Storage is Not a Commodity – a Comparison of High-End Storage Subsystems

Josh Krischer

Josh Krischer is an expert IT advisor with 37 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 talking 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 sub-systems and their functionalities is much more complicated than designing a server, therefore, lowering the discussion of these storage sub-systems to a commodity level demonstrates a severe lack of knowledge. Today, only three companies continue to develop such systems; EMC, Hitachi and IBM:

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

The fourth high-end sub-system, 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

Three different products with three 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..See fig.1

Hitachi's 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), mirrored data cache, mirrored control cache, channel host Front-end Directors (FeDs) and Back-end Directors (BeDs). The central point of this design is the Application Specific Integrated Circuits (ASICs) of the non-blocking crossbar switch architecture technology, which has

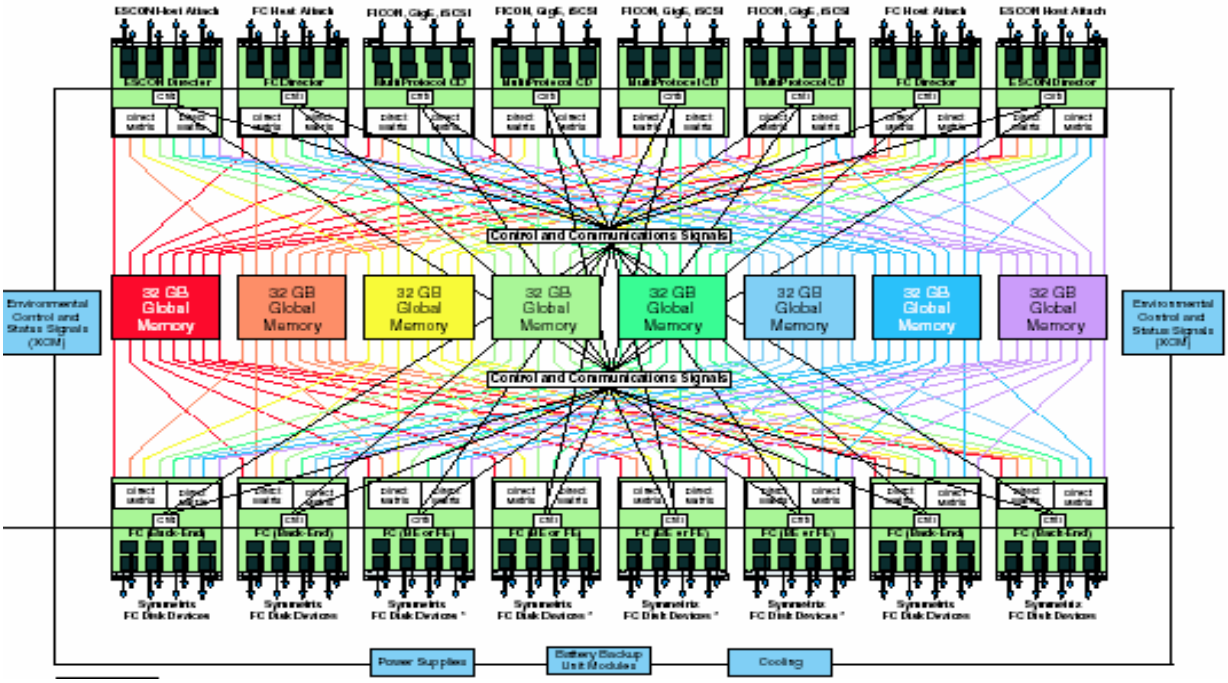

Figure 1; Source EMC

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. As opposed to 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. See fig.2

IBM's DS 8300 structure is based on a 4-way clustered p5 server (p570) with Non-Volatile Storage (NVS). See fig.3

Short History

EMC's Symmetrix DMX was announced in February, 2003 as the follow-on of 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 sub-systems to support SCSI, Fiber Channel, Remote Copy (SRDF) and Point-in-Time copies (TimeFinder), but on the other hand were lacking 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

