

February 2009



Solid state disks for enterprise storage

IBM's approach to new storage technology

Clodoaldo Barrera and Stephen Edel

Solid state disks in enterprise applications—an introduction

Enterprise storage applications present IT operations staff with a demanding mix of requirements, and historically have been the drivers of innovation by the vendor community. In recent years, the demands upon data center storage systems have escalated in all dimensions: storage capacities are growing; there is an increased need for 24X7 data; Web-based applications create high access rates; and management costs and power constraints are restraining growth. Users would like instantaneous access to “all the data, all the time,” but technology and finances set boundaries for what can be practically achieved, and these boundaries create the hurdle (and opportunity) for innovation. There continues to be a long list of unmet needs for storage systems with armies of engineers and entrepreneurs seeking to satisfy them.

The “object of desire” for storage administrators would be a much faster and cheaper storage technology that was fully compatible with existing software. Failing that, it would be useful to have a much faster technology that cost more, provided that the extra cost was not excessive. In addition, it is critical that the added complexity of a new storage type be hidden from operations staff, or the added cost of management would prevent its use. In large operations, it is unrealistic to expect administrators to manually provision storage by type, to monitor performance, and to re-position data when performance goals are not met.

Hard disk drives (HDDs), with magnetic recording as the core technology, have been the workhorses of enterprise storage for 50 years, and will continue to be. Advanced technologies including Tunnel Magnetoresistive read heads, bit-patterned media and heat-assisted magnetic recording promise to extend areal densities and thereby providing ever lower \$/GB price points.

However, there are significant limits to increasing HDD performance as measured by both input/output operations per second (IO/sec) and sustained throughput (MB/sec), driven by spindle speeds and actuator movement when accessing data. These limitations are increasingly problematic for transaction-oriented applications and specific middleware and system software. In addition, with multi-core CPUs, the gap between the processor’s ability to consume data and the hard disk drive’s ability to provide that data is widening.

The need for higher IO/sec rates in enterprise storage creates a need for a new storage type, known as solid state disk (SSD). SSDs have existed for some time, built in the past with volatile DRAM memory supplemented with battery backup. However, these were enormously expensive—often up to 1,000 times the cost of a high-performance disk drive with an equivalent capacity. Some applications do require very high IO/sec rates, and there are some military or industrial use cases that benefit from insensitivity to shock and vibration. But the vast majority of applications cannot justify the extra cost, so SSDs have remained a rarity.

A new storage device type based on non-volatile flash memory is now available for enterprise workloads that require high IO/sec. Flash memory is less than 10 percent of the cost of DRAM today, and expected innovations should lower the cost by another order of magnitude in the coming two to three years. By itself, however, flash memory is too slow performing WRITE operations and has an unacceptably short life. Clever engineering has been applied to use flash memory in combination with DRAM and embedded software to create a device for enterprise applications, one that is interface and function compatible with an HDD, but has substantially higher IO/sec performance. These devices made their appearance in 2008, and will be widely available as optional devices in disk arrays. Flash-based SSDs are still more costly than HDDs of equivalent capacity, but for high IO/sec workloads, they can cost less.

The availability of a new type of storage device raises questions, including:

- *What is the right use of the new devices?*
- *Are additional changes to the server, storage and software stack required to see the full value at the application level?*
- *How do we place the right data on these (expensive) devices at the right time to achieve performance objectives without creating an undue burden of management?*

To answer these questions, we must quantify the performance improvements SSDs can provide, and describe how they will be deployed and managed. Flash-based SSDs can be used instead of a much larger number of high-performance/low-capacity HDDs to deliver an overall IO/sec rate demanded by an application. These SSDs must be packaged in a system configuration that includes management function. IBM will deliver SSD technology in a phased rollout, with special attention to ease of management as products are introduced.

Finally, it is important to ask whether flash memory-based SSDs are the final word in storage devices. While it is premature to conclude this in the near term, IBM and many other technology companies believe that there is enormous promise in several different families of technology that could, in the span of five years or more, provide cheaper and more robust storage devices than those built on flash, and eventually rival the \$/GB of HDDs.

