Trends in Storage Technologies

E. Eleftheriou, R. Haas, J. Jelitto, M. Lantz and H. Pozidis
IBM Research – Zurich, 8803 Rüschlikon, Switzerland

Abstract

The high random and sequential I/O requirements of contemplated workloads could serve as impetus to move faster towards storage-class memory (SCM), a technology that will blur the distinction between memory and storage. On the other hand, the volume of digital data being produced is increasing at an ever accelerating pace. Recent technology demonstrations indicate that low-cost, low-power tape storage is well positioned to scale in the future and thus serve as archival medium. In this paper, we will look at the implications of these technology extremes, namely, SCM and tape, on the storage hierarchy.

1 Introduction

One of the challenges in enterprise storage and server systems is the rapidly widening gap between the performance of the hard-disk drives (HDD) and that of the remainder of the system. Moreover, trying to narrow this gap by increasing the number of disk spindles has a major impact on the energy consumption, space usage, and cost. Given that major improvements in HDD latency are not expected, research and development efforts have shifted towards semiconductor memory technologies that not only complement the existing memory and storage hierarchy but also reduce the distinction between memory (fast, expensive, volatile) and storage (slow, inexpensive, nonvolatile).

In the past, Flash memory technology has been driven by the consumer market alone. The current cost per GB and latency characteristics of NAND Flash make it an interesting candidate to bridge the widening performance gap in enterprise storage systems. In particular, the introduction of multi-level cells (MLC) further decreased the cost as compared to single-level cells (SLC), albeit at the expense of reliability and performance. The absence of moving parts in NAND Flash enhances the ruggedness and latency, and reduces the power consumption, but its operation and structure pose specific issues in terms of reliability and management overhead, which can be addressed in the hardware or software of a Flash controller. Although NAND Flash has already established itself in the memory and storage hierarchy between DRAM and HDDs, it still cannot serve as a universal memory/storage technology. In other words, it does not fully qualify as a storage-class memory (SCM) technology [1]. There are various technologies that could eventually qualify as SCM which combine the performance benefits of solid-state memories with the low cost and the large permanent storage capabilities of magnetic storage. These include ferroelectric, magnetic, phase-change, and resistive random-access memories, including perovskites and solid electrolytes, as well as organic and polymeric memories [2]. It is widely believed that the prime candidate of all these technologies to become SCM is phase-change memory (PCM).

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

On the other hand, the volume of digital data being produced is growing at an ever increasing pace. According to a recent International Data Corporation (IDC) study, 800 exabytes of data were created in 2009 [3]. In the future, this already staggering volume of data is projected to increase at an annual growth rate of more than 60%, faster than the expected growth of the total storage capacity worldwide. Moreover, new regulatory requirements imply that a larger fraction of this data will have to be preserved for extended periods of time. All of this translates into a growing need for cost-effective digital archives. Despite the significant progress of HDD technology over the past years, magnetic tape still remains the least expensive long-term archiving medium. In a recent technology demonstration by IBM in collaboration with Fujifilm, an areal density of 29.5 Gb/in^2 stored on a new BaFe medium was shown, which translates into a potential tape cartridge capacity of 35 terabytes [4]. Moreover, an analysis of the limits of current tape technology suggests that the areal density of tape can be pushed even further towards 100 Gb/in^2, leading to cartridge capacities in excess of 100 terabytes [5]. This clearly indicates that tape remains a very attractive technology for data archiving, with a sustainable roadmap for the next ten to twenty years, well beyond the anticipated scaling limits of current HDD technology. Moreover, the new long-term file system (LTFS) that has been introduced in the LTO (Linear-Tape-Open) generation-5 tape-drive systems allows efficient access to tape using standard and familiar system tools and interfaces. LTFS allows the tape cartridge to be self-describing and self-contained, making tape look and work like other storage solutions.

In this paper, we will first briefly look at today's memory storage requirements from a workload perspective. Next we will discuss HDD technology trends and various solid-state memory technologies that potentially could serve as SCM. Then, we will focus on PCM, the prime contender for SCM, review its current status and technological challenges and also discuss its impact on the system memory and storage hierarchy. We will also consider the other extreme: tape storage. Specifically, we will provide a short overview of the role of tape in the storage hierarchy and where we see the main advantages of tape. We will then briefly refer to the recent world record areal density study of 29.5 Gb/in^2, which indicates that tape has a viable roadmap for the foreseeable future. We will also discuss the recent introduction of LTFS into tape systems and its wide implications for the growing market of archival applications. Finally, we will close by discussing the implications of these two technology extremes, namely, SCM and tape, on the tiered storage hierarchy.

2 Optimized Workloads

There is a fundamental trend towards designing entire systems such that they are optimized for particular workloads, departing from the traditional general-purpose architecture. The typical system, with standard CPUs consisting of a small number of identical cores with a common set of accelerators and relying on a memory and storage hierarchy mainly composed of DRAM and HDDs, has reached its limits in terms of delivering competitive performance improvements for an increasingly diverse set of workloads: future systems will be built out of increasingly heterogeneous components.