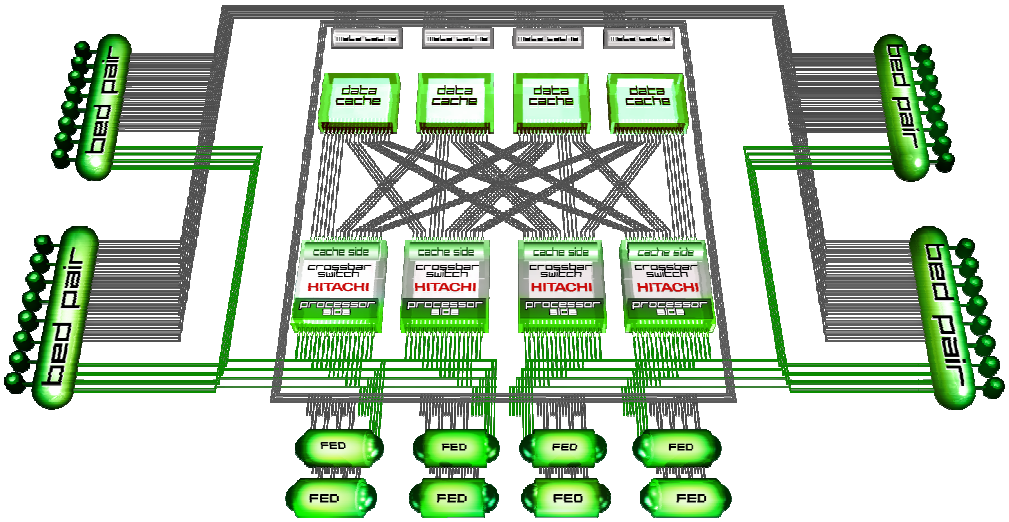

Figure2; Source Hitachi Data Systems

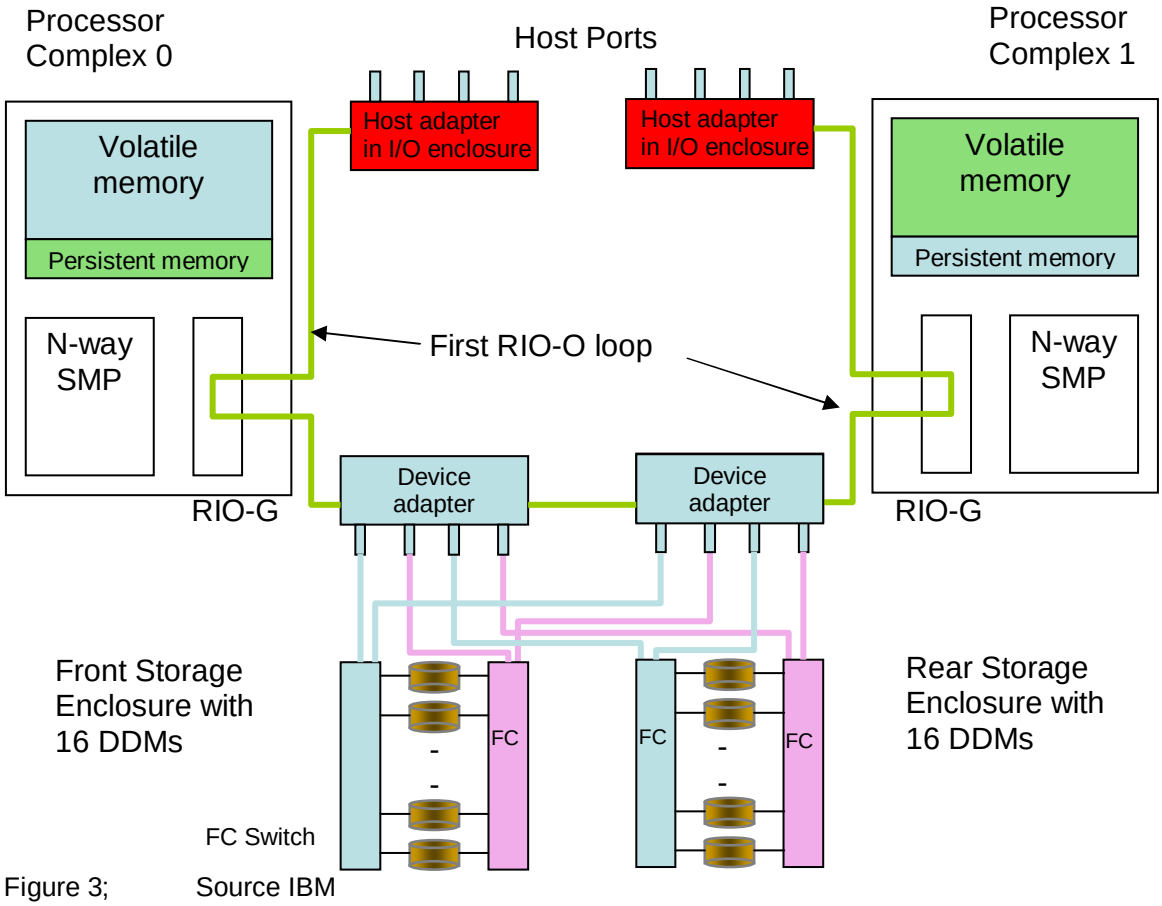

Figure 3; Source IBM

performance and enabled larger scalability In 1999 EMC's long-term road map was to continue with the shared bus structure, increasing the number of the buses and the 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 sub-systems.

