Contents

1 Storage Technology **3**
Jason Hick, John Shalf
1.1 Introduction . . . 3
1.2 Fundamentals of Magnetic Storage . . . 4
1.2.1 Storage Medium . . . 5
1.2.2 Superparamagnetism . . . 6
1.2.3 Magnetoresistive Reading . . . 6
1.2.4 Giant Magnetoresistance (GMR) . . . 6
1.2.5 Perpendicular Recording . . . 7
1.2.6 Future Trends . . . 7
1.3 Disk Storage . . . 8
1.3.1 Fundamentals . . . 8
1.3.2 Performance Characteristics . . . 9
1.3.3 Enterprise and Commodity Disk Technology . . . 10
1.3.4 Future Trends . . . 11
1.4 Tape Storage . . . 11
1.4.1 Fundamentals . . . 12
1.4.2 Enterprise Class Tape Technology . . . 13
1.4.3 Commodity Tape Technology . . . 13
1.4.4 Future Trends . . . 14
1.5 Optical Storage . . . 14
1.5.1 CDs . . . 14
1.5.2 DVDs . . . 15
1.5.3 Blu-Ray and HD-DVD . . . 15
1.5.4 Future Trends . . . 15
1.6 Composite Devices . . . 16
1.6.1 RAID . . . 16
1.6.2 Virtual Tape Libraries(VTLs) . . . 17
1.6.3 Redundant Arrays of Independent Tape (RAIT) . . . 18
1.6.4 Data Integrity . . . 18
1.7 Emerging Technologies . . . 19
1.7.1 Massive Array of Idle Disks (MAID) . . . 19
1.7.2 FLASH . . . 19
1.7.3 MRAM . . . 21
1.7.4 Phase-Change Technology . . . 21
1.7.5 Holographic Storage . . . 21
1.7.6 Direct Molecular Manipulation . . . 21
1.8 Summary and Conclusions . . . 22

0-8493-0052-5/00/$0.00+$.50

References **29**

Chapter 1

Storage Technology

Jason Hick, John Shalf

National Energy Research Supercomputing Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

1.1 Introduction 3
1.2 Fundamentals of Magnetic Storage 4
1.3 Disk Storage 8
1.4 Tape Storage 11
1.5 Optical Storage 14
1.6 Composite Devices 15
1.7 Emerging Technologies 19
1.8 Summary and Conclusions 22
Acknowledgments 24

Abstract Description of this section

1.1 Introduction

Computational science is at the dawn of petascale computing capability, with the potential to achieve simulation scale and numerical fidelity at hitherto unattainable levels. However, harnessing such extreme computing power will require an unprecedented degree of parallelism both within the scientific applications and at all levels of the underlying architectural platforms. Power dissipation concerns are also driving High Performance Computing (HPC) system architectures from the historical trend of geometrically increasing clock rates towards geometrically increasing core counts (multicore) [1], leading to daunting levels of concurrency for future petascale systems. Employing an even larger number of simpler processor cores operating at a lower clock frequency, is increasingly common as we march toward petaflop-class HPC platforms, but it puts extraordinary stress on the I/O subsystem implementation.

I/O systems for scientific computing have unique demands that are not seen in other computing environments. These include the need for very high bandwidth and large capacity, which requires thousands of devices working in parallel, and the commensurate fault resilience required to operate a storage system that contains so many components. Chapter 2 will examine the software technology required to aggregate these storage building blocks into reliable, high-performance large-scale filesystems. This chapter will focus on the characteristics of the fundamental building blocks used to construct parallel filesystems to meet the needs of scientific applications at scale.

0-8493-0052-5/00/$0.00+$.50

Understanding the unprecedented requirements of these new computing paradigms, in the context of high-end HPC I/O subsystems, is a key step towards making effective petascale computing a reality. The main contribution of this chapter is to quantify these tradeoffs in cost, performance, reliability, power, and density of storage systems by examining the characteristics of the underlying building blocks for high end I/O technology. This chapter projects the evolution of the technology as constrained by historical trends to understand the effectiveness of various implementations with respect to absolute performance and scalability across a broad range of key scientific domains.

1.2 Fundamentals of Magnetic Storage

Magnetic recording technology was first demonstrated in 1898. Magnetic storage has been closely associated with digital computing since the dawn of the industry. Its performance and cost-effectiveness ensure that it will continue to play an important role in non-volatile storage for some time to come. The primary factor that consistently drives down the cost of storage media is the ability to double storage density consistently on an annual basis as shown in the storage trends in figure 1.3. As we reach physical limits of the media, the storage density trends can no longer continue and the market may change dramatically. Until-then, magnetic storage will continue to play a leading role in non-volatile storage technology. This section elucidates the physics underlying the magnetic storage technology and the current challenges to improving storage density.