From a CPU-level perspective, technology scaling will allow 50 billion transistors to be put on a single chip in approximately five CPU generations, whereas the chip area and total power consumed will remain similar to current levels, namely, $\sim$500 mm^2 and $\sim$200 W. By exploiting the plethora of transistors available, we expect an increasing heterogeneity with dedicated fixed-functions cores (e.g., decryption, XML parsing, pattern matching), programmable cores with domain-tailored instruction sets, or even reconfigurable logic (FPGA-like).

The memory subsystem is becoming one of the most expensive parts of the system, with virtualization being one of the key factors fuelling the push towards larger amounts of memory. Such subsystems will likely be a mix of fast, expensive memory (like DRAM) and a slower, more cost-efficient memory, such as PCM. Mixed memory systems could be built in which the DRAM serves as hardware cache for the PCM memory or DRAM and PCM could both form a contiguous memory environment in which the OS or memory controller determines the optimal placement of data.

From a storage-level perspective, not all data is needed equally often or quickly. The coarse distinction between on-line, near-line, and off-line data naturally maps to heterogeneous storage tiers using the most appropriate technology in terms of access performance and cost. Automated tiering methods already offer the capability to relocate data between tiers as access patterns evolve. Solid-state nonvolatile technologies such as Flash have started to displace HDD usage both in on-line high-performance storage and in mobile computers. HDDs together with the data deduplication technologies are poised to increasingly target near-line applications, such as backup and restore, "encroaching" on the traditional stronghold of tape. On the other hand, with the introduction of LTFS, tape appears to be well positioned to offer complete system-level archive solutions. It is deemed that the low energy consumption and volumetric efficiency together with a file-level interface represent a huge growth opportunity for tape in archival as well as off-line applications.

In summary, the drive for increased performance, coupled with improved cost and power efficiency, is already leading to a greater exploitation of the heterogeneity at the CPU, memory and storage levels. We expect that further performance increases will be delivered by fully application-optimized hardware.

Let us now look at key applications and their workloads, and focus on streaming analytics as well as active archiving. Streaming analytics refers to applications that process large volumes of mainly unstructured streaming data to provide critical information rapidly. This may apply to continuous portfolio risk evaluation, real-time analysis of call data-records, etc. (for examples, see [6]). Such applications require a system designed to sustain the continuous processing of ephemeral data, with systems built out of special pre-processing frontend engines that take disparate data streams at wire rates as input, filter and analyze these data streams very efficiently, and provide the highest-value outputs to a back-end computational layer. This will be achieved by a combination of heterogeneous and dedicated processing engines complemented by a customized memory and storage hierarchy that can accommodate the throughput of such data streams while buffering a sufficient amount for processing purposes in a cost-effective fashion.

At the other extreme of the application spectrum, active archiving addresses the need to efficiently store and access all the data created by an organization. The objective is to continue to be able to leverage information even after it has been stored for good. This may apply to reprocessing or searching data that was expensive to collect in the light of new algorithms or when searching for new artifacts, such as in the case of oil-field inspections or clinical studies, or data that cannot be recreated, such as bank transactions kept for compliance reasons.

An active archive application makes all data always available without taking up expensive primary disk capacity, but instead spreading such data across multiple tiers of storage in a transparent fashion to the user (except for the variation in access performance) [7]. We expect that this will be achieved by using a mix of cache and storage technologies composed of PCM, Flash, HDDs, and tape.

3 HDD Technology Trends

The cost of storing data using HDDs is primarily governed by the achievable areal recording density. Over the past 30 years, HDD capacity and areal densities have increased at compound annual growth rates (CAGR) of about 60%, resulting in a similar exponential decrease in cost per byte of storage. In the 1990's HDD areal density grew at a CAGR of up to 100%. However, during the transition from longitudinal to perpendicular recording, this rate slowed to values as low as $\sim$20%, and currently is around 30%.

The spectacular improvement in HDD capacity has been achieved by continually scaling the size of an individual bit cell, concomitant with steadily improving the various building blocks of the drive. Unfortunately, there is a fundamental physical limitation to how far this scaling can be continued because of thermal stability effects in the storage medium, known as the super paramagnetic limit, that occurs when the magnetic energy stored in a bit becomes comparable to the thermal energy in the environment.