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 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, while the DMX-3 was announced in July 2005 and has been shipping since August of the same year. On 16th July 2007 EMC announced the DMX-4 series, mainly enhancing the performance but not adding any new functionalities. 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 enhancements are mainly targeted at reducing the gap with the competition, therefore changing the model nomenclature from DMX-3 to DMX-4 can be seen as a little exaggerated.

Hitachi's Universal Storage Platform (USP) V was announced on May, 14th 2007 and general availability began in June of this year. Hitachi having the fastest "turn-around" times in the industry brought to market in the last twenty years new control unit designs approximately every four to five years with a "mid-life" kicker two years after launching the original product. For example; it was the enterprise 7980-3 storage system in 1990, followed in 1995 by the bus structured 7700 which was the first storage sub-system with a 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 sub-system (Lightning 9900) 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 the same time.

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) - a dimension more in comparison with the other high-end sub-systems available at that time--Hitachi introduced the concept of "Virtual Ports". This embedded virtualization layer provided up to 128 Virtual Ports 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 7 September 2004, HDS 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, which is able to control third party subsystems, and supports partitioning. Hitachi OEM partner Hewlett-Packard and re-seller Sun Microsystems announced the products as the HP StorageWorks XP12000, and the Sun StorEdge 9990.

In May, 2006 HDS announced “mid-life” kicker enhancements to the USP, with a 25% performance boost bringing maximum IOPs through cache to 2.5 million, as well as new security features such as audit logging and extended business continuity and disaster recovery capabilities.

In May 2007 HDS and its partners announced the all-new control unit USP V (HP XP24000, Sun StorageTek 9990) which introduced an expanded virtualization layer, thin provisioning, large logical storage pools, performance boosted to 3.5 million peaks IOPs through cache, and greatly increased scalability (support of up to 247 Petabytes from 32).

IBM’s TotalStorage DS8300 was announced in October, 2004 with first shipments in 2Q05. This is the second version of IBM’s disk storage sub-system based on the “seascape” architecture concepts, a Storage Enterprise Architecture which was 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.

The DS8300 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 sub-system (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 was 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 which was launched in 1994. Anyhow, RAMAC which was never planned as a strategic storage sub-system and was intended to “fill the gap” till the Seascape arrives, couldn’t sell well against the EMC Symmetrix and Hitachi enterprise sub-systems and forced IBM to search for an OEM agreement with StorageTek, which was signed in June, 1996.

Under this contract IBM OEMed StorageTek’s Iceberg sub-system 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 till the end of 2000, IBM, after seeing the ESS working in its lab, practically slowed down the RVA sales in early 1999. StorageTek profited from this deal on a short-term basis, however this deal was like putting a mortgage on the future. From 1999 StorageTek tried to continue to sell the products as SVA but never regained any material market share. Sun Microsystems acquiring StorageTek in 2005 and remained loyal to Hitachi high-end systems,

practically giving SVA *coup de grace* as a storage sub-system but keeping it as a part of the Virtual Storage Management (VSM), a virtual tape sub-system.

The RVA was replaced by IBM's own product, ESS G, followed by the E and F models till the ESS 800 which was announced in June 2002. Similar to its predecessor 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 sub-systems and RAID controller cards offloading RAID functionality from the nodes. The major enhancements to the different ESS models were faster processors (following the System p developments) and larger cache. 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 sub-systems till the launch of the DS8300 in 2005.

How do the three high-end storage sub-systems compare?

Cache structure
Cache structure, bandwidth and connectivity play a crucial role in determining maximum throughput, performance and scalability of the high-end cache-centric storage sub-systems such as EMC's DMX and Hitachi USP. These are also the biggest differences among the three sub-systems.

EMC's DMX has fully mirrored cache; the cache directory is kept in cache as well which means that each access to cache will result in two accesses; one to access the directory and the second to access the data in the cache. The effective cache is less than a half of the purchased cache because of the mirroring and capacity reserved for the directory and the configuration requirements.

Moving from the original Symmetrix to the Symmetrix DMX, EMC didn't change the cache structure which remained as 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 .BIN files loading for LUN assignments, an uncomfortable process which takes time and can corrupt data if not done properly. The last is particularly true when the configuration includes Business Copy Volumes (BCVs) or remote mirrored Symmetrix Remote Data Facility (SRDF) volumes.