The underlying media for magnetic recording are ferromagnetic materials. During the write operation, a magnetic field is applied to the recording medium, that aligns the magnetic domains to the orientation of the field. The write head uses an electrical coil to induce the magnetic field in the head to magnetize regions of the medium, and the read-head senses the polarity of magnetization in the underlying media using either magnetic induction, or the magnetoresistive effect. Figure 1.1(a) shows how the current is applied to the write head to record bits onto the medium and how the read head subsequently recovers the data when the magnetically polarized media induces electrical impulses into the read head as it passes over the magnetically polarized media. The green lines shown under the read head in figure 1.1(a) show the magnetic flux lines. The current is induced in the read head when the flux lines change direction, thereby making the boundaries between bits detectable.

Figure 1.1(b) shows the hysteresis curve of typical magnetic media, which enables the magnetic flux of the write head (H) to induce a reversible change in the magnetization (M) of the underlying medium. The coercivity of the media refers to the amount of magnetic flux that must be applied to the media to change its magnetization. Higher coercivity media improves the stability and density of the storage (e.g. reducing the instabilities caused by superparamagnetism). The magnetic recording industry employs a number of technologies to improve the recording density of the media. These include improvements to the coercivity of ferromagnetic materials that comprise the media, improved read-head technology (magnetoresistive recording), and improvements to the recording technology (perpendicular recording). In the next subsections, we will go through a subset of these improvements, describe their inherent physical limitations in terms of bit density, and discuss future approaches to maintaining historical trends in storage density, which will be discussed in figure 1.3.

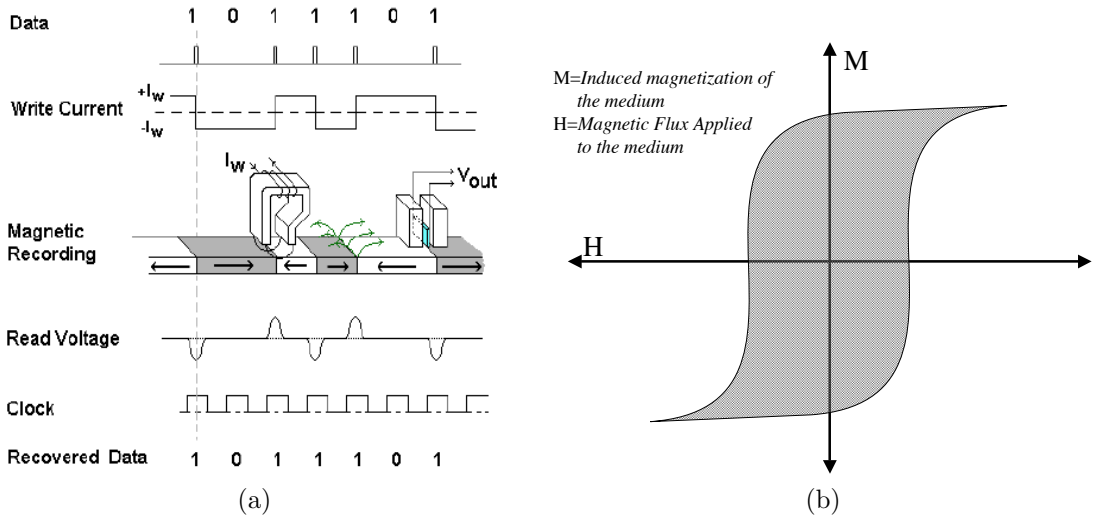


FIGURE 1.1: The left-hand figure (a) shows how data is recorded onto the ferromagnetic medium and how the read head subsequently recovers the information. The green lines shown on the figure are magnetic flux lines. It is the reversal in the magnetic flux lines that the read-head detects to determine bit boundaries. The right side (b) shows the typical hysteresis curve of typical magnetic media.

1.2.1 Storage Medium

The performance of this technology is strongly dependent on the material properties of the recording medium. The medium is composed of ferromagnetic particles suspended in a binder matrix (typically a polymer or other flexible plastic material). The ferromagnetic particles contain domains that align to the magnetic field if sufficient flux is applied. Coercivity is a measure of a material's resistance to magnetic reversal. As mentioned above, materials with a high coercivity also are better able to hold their magnetization and also tend to maintain a higher magnetic flux, which improves the signal-to-noise ratio (SNR) when reading their magnetization. The first digital storage devices employed iron oxide particles, but chromium dioxide particles, and thin-film alloy materials have substantially improved the coercivity of the medium. More recently, thin-film layering of antiferromagnetically coupled (AFM) layers have enabled order-of-magnitude improvements in coercivity over the best available homogeneous thin-films.