Currently there are two proposed technologies to extend magnetic recording beyond the super-paramagnetic limit: bit-patterned media and heat-assisted magnetic recording (HAMR). The idea of bit-patterned media is to pattern the media into magnetically isolated, single domain islands, such that each bit consists of a single "grain", which can be significantly larger than the 8 nm grain size in use today. Implementing this idea requires the development of a low-cost, large-area nano-scale patterning technique, such as nano-imprint lithography, and also necessitates changes in the way track-follow servo and timing are performed in the drive. The idea of HAMR is to use a medium with a large magnetic anisotropy and to locally reduce the magnitude of the magnetic field required to write to the medium by applying a very short and localized heat pulse with a laser. Challenges associated with this idea include the cost of the laser and the waveguide technology required to deliver the heat pulses, as well as problems associated with keeping the heat localized in the metallic storage medium, which is a good heat conductor.

To date, areal densities of over 800 Gb/in^2 have been demonstrated in the lab using perpendicular recording [8], and at the time of writing, the highest areal density shipping in a product was $\sim$739 Gb/in^2 [9]. The highest capacity HDDs currently available are 1.5 TB in the 2.5$''$ form factor and 3 TB in the 3.5$''$ form factor. The 3-TB drives contain either four 750-GB platters or five 600-GB platters. The cost of a 3-TB drive is around $250, which translates to about 8 cent/GB. However, the total costs of ownership (TCO) of disk-based storage systems are significantly higher than this because of the additional cost of power, cooling, redundancy, etc.

Conventional perpendicular recording is expected to extend to areal densities of up to $\sim$1 Tb/in^2 using exchanged-coupled composite (ECC) media. At the current rate of scaling, this areal density will be reached in about one to two years. To scale beyond this, other technologies, such as patterned media and/or HAMR, will be required.

In stark contrast to the continued exponential improvement in HDD areal densities, other attributes, such as power consumption, data rate and access times, have recently been improving at much more modest rates or in some case not at all. For example, data rate is primarily determined by the linear density and the rotation rate of the drive. Figure 1 shows a plot of the maximum sustained data rate versus time, where it can be seen that since 2000, data rates have increased by a factor of only about 2 to 3. This stagnation is due to a significant slowdown in linear density scaling and the only marginal increase in rotation rates in this period. The latter are limited by the radius of the platters and the mechanical components of the drive. Recently even a trend to decrease rotation rates to reduce power consumption emerges.

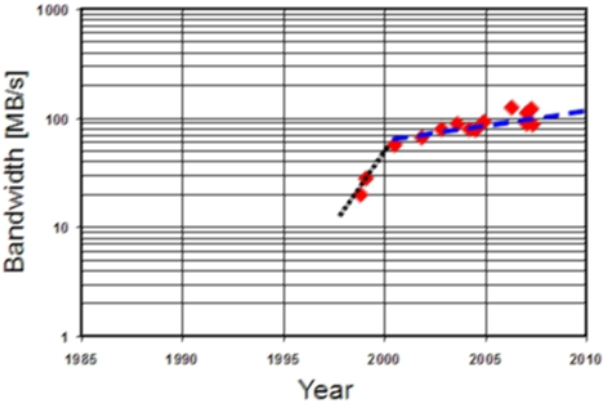

Figure 1: Max sustainable rate. Adapted from [33].

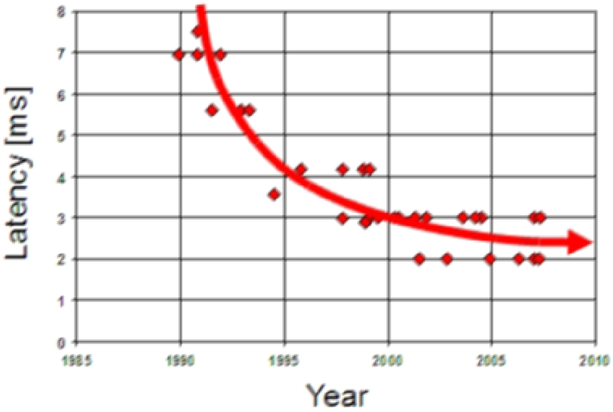

Figure 2: HDD latency. Adapted from [33].

Latency is the time interval between receipt of a read command and the time the data becomes available. It is dominated by the seek time and the disk rotation rate (typically in the range of 5400 -- 15000 rpm). In modern HDDs, access times range from around 15 ms to only a few milliseconds, depending on the class and the power consumption of the drive. Figure 2 shows a plot of latency decrease versus time, where again it can be seen that improvements since the year 2000 have been only marginal. This apparent floor results from practical limitations of the mechanical components used for rotating the disks and positioning the read-write heads.

The number of read/write operations that can be performed per second (IOPS) is given by the inverse of the latency. The stagnation in latency improvements thus means that also IOPS performance gains have stagnated. In fact, because of the continued exponential increase in capacity and the stagnation of latency, the IOPS per GB have actually been decreasing exponentially. The only way to improve the IOPS performance is to increase the number of spindles operated in parallel, which has the obvious disadvantages of a linear increase in system costs and power consumption.

