS to protect mission-critical applications. It manages application availability in and across sites for both planned maintenance and unplanned situations such as site failures or full-blown disasters. GDPS was initially designed for mainframe z/OS systems but with continuous development was later enhanced to support select open systems platforms, as well. GDPS supports any mainframe-compatible subsystem, either singular or in mixed configurations; however, GDPS deployments require a high level of storage compatibility to benefit from its annual enhancements. On March 3rd, 2009 Hitachi announced that it completed qualification testing for GDPS, and that the USP models are certified for IBM's GDPS disaster recovery solution. Almost all storage subsystems currently participating in GDPS schemes are manufactured by Hitachi or IBM.

- **High Performance FICON (zHPF)** is a new channel protocol designed for more efficient I/O operations, reducing response times particularly when accessing SSDs. The zHPF protocol was announced by IBM in October 2008 and has been supported on Hitachi's USP since January 2009. In July 2009 IBM enhanced it with zHPF Multitrack, which allows applications to read or write more than one track's worth of data in a single transfer, removing potential I/O bottlenecks.
  The following features are exclusive to the DS8000:

- **z/OS DB2 SSD placing tool** analyzes the DS8000's SMF records to produce a report that identifies data sets and volumes which can benefit from residing on SSD drives.
- **AIX/DB2 I/O priority support** increases performance of DB2 with the DS8000.
- **AIX/DB2 Cooperative Caching** enables DB2 on AIX to manage a DS8000's cache more effectively.
- **HACMP/XD Integration with DS8000 Metro Mirror** relieves users from having to write their own scripts.
- **System i POWER HA Integration** integrates replication management with IBM's System i disaster recovery solution.

Performance

Transactional performance is measured as response time in milliseconds and as maximum throughput (which is measured in number of I/O operations per second – IOPS) before the machine enters saturation, causing response time to grow exponentially. Performance of sequential operation is measured in maximum data transfer rates or MB/sec or GB/s. Storage Performance Council (SPC) benchmarks[6] are modeled on real-world applications, and therefore, help provide customers with meaningful performance results.
SPC-1 [7] simulates transactional operation and SPC-2 [8]simulates sequential access.

[6] 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.

[7] 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

According to the SPC, on December 5, 2006, IBM's System Storage DS8300 Turbo achieved 123,033 IOPS for SPC-1 and 3218 MB/sec for SPC-2. On October 1st, 2007 Hitachi Data Systems announced that the USP V achieved 200,245.73 IOPS in the SPC-1 benchmark. Almost a year later, on September 8th, 2008, Hitachi disclosed that the USP V achieved an aggregated average of 8,724.67 MB/sec in SPC-2. Since delivering the benchmark results to the SPC, IBM introduced several performance improvements utilizing new advanced cache algorithms, so current performance figures are likely to be higher than the published ones. EMC has not published absolute performance figures for its DMX or V-Max subsystems and does not participate in SPC benchmarking.

Availability

All four subsystems provide high levels of availability and reliability, non-disruptive repairs, upgrades and microcode changes, but only Hitachi and its partners, HP and Sun, are ready to provide customers with a 100-percent data availability warranty. In addition to the usual RAID techniques, Hitachi's USP models also support RAID-6 in 6D+2 P configurations. While this technique consumes 12.5 percent more storage than RAID-5 in 7D+1P configurations, it ensures almost indefinite Mean Time Between Data Loss (MTBL) and reduces the rebuild time by 60-percent in comparison to RAID-5 groups on the same system. In random writes, RAID-6 may impact performance by increasing write penalties, but no such impact should be registered in large blocks of sequential writes. Hitachi announced RAID-6 support in 2005; EMC followed with RAID 6 for the DMX-4 in 2007, and IBM followed in August 2008.