Enterprise storage—the need for technical innovation

Enterprise storage capacities have grown dramatically over the past decade, and this trend is expected to continue. In most years, the increased capacity consumption exceeds the cost takedown of data center storage, resulting in a sustained 5 to 7 percent growth rate in overall spending. There are several drivers of storage growth, including:

- *An increasing need to capture data from and serve data to Web clients, resulting in a large number of points of access for data housed in the data center. This scale of access is seen in virtually all industries, including Web 2.0 with social networking applications, and financial services with online banking and ATMs.*
- *Increasing sizes of data objects, with growth in image and video applications.*
- *A need to retain data for extended periods of time to satisfy regulatory or business process obligations.*
- *A need to replicate data to meet disaster recovery/business resiliency requirements.*

In addition to capacity growth, there is an increasing need to process data quickly. In some cases, the high volume of clients accessing a database can result in the need for a high IO/sec rate. In other cases, there is a need to quickly process data on ingest, and to index the incoming streams to allow high-speed searches and retrieval of needed records. Within the database or file system middleware there are directories, lock managers and access control records that must be accessed and updated at rates that scale with the size of the data or the number of clients being served. Finally, there are some applications that simply cannot be run fast enough to satisfy their business need. These include trading algorithms, complex simulations in aerospace or pharmaceutical design, and security video analysis. Online transaction processing (OLTP) systems are the classic example of these applications. Many of these uses of data create a need to operate at high speed, often on indices or subsets of larger collections of information. In cases where the IO/sec performance of the storage is the system bottleneck, there is a high value in faster storage.

Over the years, HDDs have maintained a dramatic rate of improvement in \$/GB, which has enabled data center administrators to keep up with storage capacity demand without greatly increasing expenses. HDDs have also performed relatively well in achieving improvements to sustained bandwidths (GB/s) with recent 15 KRPM drives advertising greater than 170 MB/sec speeds. But the rate of improvement in IO/sec has lagged far behind other system elements. While \$/GB has improved at a rate of 50 percent per year or more over the last decade, IO/sec has limped along at a rate of 5 percent per year, and has slowed even further in recent years.¹

SEAGATE 15 KRPM/3.5" DRIVE SPECIFICATIONS¹

Product Availability	2002	2008	CGR
Generation	15K.3	15K.6	
Capacity (GB)	73	450	35%
Max Sustained DR (MB/sec)	75	171	15%
Read Seek Times (ms)	3.6	3.4	1%

This parameter of performance is dominated by the mechanical elements of the drive: the drive's rotational speed and the seek time of the arm. Substantial improvements for each drive form factor (for example, 3.5") are no longer attainable. Although some minor improvement is possible by using smaller drives (for example, 2.5", 15 KRPM), this comes at a higher \$/GB cost. This lack of improvement in IO/sec means that drives are actually getting worse in access density as defined by IO per GB per second (IO/GB/sec).

For applications that do not have strong performance needs, or that measure performance in sustained throughput (GB/sec), HDDs can continue to do relatively well. The bandwidths of data from disk storage can be made even better when striping data across multiple disks. But for applications or middleware requiring high IO/sec, there is a large gap between need and capability.

Another way to look at the IO performance problem is to ask how long a single access to storage takes. There is a large disparity between the access time for data from volatile memory compared to the access time from non-volatile memory or storage. Today an access to DRAM is 60 nsec, compared to a 5 msec access to disk, or a 40 second access to tape. In human terms, this would mean waiting one second to access data from memory, two months for disk, and 1,600 years for data from tape.

Designers of computer systems and software are aware of the performance disparity between memory and storage today, and they make decisions to minimize its impact. Systems are built with large memories and designed to keep a working set of data in memory. Disk accesses are aggregated, and large sequential blocks of data are rolled in and out of memory whenever possible. Disk arrays include DRAM caches and improved algorithms that minimize accesses to the drives themselves. HDDs also contain a modest amount of memory to reduce mechanical access to the media. Even with these designs, there are some workloads that require large numbers of small record accesses, and suffer from the limitations of drive performance.

An example is useful to understand the plight of a storage administrator trying to meet user needs. Let us suppose that a database can be housed in 5 TB of disk storage, that there are an equal number of READs and WRITEs, and that an effective rate of 15,000 reads and 15,000 writes per second are needed from the storage devices to support the database. If we further assume that the data is protected by a RAID 5 algorithm, so that every READ creates a single IO and every WRITE creates four additional IO accesses (READ original data, READ original parity, WRITE new data, WRITE new parity), the effective IO rate at the HDDs needs to be 75,000 IO/sec. Assuming 150 IO/sec per drive, this would require 500 drives, which could consume as much as two to three full racks of disks.