The power consumption of an HDD depends on the operations performed and varies with the type of HDD. Number and size of the platters, rotation rate, interface type, data rate and access time all impact power consumption. For example, in a 2005 study of 3.5″ disks, the power consumption was found to vary from as much as ∼30 W during spin-up to 5 — 10 W during idle, with an average of approx. 10 W for intensive work loads [10]. More recently, HDD manufacturers have begun to produce more “eco-focused” drives in which the power consumption during read/write operations for a 3.5″ drive is on the order of 5 W and can be as low as ∼2 W for a 2.5″ drive. One strategy to reduce power in a large array of disks is to spin down unused disks, a strategy known as MAID (massive array of idle disks). However, this concept has not yet become very successful, most likely because of the large latency penalty and the reliability concerns associated with continually powering disks on and off.

4 Flash Technology and Storage Class Memory (SCM) Trends

Today’s memory and storage hierarchy consists of embedded SRAM on the processor die and DRAM as main memory on one side, and HDDs for high-capacity storage on the other side. Flash memory, in the form of solid-state-disks (SSD), has recently gained a place in between DRAM and HDD, bridging the large gap in latency ($\sim 10^5$ times) and cost (∼100 times) between them. However, the use of Flash in applications with intense data traffic, i.e., as main memory or cache, is still hampered by the large performance gap in terms of latency between DRAM and Flash (still ∼1000 times), and the low endurance of Flash (10^4 – 10^5 cycles), which deteriorates with scaling and more aggressive MLC functionality. MLC technology has become the focus of the Flash vendors because they target the huge consumer-electronics market, where the low cost per gigabyte of MLC plays a very important role, but it suffers from endurance and latency issues, which could be problematic for enterprise-class applications. For example, at the 32-nm technology node and 2 bits per cell, it is expected that the standard consumer-grade MLC will offer a write/erase endurance of approx. 3,000 cycles, which clearly will not suffice for enterprise-class storage applications. On the other hand, an enterprise-grade MLC with higher cost per gigabyte could offer a write/erase endurance of 10^4 to 3×10^4, albeit with a slower programming latency of approx. 1.6 ms. These limitations of the MLC technology necessitate the use of more complex error-correction coding (ECC) schemes and Flash management functions, which, depending on the workload, could improve the reliability and hide the latency issues to a certain extent — but certainly not to full satisfaction.

Moreover, as we go down on the technology node, these issues will be further aggravated, and new challenges will have to be resolved. For example, the stringent data-retention requirements, in particular for enterprise-storage systems, impose a practical limit for the thickness of the tunnel oxide. Another challenge in the scaling of floating-gate NAND is floating-gate interference. To resolve this issue, a charge-trapping layer has been proposed as an alternative technology to the floating gate [2]. In general, it was believed for a long time that by moving to charge-trapping storage it would be possible to scale at least to the 22-nm lithography generation. However, recently a very promising trend towards stacking memory cells in three dimensions in what is called 3D memory technology has emerged, and leading NAND Flash memory manufacturers are already pursuing it [11]. Of course, this 3D memory technology will not truly have an impact on reliability, endurance and latency, but it will offer much larger capacities at even lower cost in the future. For all these reasons, NAND Flash is not expected to become an SCM technology in general.

Scaling issues are also critical for other solid-state memories, such as SRAM and DRAM. Specifically, SRAM suffers from signal-to-noise-ratio degradation and 10x leakage increase with every technology node, and DRAM faces a continuous increase of the refresh current.

Hence, there is a large opportunity for new solid-state nonvolatile memory technologies with "universal memory" characteristics. These technologies should not only extend the lifetime of existing memories, but also revolutionize the entire memory-storage hierarchy by bridging the gap between memory (fast, expensive, volatile) and storage (slow, inexpensive, permanent). The requirements of this new family of technologies called SCM [1] are nonvolatility, solid-state implementation (no moving parts), low write/read latency (tens to hundreds of nanoseconds), high endurance (more than 10^8 cycles), low cost per bit (i.e., between the cost per bit of DRAM and Flash), and scalability to future technology nodes.

Many new nonvolatile solid-state memory technologies have recently emerged. The objective has not only been to realize dense memory arrays and show a viable scalability roadmap, but also to achieve a performance superior to that of Flash memory in many aspects. The catalog of new technologies is very long, and they may be broadly categorized into charge-trap-based, capacitance-based and resistance-based memories. Charge-trap based memories are basically extensions of the current floating-gate-based Flash and, while offering advantages in reliability, suffer from the same drawbacks that afflict Flash technology, namely, low endurance and slow writing speeds. Capacitance-based memories, in particular ferroelectric memories (FeRAM), exhibit practically infinite cycling endurance and very fast read/write speeds, but are hampered by short retention times and, even more importantly, their limited scaling potential up to the 22-nm node [12].