The DMX cache is built from 2 to 8 cache modules with capacities between 16 and 512 GBytes. Each Cache module has 8, 1Gbit/s connections 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 GBytes/s bi/directional connections. However, this doesn't mean that the maximum cache bandwidth is 128 GBytes/s because the DMX cache supports a maximum of only 32 concurrent operations[1] (4 concurrent memory transfers per cache module) which only

[1] *" 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"* from EMC Symmetrix DMX architecture guide.

result in a total theoretical 32 GBytes/s for data and control traffic. Each of the 1GB/s serial connections is composed of a pair of full-duplex unidirectional serial links—two 250MB/s serial transfer links (TX), and two 250MB/s serial receive links (RX) which means 0.5 GByte/s in each direction, which, depending on the type of workload may further reduce the practical 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 the cache and each access to cache requires additional access to fetch the metadata, the effective bandwidth is even lower. Considering all the above the maximum achievable bandwidth of the DMX is below 16 GByte/s, much lower than the 128 GByte/s stated in DMX documentation.

Hitachi's USP V has up to 4 data cache modules with a maximum capacity of 256 GBytes for data and up to 32 GBytes of separate, dedicated cache for the metadata. Only the "write portion" of the cache is mirrored when the threshold is automatically adjusted dependents on the activity therefore the effective cache size is reduced by ca.20% for a typical workload. The Dynamic Cache structure and the separate control cache allows dynamic configuration changes in the data cache by changing bits in the control store through a service processor. Hitachi's cache design is the most advanced in the industry and provides any-to-any connectivity between any host port and disk array. 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 GBytes/s for data and 38 GBytes/s for the metadata which aggregates to a total of 106 GBytes/s for the whole cache. The Massively Parallel Universal Star Network Crossbar Switch Architecture supports up to 320 concurrent internal cache and control cache operations which is 10x maximum number of cache operations of the DMX.

IBM's DS8300 SMP cluster structure is difficult to compare with the EMC DMX and the Hitachi USP because the cache is allocated as part of the System p server memory. Instead of using a dedicated cache the DS8300 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 is 32-256 Gbytes. Each side of the cluster has its own cache and the Non Volatile Storage (NVS) of the other cluster, therefore the effective cache size equals the installed capacity. During normal operation, the DS8300 preserves fast writes using the NVS copy in the alternate server. This "cross connection" protects write data loss in case of power loss or other malfunctions.

Front-end connectivity

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

Port Type	EMC DMX-4	HDS USP V	IBM DS8300
FC Ports (4 Gb/s)	64	224	128
Max. FICON Ports	48	112	128
Max. ESCON Ports	64	112	64
iSCSI Ports	48	Planned for 2008	-

Table 1. The DMX-4 specifications were taken on 17th July 2007 from Specification Sheet C1166 published on the EMC web site.

Please pay attention that on 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. 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 sub-system 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 appears as a separate virtual array despite using the same physical port. This ensures safe multi-tenancy as there is no danger of overwriting each 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 supporting heterogeneous clusters and server virtualization.

This is a similar idea as the NPIV (N-Port ID Virtualization) Fiber channel standard which allows a single Fibre 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. The adoption of the NPIV standard is planned for 2008.

Back-end connectivity
EMC's DMX-4 supports in full configuration up to 64 back-end 4Gbps switched Fibre Channel.paths.
Hitachi's USP V supports up to 64, 4Gbps switched fabric paths.
IBM's DS8300 back-end is comprised of 2Gbps dual-redundant switched FC Arbitrated Loops.
The benefit of a switched fabric (point-to-point connectivity) back-end is potentially better performance and easier troubleshooting of faulty drives.

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

Max. Bandwidth (GB/s)	EMC DMX	Hitachi USP V	IBM DS8300
Front-end	256	896	256
Back-end	256	256	128
Cache	32*	106	

* Effective less

Out of the three bandwidths is the smallest one which will have the biggest impact on 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 fact is that the sub-system cannot send more data to the hosts that it receives from the cache.

Scalability

From the three products EMC claims to have the largest scalability supporting up to 1,920 HDDs and a maximum raw capacity of 1,103 TBytes using 300 GB, FC drives or 2,400 Tbytes with the 500 GB "low-cost FC drives" alias FATA.

The Hitachi USP V supports up to 332 TBytes of internal capacity and 247 Petabytes of externally virtualized storage.

The IBM DS8300 supports up to 1,024 300 GB FC or 500 GB FATA drives with a maximum capacity of 512 TBytes.