Even if the fastest, smallest capacity available disk drives were chosen (for example, 73 GB/15 KRPM), the administrator would actually deploy 36 TB of capacity—much larger than the required 5 TB. In a real deployment, we would use substantially more than 500 drives, because we cannot count on a perfectly even distribution of IO across all the spindles. So HDDs continue to provide value on a \$/GB metric but are getting worse in IO/GB—what will be done to fix this problem?

Solid state disks—historical perspectives and the promise of new technology

The example above allows us to describe the attributes that a better technology would need. The desirable characteristics would include a non-volatile device in the general capacity range of an HDD that is capable of substantially higher IO/sec rates—100 times higher or more. The technology would provide a good balance of READ and WRITE performance. The device should be at least as reliable and last as long as an HDD, and it should consume equal or less power per GB. Finally, the cost ideally should be no more than 10 times (hopefully less) than the cost of an equivalent capacity HDD. At 10 times the price and 100 times the performance, the value proposition is very compelling for high IO/sec applications. Devices that get close to these parameters are now available, and show great promise in addressing the niche we describe above. These are SSDs built mainly from flash memory and augmented with DRAM along with a rather sophisticated internal controller.

Flash memory based on the NAND (the logical “Not And” operation) has been available for nearly two decades, and is used in several high-volume consumer electronics applications, including cell phones, PDAs and MP3 players. Flash memory is non-volatile (it retains data without a power source), involves no mechanical parts, and can be manufactured as standard components in high volume. Current flash technology manipulates a charge on the floating gate of specially designed transistors to allow representation of two (voltage) states, which translates to a single bit per cell, and is called single layer cell or SLC. SLC NAND-based flash has been dropping in price faster than DRAM and HDDs, and now sits between them in a price range approximately 20 to 30 times more expensive than HDDs but up to 10 times cheaper than DRAM.

The next evolution in flash technology is multiple layer cell or MLC. MLC is designed to allow a more precise amount of charge on the floating gate of the transistor to represent four different states, thereby translating to two bits of information per cell. This higher density per cell and the potential to store three or more bits of information per cell provides for continuing cost improvement in the coming years. If MLC is successful, flash memory could reach down to only two to three times the price of high-performance HDDs (FC or SAS 15 KRPM) in the foreseeable future, close enough to drive a significant substitution rate of low-capacity/high-performance HDDs.

As mentioned earlier, SSDs have also been in use for more than 20 years in applications that include military and other industrial deployments where extreme temperatures, shock and vibration were too significant for HDDs to handle. However, several new uses for flash-based SSDs have begun to emerge.

Servers have begun to use modest capacity SSDs (10 to 30 GB) as boot devices and to hold paging data. Although adding SSDs to servers certainly improves performance, there are additional attractive features of SSDs, including improved reliability (compared to mechanical HDDs) and significantly lower power. For example, IBM ships BladeCenter® servers that can utilize both 16 GB solid state drives and 8 GB flash

drives. The flash drive can be used as a Linux® operating system boot drive and as a storage device to complement shared storage on the IT network. SDDs are also being adopted by notebook manufacturers for their low weight and to increase battery life while accelerating application performance to rival desktops. In these uses, SDDs are sold at a premium compared to high-capacity HDDs. These non-enterprise applications will generate increasing volumes for SDDs, which in turn will decrease the price of enterprise SSDs, a related but quite different animal.

Solid state disk challenges

One important fact about enterprise computing today is that it can no longer afford unique basic components. Almost all of the core technologies found in data centers are versions of technology developed for higher volume markets—typically consumer electronics or desktop and mobile computing devices. This means that enterprise equipment designers prefer to take advantage of high-volume technologies (warts and all); they can achieve scale while masking the deficiencies of the underlying technology. This also applies to solid state disks for the enterprise.

From an enterprise storage perspective, NAND-based flash memory technologies have some challenges.

Performance: WRITEs to a flash cell require that the cell first be “erased,” after which it can be programmed with new data. The erasure procedure (typically 1.5 to 2.0 ms) is quite long compared to everything else in a memory-based system and nearly as long as a seek on an HDD. While SSD READ performance (on a sector basis) can be sustained at 20 MB/sec to 40 MB/sec, WRITE performance significantly lags at only 1 MB/sec to 5 MB/sec sustained. The good news is that lower WRITE performance can be compensated for by connecting multiple flash devices together on multiple flash ports to allow data to be striped and operations performed in parallel.