Resistance-based memories encompass a very broad range of materials, switching mechanisms and associated devices. Following the International Technology Roadmap for Semiconductors (ITRS), one may categorize resistance-based memories into nanomechanical, spin torque transfer, nanothermal, nanoionic, electronic effects, macromolecular and molecular memories [12]. Of these technologies, those that have received more attention by the scientific community and the semiconductor industry and are thus in a more advanced state of research and/or development, are spin torque transfer, nanoionic and thermal memories. We will take a closer look at these technologies next.

Spin-torque transfer memory (STTRAM) [13]–[15] is an advanced version of the magnetic random access memory (MRAM) in which the switching mechanism is based on the magnetization change of a ferromagnetic layer induced by a spin-polarized current flowing through it. The most appealing features of STTRAM are its very high read/write speed, on the order of 10 ns or less, and its practically unlimited endurance. Important challenges are overcoming the small resistance range (between low and high resistance), which limits the possibility of MLC storage, and achieving adequate margins not only between read and write voltages but also between write and breakdown voltages for reliable operation, especially at high speeds.

Nanoionic memories [16]–[21] are characterized by a metal-insulator-metal (MIM) structure, in which the "metal" typically is a good electrical conductor (possibly even dissimilar on the two sides of the device) and the "insulator" consists of an ion-conducting material. Typical insulator materials reported so far include binary or ternary oxides, chalcogenides, metal sulfides, and even organic compounds. The basic switching mechanism in nanoionic memories is believed to be the combination of ionic transport and electrochemical redox reactions [19]. Most of these technologies—with very few exceptions—are still in a very early stage of research, with many of their interesting features derived from projections or extrapolations from limited experimental data. As is generally known, the actual issues associated with a particular technology will likely only manifest themselves in large demonstration device test vehicles, so that it may well be that the road to widespread adoption of nanoionic memories is still a long one.

The best known thermal memory is phase-change memory (PCM). This discussion focuses on PCM, mainly because of the very large activity around it and the advanced state of development it has reached, allowing credible projections regarding its ultimate potential. The active material in PCM is a chalcogenide, typically involving at least two of the materials Ge, Sb and Te, with the most common compound being $Ge_2Sb_2Te_5$

or, simply, GST. The active material is placed between two electrically conducting electrodes. The resistance switching is induced by the current flowing through the active material, which causes a structural change of the material due to Joule heating. Phase-change materials exhibit two meta-stable states, namely, a (poly)-crystalline phase of long-range order and high electrical conductivity and an amorphous phase of short-range order and low electrical conductivity. Switching to the amorphous phase (the RESET transition) is accomplished by heating the material above its melting temperature followed by ultra-fast quenching, whereas the crystalline phase (SET transition) is reached by heating the materials above its crystallization temperature and subsequent annealing. The RESET transition necessitates high current, but this current has been shown to scale linearly with the technology node as well as decrease significantly in confined memory cell architectures [22]. The RESET transition is fast, typically less than 50 ns in duration, whereas the SET transition is on the order of 100 ns, although very fast materials exhibiting sub-20-ns switching times have been reported [23].

PCM scores well in terms of most of the desirable attributes of a SCM technology. In particular, it exhibits very good endurance, typically exceeding 10^8 cycles, excellent retention, and superb scalability to sub-20-nm nodes and beyond. Most importantly, these characteristic numbers have been measured on large prototype devices and thus provide confidence regarding the true performance of the memory technology. On a smaller-scale device level, PCM has been shown to possess all the necessary characteristics of a SCM technology. Specifically, sub-20-ns SET switching times have been reported with doped SbTe materials [23]. Furthermore, an impressive device has been fabricated at the 17-nm design rule at $4F^2$ size, with further scaling prospects not limited by lithography but only by the material film thickness [24]. The same device showed an extrapolated cycling endurance exceeding 10^{15} cycles. The ultimate scaling limits of phase change in chalcogenide materials provide an indication regarding the future scaling of PCM. In a recent study, GST films that are a mere 2 nm thick have been shown to crystallize when surrounded by proper cladding layers [25].

Apart from the necessary RESET current reduction and SET speed improvement discussed above, a significant challenge of PCM technology is a phenomenon known as (short-term) resistance drift: The resistance of a cell is observed to drift upwards in time, with the amorphous and partially-amorphous states drifting more than their crystalline counterparts. This drift is believed to be of electronic nature, manifests itself as noise, and seriously affects the reliability of MLC storage in PCM because of the reduced sensing margin between adjacent tightly-packed resistance levels. Therefore, effective solutions of the drift issue are a key factor of the cost competitiveness of PCM technology and thus of its suitability as SCM.

In summary, PCM is the only one of numerous emerging memory technologies that has evolved from the basic research stage to the advanced development and late prototyping stage without encountering any fundamental roadblocks. Advanced prototype PCM chips that at least partially meet the requirements for SCM already exist today [26, 27], and new and exciting device demonstrations have shown tremendous potential for further improvement. These developments render PCM the leading technology candidate for SCM today, with the potential to play an extended role in the memory and storage hierarchy of future computing systems.