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 practical installable capacity. The DMX -3,4 with its 64 host ports and limited cache bandwidth will be able to support 1,920 drives only for very low activity environments, therefore EMC's claim: "Symmetrix DMX-3: World's Largest High-end Storage Array" is questionable. The same applies to the DMX-4.

Functions and features

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

- **Hitachi Dynamic Provisioning or "Thin provisioning"** enables allocation of virtual storage as needed without the need 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 the 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.

- **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 Hitachi’s or from a third party. 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 and other applications 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.

- **Virtual Partition Manager** is sub-system 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 can be created, each separately managed and password-protected, to provide better resource allocation and enhanced protection by isolation between the various partitions. This capability enables users to build different internal service levels, to separate test from production and to reduce the costs for users that previously, for data security reasons, may have required separate storage sub-systems.

  IBM’s TotalStorage DS8000 Series supports two storage system partitions as well. As opposed to the Hitachi USP V, each of the DS8000 partitions run separate copies of the DS8000 microcode, which may simplify testing different versions of the microcode, for example.

- **Storage Security Services,** which includes several functions, some of them introduced as early on as the 7700 subsystem of the mid ‘90s. These functions include:
  Controller-based data shredding;
  Write Once Read Many (WORM) software for tamperproof data protection (required by most of the 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. EMC 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.

- **Hitachi Universal Replicator** which is asynchronous, storage-agnostic data replication for internal and externally attached storage. Instead of using cache as other techniques do, this advanced technique is using 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 three high-end subsystems, which may be the subject of another research report.

Performance
Performance has two dimensions; throughput which is measured in the number of I/O operation per second (IOPS), and an acceptable response time in milliseconds. All the vendors claim to have the best performance but only IBM currently participates in the Storage Performance Council (SPC). According to the SPC BENCHMARK1™ from December 5, 2006 the IBM System Storage DS8300 Turbo achieved 120,000 IOPS. Hitachi Data Systems claims that the USP V maximum throughput is 3,500,000 IOPS. This throughput was achieved with 100% cache hit ratio which benchmarks the cache and the front-end but does not represent typical workloads. EMC didn't publish its performance figures for the DMX line.

Availability
All three sub-systems 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 guarantee. In addition to the usual RAID techniques Hitachi's USP models also support RAID-6 in 6D+2 P configurations. This technique consumes 12.5 percent more storage than RAID-5 in 7D+1P configurations, however, 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 the "write penalties," but no such impact should be registered in large blocks of sequential writes. Hitachi announced RAID-6 support in 2005; EMC recently announced RAID 6 for the DMX-4 as well.

Technology

EMC's DMX -3 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 and the Hitachi USP V uses MIPS processors and custom fabricated ASICs as well.
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

Future development

EMC may enhance the DMX in the future with faster processors and a larger cache, but the is if cumbersome static cache architecture design will remain in place; IBM will deploy POWER6 technology, announced on 21st May 2007, which is twice as fast as the POWER5+; and if Hitachi will continue with its current development cycle it will most probably introduce a new high-end product in 2009 or 2010. Performance *per se* is not an issue anymore, in most of the cases the available performance from these three high-end storage sub-systems is acceptable by the end-users; therefore, performance, connectivity and throughput can be seen as maximum scalability enablers.

But scalability is not an issue as well because the capacity of the majority of the shipped sub-systems is below the maximum practical scalability, therefore these vendors will likely concentrate on future functionality.

Virtualization and partitioning are the basis to transform the high-end array control unit into a ubiquitous storage server to support 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. These features increase the functionality gap between high-end and midrange storage systems, 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. 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

As you can read above high-end storage sub-systems are not commodity products; there are significant differences between the three storage sub-systems, however, all are viable solutions and have proven records in the field. Looking at the architectural developments since the '90s it appears that Hitachi was the only company constantly developing its high-end storage subsystems, addressing the changing demands of enterprise customers. Hitachi storage sub-systems have been leading for many years in hardware design and several years ago took the lead in functionality as well. IBM lost several years in the '90s but recovered successfully in the current decade. EMC "stretched" the original Symmetrix design a few years too long which allowed IBM, Hitachi and its partners HP and Sun Microsystems to re-gain some "lost territories." These created a balanced market situation ultimately for the benefits of 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 costs of ownership in the lifetime of the product as well.