Technology

EMC's DMX uses standard PowerPC processors and other standard off-the-shelf components. IBM's DS8300 Turbo is built from IBM System p p5 570 clusters using P5+ processors and custom-fabricated ASICs. The Hitachi USP V uses MIPS processors and custom-fabricated ASICs as well. The V-Max employs industry-standard modules with two quad-core 2.33 GHz Intel Xeon processors.

As mentioned earlier, Hitachi Ltd., being a large technology corporation, leverages other branches of technologies in its storage products such as tailor-made ASIC chips or the Universal Star Network V crossbar switch architecture, which was designed by multiple IT groups within the company. IBM, another technology producer, uses ASICs such as the RAID Data protection Data Mover ASIC as well. These ASICs are responsible for managing, monitoring, and rebuilding the RAID arrays, for example.

operations and require both queries as well as update operations. Examples of those types of applications include OLTP, database operations, and mail server implementations.

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

Future developments

EMC may (though not likely) enhance the DMX in the future with faster processors and a larger cache, but the cumbersome static cache architecture design will remain in place. IBM will deploy POWER6 technology, announced on May 21st, 2007, which is twice as fast as the POWER5+.

Performance *per se* is not an issue anymore as, in most cases, the available performance from these high-end storage subsystems is acceptable for most end-users. Hence, performance, connectivity, and throughput can be seen as maximum scalability enablers. But scalability is not an issue because the capacity of the majority of shipped subsystems is below the maximum practical scalability, so these vendors will likely concentrate on functionality in the future.

Virtualization and partitioning form the basis to transform the high-end array control unit into a ubiquitous storage server for the support of other storage media such as tape or optical libraries. These features will allow deployments of real LAN-less, server-less backup, embedded de-duplication or turn-key systems such as medical scanning and archiving systems. Such features increase the functionality gap between high-end and mid-range storage subsystems, which has narrowed over the past few years, and will stem the market-share erosion of high-end enterprise systems.

IBM, which is using System p clusters, is well-positioned to exploit its future server functionality by adding applications to run in DS8000 partitions. Hitachi, using its virtualization layer as a basis, will continue to develop features such as its thin provisioning software to exploit it even further.

Summary

> *"Hitachi storage subsystems have been leading for many years in hardware design and several years ago took the lead in functionality, as well."*
>
> *"In fact, for several years, EMC, once a feature innovation leader, has been but a follower"*

As outlined in this document, high-end storage subsystems are not commodity products as there are significant design and functional differences between the four major storage subsystems. The DMX, the DS8000, and the Universal Storage Platform V are viable solutions and have proven track records in the field. The newcomer, V-MAX, has yet to prove itself, so solid users' testimonies should be requested, particularly when planning for large configurations. Looking at the architectural developments since the '90s until October 2007, Hitachi has been the only company constantly developing its high-end storage subsystems, addressing the changing demands of enterprise customers. Hitachi storage subsystems have been leading for many years in hardware design and several years ago took the lead in functionality, as well. IBM lost ground for several years in the '90s but recovered successfully in the current decade. In October 2007 IBM filled the gap in functionality and has since then been competing with Hitachi's pace of enhancements and new architectural features for its subsystem every few months. EMC stretched the original

Symmetrix design a few years too long, which allowed IBM, Hitachi, and its partners HP and Sun Microsystems (acquired by Oracle) to re-gain some lost territory, particularly in the mainframe storage market. EMC is repeating the “stretching” with the DMX as well as it has remained virtually unchanged since 2005. The only hardware changes between the DMX-3 and DMX-4 were the 4Gb/s back-end, RAID 6, and the solid-state disks, but no new innovative features emerged. In fact, for several years, EMC, once a feature innovation leader, has been but a follower.

Nonetheless, the existence of three companies delivering high-end storage creates a balanced market situation, which ultimately benefits end-users. While hardware, software and overall functionality are important criteria in storage procurement, users should evaluate local support, problem escalation procedures, company culture, and the total cost of ownership over the lifetime of the product as well.