5 Tape Technology Trends

Tape systems constitute an integral part of current tiered storage infrastructures. They are especially suited for low-cost, long-term storage of data as well as for backup and disaster recovery purposes. Tape technology offers several important advantages, including energy savings, security, lifetime, reliability and cost. Moreover, in such applications, the main drawback of tape, its slow access time, does not have a major impact on the system., Once data has been recorded in tape systems, the medium is passive; it simply sits in a rack and no power is needed. Compared with similar disk-based systems, a tape-based archive consumes approximately 290 times less power [28]. In terms of security, once data has been recorded and the cartridge removed from the access system, the data is inaccessible until the cartridge is reinstalled in the active system. Security is further enhanced by drive-level encryption, which was introduced in Linear Tape Open generation-4 drives (LTO-4) and

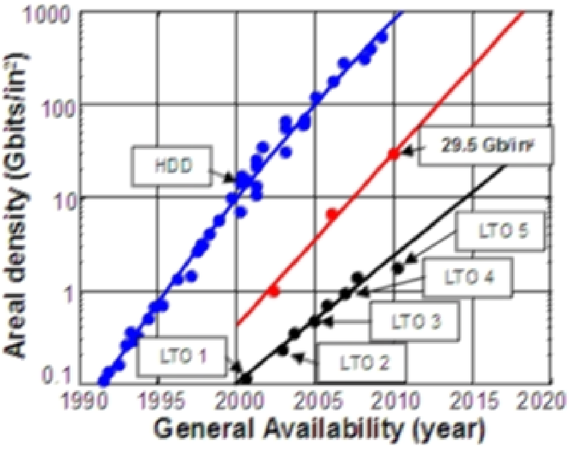

Figure 3: Figure 3. Recording of areal density of HDD and tape products and tape demonstrations. Adapted from [5].

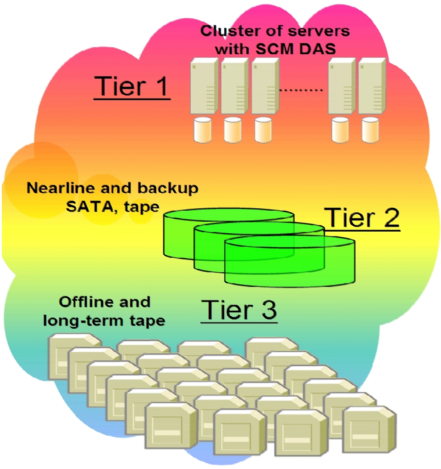

Figure 4: Tiered storage.

is also standard in enterprise-level tape drives. The tape medium has a lifetime of 30+ years; however, this is rarely taken advantage of because of the rapid advances in tape hardware and the cost savings associated with migration to higher-capacity cartridges. In terms of reliability, LTO-4 tape has a bit error rate that is at least an order of magnitude better than that of a SAS HDD [29]. Moreover, the fact that tape media is removable and interchangeable, means that, in contrast to HDDs, mechanical failure of a tape drive does not lead to data loss because a cartridge can simply be mounted in another drive. All of the above advantages contribute to the major net advantages of tape, which are cost and reliability. Estimates of the savings of the total cost of ownership of a tape backup system, relative to HDD backup, range from a factor of three to more than 20, even if developments such as data deduplication are taken into account. In archival applications, where no effective use of deduplication is possible, cost savings can be even higher.

Figure 3 compares the evolution of the areal densities of HDD and tape products over time, including recent tape areal density demonstrations. The plot indicates that even though the gap between the areal densities of HDD and tape products has remained essentially constant in recent years, tape areal density demonstrations exhibit a slope of about 60% CAGR, indicating the potential of reducing the areal density gap between tape and HDD. Insight into how this was achieved can be gained by comparing the bit aspect ratios (BAR) of tape drives and HDDs. The typical BAR (bit width to bit length) in recent HDD technology demonstrations is on the order of 6:1 [30] and, as mentioned in Section 3, currently there are HDD products on the market that operate at linear densities of more than 700 Gb/in^2. In contrast, the latest LTO-5 tape drives operate at an areal density of 1.2 Gb/in^2 with a BAR of about 123:1 [31]. This comparison indicates that there is considerable room to reduce the track width in tape systems. The potential to continue to scale tape areal densities was recently experimentally confirmed in a world record tape areal density recording demonstration of 29.5 Gb/in^2, performed jointly by IBM and Fujifilm [4]. What is clear from this density demonstration is that there is ample room to continue scaling the areal density and cartridge capacity of tape systems for at least the next 10 years. Moreover, a recent study shows that further improvements in technology may increase this density to 100 Gb/in^2, indicating the potential to continue scaling tape for many years to come without a fundamental change of the tape-recording paradigm [5]. Finally, considering that the areal density of HDDs is currently scaling at a CAGR of $\sim$30% and that tape has the potential to continue scaling at 40% or more, the cost advantages of tape over HDD may become even greater.