Storage is Not a Commodity – a Comparison of High-End Storage Subsystems

Josh Krischer

Josh Krischer is an expert IT advisor with 37 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 talking 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 sub-systems and their functionalities is much more complicated than designing a server, therefore, lowering the discussion of these storage sub-systems to a commodity level demonstrates a severe lack of knowledge. Today, only three companies continue to develop such systems; EMC, Hitachi and IBM:

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

The fourth high-end sub-system, 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

Three different products with three 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.
Hitachi's 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), mirrored data cache, mirrored control cache, channel host Front-end Directors (FeDs) and Back-end Directors (BeDs). The central point of this design is the Application Specific Integrated Circuits (ASICs) of the non-blocking crossbar switch architecture technology, 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. As opposed to 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 Non-Volatile Storage (NVS).

Short History

EMC's Symmetrix DMX was announced in February, 2003 as the follow-on of 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 sub-systems to support SCSI, Fiber Channel, Remote Copy (SRDF) and Point-in-Time copies (TimeFinder), but on the other hand were lacking 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 of the buses and the 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 sub-systems.

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 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, while the DMX-3 was announced in July 2005 and has been shipping since August of the same year. On 16th July 2007 EMC announced the DMX-4 series, mainly enhancing the performance but not adding any new functionalities. 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 enhancements are mainly targeted at reducing the gap with the competition, therefore changing the model nomenclature from DMX-3 to DMX-4 can be seen as a little exaggerated.

Hitachi's Universal Storage Platform (USP) V was announced on May, 14th 2007 and general availability began in June of this year. Hitachi having the fastest "turn-around" times in the industry brought to market in the last twenty years new control unit designs approximately every four to five years with a "mid-life" kicker two years after launching the original product. For example; it was the enterprise 7980-3 storage system in 1990, followed in 1995 by the bus structured 7700 which was the first storage sub-system with a 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 sub-system (Lightning 9900) 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 the same time.

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) - a dimension more in comparison with the other high-end sub-systems available at that time--Hitachi introduced the concept of "Virtual Ports". This embedded virtualization layer provided up to 128 Virtual Ports 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 7 September 2004, HDS 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, which is able to control third party subsystems, and supports partitioning. Hitachi OEM partner Hewlett-Packard and re-seller Sun Microsystems announced the products as the HP StorageWorks XP12000, and the Sun StorEdge 9990.

In May, 2006 HDS announced "mid-life" kicker enhancements to the USP, with a 25% performance boost bringing maximum IOPs through cache to 2.5 million, as well as new security features such as audit logging and extended business continuity and disaster recovery capabilities.

In May 2007 HDS and its partners announced the all-new control unit USP V (HP XP24000, Sun StorageTek 9990) which introduced an expanded virtualization layer, thin provisioning, large logical storage pools, performance boosted to 3.5 million peaks IOPs through cache, and greatly increased scalability (support of up to 247 Petabytes from 32).

IBM's TotalStorage DS8300 was announced in October, 2004 with first shipments in 2Q05. This is the second version of IBM's disk storage sub-system based on the "seascape" architecture concepts, a Storage Enterprise Architecture which was 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.

The DS8300 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 sub-system (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 was 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 which was launched in 1994. Anyhow, RAMAC which was never planned as a strategic storage sub-system and was intended to "fill the gap" till the Seascape arrives, couldn't sell well against the EMC Symmetrix and Hitachi enterprise sub-systems and forced IBM to search for an OEM agreement with StorageTek, which was signed in June, 1996.

Under this contract IBM OEMed StorageTek's Iceberg sub-system 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 till the end of 2000, IBM, after seeing the ESS working in its lab, practically slowed down the RVA sales in early 1999. StorageTek profited from this deal on a short-term basis, however this deal was like putting a mortgage on the future. From 1999 StorageTek tried to continue to sell the products as SVA but never regained any material market share. Sun Microsystems acquiring StorageTek in 2005 and remained loyal to Hitachi high-end systems, practically giving SVA *coup de grace* as a storage sub-system but keeping it as a part of the Virtual Storage Management (VSM), a virtual tape sub-system.

The RVA was replaced by IBM's own product, ESS G, followed by the E and F models till the ESS 800 which was announced in June 2002. Similar to its predecessor 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 sub-systems and RAID controller cards offloading RAID functionality from the nodes. The major enhancements to the different ESS models were faster processors (following the System p developments) and larger cache. 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 sub-systems till the launch of the DS8300 in 2005.

How do the three high-end storage sub-systems compare?

Cache structure
Cache structure, bandwidth and connectivity play a crucial role in determining maximum throughput, performance and scalability of the high-end cache-centric storage sub-systems such as EMC's DMX and Hitachi USP. These are also the biggest differences among the three sub-systems.

EMC's DMX has fully mirrored cache; the cache directory is kept in cache as well which means that each access to cache will result in two accesses; one to access the directory

and the second to access the data in the cache. The effective cache is less than a half of the purchased cache because of the mirroring and capacity reserved for the directory and the configuration requirements.