Improvements in storage density are achieved by reducing the thickness of the magnetic medium to maintain a high aspect ratio of the recorded bits, but this is done at the expense of the signal-to-noise ratio of the medium. Even if the lowered SNR can be compensated by using more sensitive read head technology, the superparamagnetic effect (explained below) sets a lower limit on bit size for conventional longitudinal recording as bit stability becomes increasingly subject to changes due to random temperature fluctuation.

1.2.2 Superparamagnetism

Magnetic media are composed of discrete crystaline grains of ferromagnetic material embedded in some form of binder matrix (polymer or similar materials). The magnetic bit value is determined by the magnetic polarity of a population of about 1000 of these magnetic "grains" in order to maintain a reasonable signal-to-noise ratio. Continued improvements in areal density require proportional decreases in magnetic grain size for the underlying magnetic material. However, as these grains get smaller, they are more susceptible to random changes in their magnetic polarity due to random thermal fluctuations in the medium, in a behavior known as the superparamagnetic effect [2]. Improvements in the coercivity of storage media has kept the superparamagnetic effect at bay since the mid 1990's, but many believe these improvements will reach their limit at 150-200 Gb/mm^2, forcing a move towards alternative approaches to recording, including magneto-optical, heat-assisted recording to reduce coercivity for writes, and perpendicular recording technology.

1.2.3 Magnetoresistive Reading

Another aspect that can limit the density of recorded data is the sensitivity, size, and consequent flying-height of the read heads. In early devices, the read operation employed an inductive head that depended on the magnetized medium inducing a current in coiled wire on the read head as the medium moves under the read head. Magnetoresistive materials change their resistance in response to the magnetization of the substrate. Magnetoresistive read-heads offered order of magnitude improvements in sensivity compared to inductive heads, enabling increased recording densities. The performance of a magnetoresistive head is independent of the rate of movement of the underlying medium, which is particularly useful for tape heads where the rate of the medium is very low or subject to velocity changes. The magnetoresistive effect cannot be used to induce magnetization in the medium, so the conventional inductive heads must remain, but magnetoresistive heads are substantially smaller than the typical inductive heads and can often be embedded within the structure of a typical thin-film recording head. The magnetoresistive heads are typically narrower than the recorded track in the medium, which can dramatically reduce the cross-talk between neighboring tracks during the read operation.

1.2.4 Giant Magnetoresistance (GMR)

The "Giant Magneto-Resistive" (GMR) effect was discovered in the 1980s by Peter Gruenberg and Albert Fert, and led to their 2007 nobel prize in physics. In their research on materials that are comprised of alternating layers of magnetic and non-magnetic materials, they observed that sensitivity to magnetization improved between 6% and 50% in comparison to conventional magnetoresistive materials. Further refinement of the GMR technology led to read-heads that were an order of magnitude more sensitive than its conventional magnetoresistive approach. The improvements in read sensitivity enabled further reductions in bit-sizes and enabled continued annual doubling of storage densities from 1997 to approximately 2005.

The GMR effect also gave birth to a new form of magnetism-based computer logic referred to as "spintronics," [4] which refers to the underlying mechanism of the effect, which is spin-dependent scattering of electrons. The GMR effect is also being used for new solid state devices such as MRAM, which we will discuss in the future technologies section.

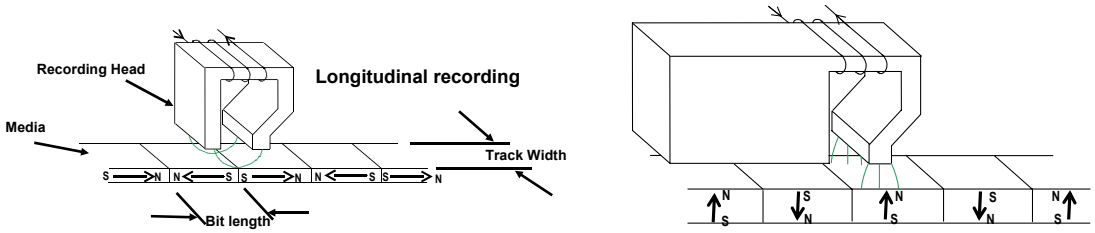


FIGURE 1.2: Longitudinal recording devices (a) currently dominate the market. However, the superparamagnetic effect limits areal densities of this technique to 150Gbits/mm^2. Perpendicular recording (b) promises to support continued scaling of areal density up to 1Tbit/mm^2 because neighboring domains complement rather than counteract each other's polarity. This improves the stability of the recording and thereby overcomes the challenges of superparamagnetism.