One of the obstacles to more widespread adoption of tape in the past has been the difficulty of using tape in a general or stand-alone context. Hard disks provide random access to data and generally contain a file index managed by a file system. These files can be accessed by means of standard sets of application programming interfaces (APIs) using various operating systems and applications. Tape, in contrast, is written in a linear sequential fashion typically using a technique called "shingling" which provides backward write compatibility, but also implies that new data can only be appended and that previously written areas can only be reclaimed if the entire cartridge is reclaimed and rewritten. In traditional tape systems, an index of the files written on a given cartridge is usually only kept in an external database managed by an application such as a proprietary back-up application. The need to access an external database to retrieve data renders data on tape much less portable and accessible than with alternative storage methods, such as a HDD or a USB drive.

To address these deficiencies, a new long-term file system (LTFS) has recently been introduced into the LTO-5 tape-drive systems to enable efficient access to tape using standard and familiar system tools and interfaces [32]. LTFS is implemented using the dual-partition capabilities supported in the new LTO-5 format. A so-called index partition is used for writing the index, and the second, much larger partition for the data itself. This new file system makes files and directories show up on the desktop with a directory listing. Users can "drag and drop" files to and from tape and can run applications developed for disk systems. In library mode, the content of all volumes in the library can be listed and searched without mounting the individual cartridges. All these features help reduce tape, file management and archive costs and eliminate the dependency on a middleware layer. Hence the cost per GB stored is reduced. In addition, tape becomes cross-platform-portable (Linux*, Mac*, Windows*), enabling and facilitating the sharing of data between platforms. These features enable significant new use cases for tape, such as video archives, medical images, etc.

Considering the cost advantages of tape over other storage solutions, the demonstrated potential for the continued scaling of tape-cartridge capacity and cost per GB as well as the increasing usability of tape provided by advances such as the LTFS, tape appears set to play an important role in the exploding market for archival data storage solutions.

6 Implications of SCM and Tape on Tiered Storage Hierarchy – Conclusion

Figure 4 shows how SCM and tape will affect the tiered storage hierarchy. The advent of SCM as well as the continuous density scaling of the low-cost tape technology will impact the storage tiering hierarchy, fitting the workloads envisaged in future applications. Specifically, it is expected that, for the on-line tier, the direct-attached storage (DAS) model will resurface in the form of server clusters with low-latency SCM storage in each server. This will provide the high performance required by IO-limited workloads by leveraging the huge internal bandwidth necessary for streaming analytics, for instance. The co-location of computation workload and corresponding data sets will in turn reduce the need for caching. The current benefits in provisioning and utilization of virtualized SAN-attached storage will need to be leveraged in the new DAS model as well.

For the near-line tier, both low-cost SATA disks and tape will compete to support backup and active archive applications, whereas for the off-line tier, which primarily targets archiving applications, power-efficient high-capacity, low-cost tape will be the greenest technology and the ultimate insurance policy for the enterprise.

*Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both. Microsoft is a trademark of Microsoft Corporation in the United States, other countries, or both. IBM is a registered trademark of International Business Machines Corporation in the United States, other countries, or both. Other product and service names might be trademarks of IBM or other companies.

References

[1] R. F. Freitas, W. W. Wilcke. Storage-Class Memory: The Next Storage System Technology. *IBM J. Res. Develop.* 52(4/5), 439–447 (2008).

[2] G. W. Burr *et al.* Overview of Candidate Device Technologies for Storage-Class Memory. *IBM J. Res. Develop.* 52(4/5), 449–464 (2008).

[3] 2010 Digital Universe Study: A Digital Universe Decade Are You Ready?
http://gigaom.files.wordpress.com/2010/05/2010-digital-universe-iview_5-4-10.pdf

[4] G. Cherubini *et al.* 29.5 Gb/in^2 Recording Areal Density on Barium Ferrite Tape. *IEEE Trans. Magn.* 47(1) (January 2011, in press) DOI: 10.1109/TMAG.2010.2076797; also in *Digest of Magnetic Recording Conf. "TMRC,"* San Diego, CA, pp. 29–30 (August 2010).

[5] A. J. Argumedo *et al.* Scaling Tape-Recording Areal Densities to 100 Gb/in^2. *IBM J. Res. Develop.* 52(4/5), 513-527 (2008).