Moving from the original Symmetrix to the Symmetrix DMX, EMC didn't change the cache structure which remained as 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 .BIN files loading for LUN assignments, an uncomfortable process which takes time and can corrupt data if not done properly. The last is particularly true when the configuration includes Business Copy Volumes (BCVs) or remote mirrored Symmetrix Remote Data Facility (SRDF) volumes.

The DMX cache is built from 2 to 8 cache modules with capacities between 16 and 512 GBytes. Each Cache module has 8, 1Gbit/s connections 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 GBytes/s bi/directional connections. However, this doesn't mean that the maximum cache bandwidth is 128 GBytes/s because the DMX cache supports a maximum of only 32 concurrent operations[1] (4 concurrent memory transfers per cache module) which only result in a total theoretical 32 GBytes/s for data and control traffic. Each of the 1GB/s serial connections is composed of a pair of full-duplex unidirectional serial links—two 250MB/s serial transfer links (TX), and two 250MB/s serial receive links (RX) which means 0.5 GByte/s in each direction, which, depending on the type of workload may further reduce the practical 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 the cache and each access to cache requires additional access to fetch the metadata, the effective bandwidth is even lower. Considering all the above the maximum achievable bandwidth of the DMX is below 16 GByte/s, much lower than the 128 GByte/s stated in DMX documentation.

Hitachi's USP V has up to 4 data cache modules with a maximum capacity of 256 GBytes for data and up to 32 GBytes of separate, dedicated cache for the metadata. Only the "write portion" of the cache is mirrored when the threshold is automatically adjusted dependents on the activity therefore the effective cache size is reduced by ca.20% for a typical workload. The Dynamic Cache structure and the separate control cache allows dynamic configuration changes in the data cache by changing bits in the control store through a service processor. Hitachi's cache design is the most advanced in the industry and provides any-to-any connectivity between any host port and disk array. 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

[1] *" 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"* from EMC Symmetrix DMX architecture guide.

available bandwidth is 68 GBytes/s for data and 38 GBytes/s for the metadata which aggregates to a total of 106 GBytes/s for the whole cache. The Massively Parallel Universal Star Network Crossbar Switch Architecture supports up to 320 concurrent internal cache and control cache operations which is 10x maximum number of cache operations of the DMX.

IBM's DS8300 SMP cluster structure is difficult to compare with the EMC DMX and the Hitachi USP because the cache is allocated as part of the System p server memory. Instead of using a dedicated cache the DS8300 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 is 32-256 Gbytes. Each side of the cluster has its own cache and the Non Volatile Storage (NVS) of the other cluster, therefore the effective cache size equals the installed capacity. During normal operation, the DS8300 preserves fast writes using the NVS copy in the alternate server. This "cross connection" protects write data loss in case of power loss or other malfunctions.

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

Port Type	EMC DMX-4	HDS USP V	IBM DS8300
FC Ports (4 Gb/s)	64	224	128
Max. FICON Ports	48	112	128
Max. ESCON Ports	64	112	64
iSCSI Ports	48	Planned for 2008	-

Table 1. The DMX-4 specifications were taken on 17th July 2007 from Specification Sheet C1166 published on the EMC web site.

Please pay attention that on 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. 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 sub-system 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 appears as a separate virtual array despite using the same physical port. This ensures safe multi-tenancy as there is no danger of overwriting each 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 supporting heterogeneous clusters and server virtualization.

This is a similar idea as the NPIV (N-Port ID Virtualization) Fiber channel standard which allows a single Fibre 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. The adoption of the NPIV standard is planned for 2008.

Back-end connectivity
EMC's DMX-4 supports in full configuration up to 64 back-end 4Gbps switched Fibre Channel.paths.
Hitachi's USP V supports up to 64, 4Gbps switched fabric paths.
IBM's DS8300 back-end is comprised of 2Gbps dual-redundant switched FC Arbitrated Loops.
The benefit of a switched fabric (point-to-point connectivity) back-end is potentially better performance and easier troubleshooting of faulty drives.

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

Max. Bandwidth (GB/s)	EMC DMX	Hitachi USP V	IBM DS8300
Front-end	256	896	256
Back-end	256	256	128
Cache	32*	106	

* Effective less
Out of the three bandwidths is the smallest one which will have the biggest impact on 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 fact is that the sub-system cannot send more data to the hosts that it receives from the cache.

Scalability

From the three products EMC claims to have the largest scalability supporting up to 1,920 HDDs and a maximum raw capacity of 1,103 TBytes using 300 GB, FC drives or 2,400 Tbytes with the 500 GB "low-cost FC drives" alias FATA.