Endurance: Flash-based SSDs by their design wear out after repetitive data WRITES due to charges that get trapped in the dielectric oxide layer between the substrate and the floating gate of the cell. In SLC technology, the limit on the number of WRITES is approximately 1,000,000, which, under some stressed usage cases could be exceeded in a matter of days. For MLC, the number of WRITES can range from 10,000 to 100,000, making MLCs even 10 to 100 times less durable than SLCs. In addition, SSDs (like HDDs) have bad sectors or blocks that can occur over time. Other failures include “program disturb” errors (bits accidentally changing state during programming) “READ disturb” (bits accidentally changing state during a READ), and other similar scenarios.

Fortunately engineers have been working on creative methods to alleviate these issues. As with HDDs, SSDs include controllers that employ error correction codes (ECC) to correct bit errors as they are read back from the flash memory. The controllers also track and manage bad blocks. Wear-leveling techniques are used to map different physical locations of the memory to logical addresses during re-programming to reduce the number of WRITES to any one cell. Extra flash capacity is reserved in the device to improve life and act as spare space as cells die.

For enterprise storage, highly specialized devices will combine the benefits of DRAM memory, flash memory, with sophisticated controllers to buffer frequently accessed or critical data in DRAM (demand paging) and to eliminate WRITES while evenly distributing data across the flash memory to deliver a more robust SSD. The result is an SSD that is much more than just a large flash memory, providing higher and more balanced READ and WRITE performance (for example, one device claims 50 K READS/sec and 17 K WRITES/sec), and provide an equivalent life (five years or longer) to HDDs under heavy WRITE workloads. With additional technology and extra engineering, these devices will be priced higher than the price of their flash content, but will track downward in price as flash prices decline and as multiple vendors compete for share. These devices are new, so some problems are to be expected. And, like other new technologies, they will experience a learning curve that will improve reliability and price.

Solid state disks—enterprise storage applications

Once we have an enterprise-class SSD, we need to understand how to use it optimally in a data center. As discussed above, SSDs can deliver high IO/sec, but currently at a higher price. It is important to place the right kind of data on the SSD, data that needs the higher performance, and not waste it on those files or workloads that will not benefit substantially. High-transaction performance databases, database logs and large file directories are typically shared across multiple servers in a SAN configuration. This data requires snapshots, backup services and remote replication typically provided by enterprise disk array controllers. Therefore, inclusion of SSDs in a SAN-attached disk array seems like a reasonable thing to do.

If we revisit the example of a 5 TB database with a 15 K IO/sec READ and 15 K IO/sec WRITE requirement, users can potentially realize substantial savings as the READ and WRITE performance needs can be met by a single SSD. To reach the 5 TB capacity, we will need to use 20 drives of 250 GB each. In a RAID 5 configuration of 5+P+S, this would require 28 drives. (Since SSDs can fail, they should be protected with a RAID configuration just as HDDs are when managed by an array controller.) An SSD is commonly designed to conform to the same package size as an HDD (2.5", 3.5"), while consuming only 50 to 60 percent of the power. Twenty-eight SSD drives could fit into a 4U enclosure. In 2008 the SSD configuration cost more in purchase price than the HDDs. However, the 28 SSDs can deliver 10 times the IO/sec rate of the 500 HDDs, while consuming less than 5 percent of the power and space.

In the example above, we are only comparing the activity at the device level. To understand the benefits at the application level, the behavior of the entire system must be modeled or measured.

IBM's roadmap for SSDs in enterprise applications

IBM engineers and researchers have investigated the question of the optimal use of SSDs in enterprise applications, and laid out a roadmap for their use. We see three distinct types of enterprise architecture that will coexist in the market but will be staggered somewhat in the timing of their introduction. In addition to the system architectures, there is a need for middleware awareness of SSDs, and management software to control data placement, load balancing and error recovery. Finally, there will be an evolution of the basic devices themselves, discussed in the section below.

The first phase of SSD use will be as devices in one or more RAID ranks managed by an existing cached RAID controller. This design benefits from the fact that SSD vendors have carefully duplicated the form factor and function of an enterprise HDD. The failure characteristics of an SSD will be somewhat new; systems will need to monitor different SMART attributes that could track cell failure and excessive ECC correction activity.