[6] http://www.netezza.com/documents/whitepapers/streaming analytic white paper.pdf

[7] Active Archive Alliance whitepapers: http://www.activearchive.com/common/pdf/TimeValue WP US 4229.pdf
http://www.activearchive.com/common/pdf/ActiveArchiveWPInterSect360201004.pdf

[8] Naoki Asakawa. TDK Achieves World's Highest Surface Recording Density of HDD. *Nikkei Electronics*, Sept. 30, 2008, http://techon.nikkeibp.co.jp/english/NEWS EN/20080930/158806/

[9] http://www.conceivablytech.com/2106/products/hard-drives-get-new-record-density-where-is-the-limit/

[10] HDD Diet: Power Consumption and Heat Dissipation. http://ixbtlabs.com/articles2/storage/hddpower.html

[11] M. Kimura. 3D Cells Make Terabit NAND Flash Possible. *Nikkei Electronics Asia*, Sept. 17, 2009.

[12] *International Technology Roadmap for Semiconductors, Emerging Research Devices*, 2009 edition.

[13] J-G. Zhu. Magnetoresistive Random Access Memory: The Path to Competitiveness and Scalability. *Proc. IEEE* 96, 1786–1797 (2008).

[14] T. Kishi *et al.* Lower-Current and Fast Switching of a Perpendicular TMR for High Speed and High Density Spin-Transfer-Torque MRAM. In *Proc. IEDM 2008*, pp. 1–4 (2008).

[15] M. Hosomi *et al.* A Novel Nonvolatile Memory with Spin Torque Transfer Magnetization Switching: Spin RAM. In *2005 IEDM Technical Digest*, p. 459 (2005).

[16] A. Asamitsu, Y. Tomioka, H. Kuwahara, Y. Tokura. Current-Switching of Resistive States in Colossal Magnetoresistive Oxides. *Nature* 388, 50–52 (1997).

[17] M. N. Kozicki, M. Yun, L. Hilt, A. Singh. Applications of Programmable Resistance Changes in Metal-Doped Chalconides. *Electrochem. Soc. Proc.* 99-13, 298-309 (1999).

[18] A. Beck, J. G. Bednorz, C. Gerber, C. Rossel, D. Widmer. Reproducible Switching Effect in Thin Oxide Films for Memory Applications. *Appl. Phys. Lett.* 77(1), 139–141 (2000).

[19] R. Waser, M. Aono. Nanoionics-based Resistive Switching Memories. *Nature Materials* 6, 833–840 (2007).

[20] R. Meyer, et al. Oxide Dual-Layer Memory Element for Scalable Non-Volatile Cross-Point Memory Technology. In *Proc. NVMTS 2008*, pp. 54-58 (2008).

[21] J. Borghetti *et al.* "Memristive" Switches Enable 'Stateful' Logic Operations via Material Implication. *Nature* 464(8), 873–876 (2010).

[22] D. H. Im *et al.* A Unified 7.5nm Dash-Type Confined Cell for High Performance PRAM Device. In *Proc. Int'l. Electron Devices Meeting (IEDM) 2008*, San Francisco, CA, pp. 211–214 (2008).

[23] B. Cheong *et al.* Characteristics of Phase Change Memory Devices Based on Ge-doped SbTe and its Derivative. In *Proc. European Phase Change and Ovonics Symposium (E*PCOS)* (2007).

[24] I. S. Kim. High Performance PRAM Cell Scalable to Sub-20nm Technology with below 4F^2 Cell Size, Extendable to DRAM Applications. In *2010 Symp. on VLSI Technology Digest of Technical Papers*, pp. 203–204 (2010).

[25] R. E. Simpson *et al.* Toward the Ultimate Limit of Phase Change in Ge$_2$Sb$_2$Te$_5$. *Nano Lett.* 10(2), 414–419 (2010).

[26] F. Bedeschi *et al.* A Bipolar-Selected Phase Change Memory Featuring Multi-Level Cell Storage. *IEEE J. Solid-State Circuits* 44(1), 217–227 (2009).

[27] K.-J. Lee *et al.* A 90 nm 1.8 V 512 Mb Diode-Switch PRAM with 266 MB/s Read Throughput. *IEEE J. Solid-State Circuits* 43(1), 150–161 (2008).

[28] D. Reine, M. Kahn. Disk and Tape Square Off Again — Tape Remains King of the Hill with LTO-4. *Clipper Notes* (2008). www.clipper.com/research/TCG2008009.pdf

[29] H. Newman. The Tape Advantage: Benefits of Tape over Disk in Storage Applications. White paper, Instrumental (2008).

[30] Z. Bandic, R. H. Victora. Advances in Magnetic Data Storage Technologies. *Proc. IEEE* 96(11), 1749–1753 (2008).

[31] https://www.bluestoragemedia.com/External/BigBlueBytes/LTO5

[32] D. Pease *et al.* The Linear Tape File System. In *Proc. 2010 IEEE 26th Symp. on Mass Storage Systems and Technologies (MSST)*, Incline Village, NV, pp. 1–8 (2010).

[33] R. Freitas, W. Wilcke, B. Kurdi. Storage Class Memory Technology and Use. *Proc. FAST '08*, 26–29, (2008).