The Hitachi USP V supports up to 332 TBytes of internal capacity and 247 Petabytes of externally virtualized storage.

The IBM DS8300 supports up to 1,024 300 GB FC or 500 GB FATA drives with a maximum capacity of 512 TBytes.

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 practical installable capacity. The DMX -3,4 with its 64 host ports and limited cache bandwidth will be able to support 1,920 drives only for very low activity environments, therefore EMC’s claim: “Symmetrix DMX-3: World’s Largest High-end Storage Array” is questionable. The same applies to the DMX-4.

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

- **Hitachi Dynamic Provisioning or “Thin provisioning”** enables allocation of virtual storage as needed without the need 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 the 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.

- **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 Hitachi’s or from a third party. 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 and other applications 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.

- **Virtual Partition Manager** is sub-system 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 can be created, each separately managed and password-protected, to provide better resource allocation and enhanced protection by isolation between the various partitions. This capability enables users to build different internal service levels, to separate test from production and to reduce the costs for users that previously, for data security reasons, may have required separate storage sub-systems.

IBM's TotalStorage DS8000 Series supports two storage system partitions as well. As opposed to the Hitachi USP V, each of the DS8000 partitions run separate copies of the DS8000 microcode, which may simplify testing different versions of the microcode, for example.

- **Storage Security Services,** which includes several functions, some of them introduced as early on as the 7700 subsystem of the mid '90s. These functions include:
  Controller-based data shredding;
  Write Once Read Many (WORM) software for tamperproof data protection (required by most of the 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. EMC 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.

- **Hitachi Universal Replicator** which is asynchronous, storage-agnostic data replication for internal and externally attached storage. Instead of using cache as other techniques do, this advanced technique is using 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 three high-end subsystems, which may be the subject of another research report.

Performance
Performance has two dimensions; throughput which is measured in the number of I/O operation per second (IOPS), and an acceptable response time in milliseconds. All the vendors claim to have the best performance but only IBM currently participates in the Storage Performance Council (SPC). According to the SPC BENCHMARK1™ from December 5, 2006 the IBM System Storage DS8300 Turbo achieved 120,000 IOPS. Hitachi Data Systems claims that the USP V maximum throughput is 3,500,000 IOPS. This throughput was achieved with 100% cache hit ratio which benchmarks the cache and the front-end but does not represent typical workloads. EMC didn't publish its performance figures for the DMX line.

Availability
All three sub-systems 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 guarantee. In addition to the usual RAID techniques Hitachi's USP models also support RAID-6 in 6D+2 P configurations. This technique consumes 12.5 percent more storage than RAID-5 in 7D+1P configurations, however, 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 the "write penalties," but no such impact should be registered in large blocks of sequential writes. Hitachi announced RAID-6 support in 2005; EMC recently announced RAID 6 for the DMX-4 as well.

Technology

EMC's DMX -3 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 and the Hitachi USP V uses MIPS processors and custom fabricated ASICs as well.
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

Future development
EMC may enhance the DMX in the future with faster processors and a larger cache, but the is if cumbersome static cache architecture design will remain in place; IBM will deploy POWER6 technology, announced on 21st May 2007, which is twice as fast as the POWER5+; and if Hitachi will continue with its current development cycle it will most probably introduce a new high-end product in 2009 or 2010. Performance *per se* is not an issue anymore, in most of the cases the available performance from these three high-end storage sub-systems is acceptable by the end-users; therefore, performance, connectivity and throughput can be seen as maximum scalability enablers.

But scalability is not an issue as well because the capacity of the majority of the shipped sub-systems is below the maximum practical scalability, therefore these vendors will likely concentrate on future functionality.

Virtualization and partitioning are the basis to transform the high-end array control unit into a ubiquitous storage server to support 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. These features increase the functionality gap between high-end and midrange storage systems, 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. 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

As you can read above high-end storage sub-systems are not commodity products; there are significant differences between the three storage sub-systems, however, all are viable solutions and have proven records in the field. Looking at the architectural developments since the ’90s it appears that Hitachi was the only company constantly developing its high-end storage subsystems, addressing the changing demands of enterprise customers. Hitachi storage sub-systems have been leading for many years in hardware design and several years ago took the lead in functionality as well. IBM lost several years in the ’90s but recovered successfully in the current decade. EMC “stretched” the original Symmetrix design a few years too long which allowed IBM, Hitachi and its partners HP and Sun Microsystems to re-gain some “lost territories.” These created a balanced market situation ultimately for the benefits of 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 costs of ownership in the lifetime of the product as well.