IBM has announced the availability of SSDs for use within the DS8000® Storage System, with attachment to both System z® through FICON®, and Microsoft® Windows®/UNIX®/Linux systems through FCP. SSDs provide a faster device, but as noted above, application benefits depend on the entire system. With High-Performance FICON, Parallel Access Volumes, advanced caching technology, and Pre-Deposit Writes with FCP remote copy, the DS8000 provides an end-to-end system that enables applications to take full advantage of these higher performance devices. Given their high costs, such full-system optimization is essential when evaluating their cost/performance value. No other vendor can offer this optimized approach. IBM will introduce SSDs into other disk array models in 2009.

In addition to the DS8000 hardware system, IBM has added SSD-awareness to the automated data placement facility on System z. The DFSMS (Data Facility–System Managed Store) management software allows users to define Storage Pools, made up of specific storage types that now include SSD storage in DS8000. DFSMS also defines Data Classes, with which a user identifies the needs of data sets. The DFSMS facility then automatically provisions the data requiring the highest performance to the fastest storage. To aid administrators in assigning Data Classes, analysis tools are available to characterize the performance of data with its current storage assignment, and recommend which data would benefit the most from placement on SSDs.

IBM has also begun investigation into other models for SSD use in Enterprise storage. In 2008, IBM showed a prototype storage system optimized for SSDs, based on IBM's powerful SAN Volume Controller (SVC). SVC in combination with IBM DS4000® arrays is already the world leader in Storage Performance Council (SPC) benchmarks. Although it is not yet an official SPC benchmark, an SVC controller was demonstrated by the Project Quicksilver "Proof of Concept" to provide LUN management and SAN access to SSD data with over 1,000,000 IO/sec—four times the current SPC record, also owned by SVC. Products based on Quicksilver technology will be introduced by IBM as the second phase of enterprise SSDs: external array controllers optimized for SSD devices.

Finally, SSDs can be used as an extension of the memory space of servers or of cache in storage controllers. This is a different thought than the current use of SSDs as a boot image and private paging store for servers. In this model the SSDs within a server cluster represent shared system data, and will have high-performance memory model access. This use case will require modifications to the operating system or middleware to be properly exploited, but when used should offer substantial performance benefits to clustered applications.

For all of these models, management software will be needed to understand the capability of the devices, and to correctly place data to achieve the most benefit while extending endurance. SSDs will remain expensive for the immediate future, so management awareness will be a requirement. As more complex use cases are introduced, management software will be called upon to hide the complexity of an additional storage tier, and to automate data placement and movement between tiers.

The future of SSDs

IBM believes that there is still a long road ahead in the use of SSDs. The current generation of SLC NAND flash devices will offer important value as described in our example above, but this is hardly the final word. SSDs based on SLC flash will complement HDDs, but probably will never be inexpensive enough to significantly replace high-performance FC and SAS 15 KRPM drives. MLC with the capability of two or more bits per cell will lower the cost of the basic components and more rapidly replace high-performance HDDs, but will introduce another round of challenges with cell life and device reliability. It is possible that these challenges can be addressed with yet more clever algorithms with SSD controllers. It is also possible that the expected time to market of these MLC-based devices will be delayed.

If the issues with MLC-based devices can be successfully resolved, we could see a return to a three-tier storage hierarchy in the enterprise: MLC-Flash SSDs as the top tier, capacity-optimized SATA HDDs as the middle tier, and tape as the third. A major substitution of SSDs for performance optimized HDDs is probably three to five years away at best.

There are other technologies, less mature than flash, but more promising in their ultimate cost and reliability. IBM has a research investment in two of these: Racetrack Memory, and Phase Change Memory. Either of these may be five or more years away from product readiness, but they have the promise of achieving the long-term goal of lower cost than spinning magnetic technology with access speeds closer to DRAM and long device life. If this goal is finally realized, a major transformation in data center technology will be upon us. And a venerable technology that has served the enterprise data center and many other industries already for more than 50 years may finally be replaced.



For more information

To learn more about solid state disks from IBM, please contact your IBM marketing representative or IBM Business Partner, or visit the following Web site: ibm.com/systems/storage

© Copyright IBM Corporation 2009

IBM Systems and Technology Group
Route 100
Somers, NY 10589

Produced in the United States of America
February 2009
All Rights Reserved

IBM, the IBM logo and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at ibm.com/legal/copytrade.shtml

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft and Windows are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other product, company or service names may be trademarks or service marks of others.

¹ Source: Seagate data sheets @ www.seagate.com



Recyclable, please recycle