1.2.5 Perpendicular Recording

Superparamagnetic effects limit recording density to 150 gigabits/mm^2 for conventional longitudinal recording where the polarization is parallel to the magnetic storage medium. In such an arrangement, the areas between oppositely polarized regions tend to demagnetize their neighbors. As shown in figure 1.2, perpendicular recording technology induces the magnetic field perpendicular to the storage medium where the demagnetizing fields tend to have less interaction between neighboring bits. Perpendicular recording is expected to support continued improvements to storage densities that top out at about 1 Terabit/mm^2. The first commercial products incorporating vertical recording technology were introduced in 2005 with bit densities of 170 Gigabits/mm^2, with density doubling approximately every two years.

1.2.6 Future Trends

Figure 1.3 summarizes trends in storage density and cost per bit. The cost-trends depicted in figure 1.3(b) are strongly dependent on the ability to improve bit densities in figure 1.3(a) at exponential rates. When a storage media reaches the limits imposed by physics, it opens opportunities for alternative technology to take over. For example, magnetic disk storage density was eclipsed by optical storage in the late 1980's, but was able to improve performance at a far faster pace and achieve higher bit-densities than competing optical solutions. However, as we edge closer to the limits of perpendicular recording technology (1 Terabit/mm^2), multiple technology options appear to compete well. At atomic scales, existing approaches to magnetic storage technology, including perpendicular recording, are clearly not viable. However, spintronic devices are able to push past the superparamagnetic limit, and may yet reset the clock again for magnetic technology.

Notice that tape is far below the storage density curve traced by the disk technology in figure 1.3. Helical tape technology will not reach the superparamagnetic limits to areal storage density until nearly 2020 if densities improve at historical rates. The cost per bit of tape storage remains far superior and differentiated from the cost of low-end disk storage in figure 1.3, so the tape systems are unlikely to jump to the kinds of storage densities seen in the disk storage systems because there is currently little economic incentive to do so. It also indicates that if there is any market pressure

on tape storage, the technology has a lot of margin to improve storage density relative to the physical limits imposed by the underlying recording media were it to be challenged by holographic or enhance optical storage media in the future.

In the area of desktop disk storage, the storage trends have followed a very consistent slope on the exponential graph except for the non-volatile solid-state storage, such as FLASH. The consumer electronics market applications for FLASH storage have created such rapid increases in manufacturing volume, that the cost/bit has been decreasing far faster than any of the other technology options. The costs of such devices are now within striking distance of the upper-end of the consumer grade ATA disk systems, and may well be a direct competitor to that market segment in the coming years if this accelerated trend can continue. If FLASH becomes competitive with mechanical disk, it will dramatically change the landscape of storage systems – affecting fundamental design decisions for the parallel filesystems described in Chapter 2. However much work still needs to be done to characterize the failure modes of non-volatile storage solutions to understand how they will behave for large-scale high-bandwidth scientific storage subsystems, before progress can be made in this area.

1.3 Disk Storage

The very first rotating disk storage device was the IBM 350 RAMAC (Random Access Memory Accounting system), which was introduced in 1956. It contained 50 platters spinning at 1200RPMs and had a total storage capacity of 5 megabytes. Whereas modern disk technology can pack 150 billion bits per mm^2, the RAMAC supported a storage density of approximately 4 bits per mm^2 (100 bits per inch on each track with 20 tracks per inch). In order to improve the signal-to-noise ratio for the recordings, early disk technology attempted to reduce the flying height of the head to the disk surface, which made their debut in the 1960's. Approximately 15 years after the RAMAC, IBM introduced the 3340 using what was termed "Winchester" disk technology, which is considered the ancestor of modern disk technology. The 3340 introduced a thin-film inductive head that flew just 18 um from the surface on a cushion of air. The bernoulli effect enabled the head to fly that close to the disk surface. Modern hard drives continue to employ this same basic flying-head approach of these early washing-machine-sized disk units. The flying height for modern disks is now on the order of 15nm (just 15 atoms of air between the head and the disk surface) with the medium flying at 50-150MPH beneath the head for a typical 7,500-15,000 RPM rotational rate.

Disks continue to be the preferred technology for secondary storage on computing devices from desktop computers to the largest scale supercomputing systems. In the following subsections, we will discuss the organization of disk storage devices, technology trends, and their ramifications on design considerations for storage devices for scientific computing systems.

1.3.1 Fundamentals

Ever since the very first RAMAC disk device, disks have been organized as a spinning magnetic media on a platter – with multiple platters in the disk often referred to as a *spindle*. On the platter are concentric *tracks* of recorded data that are often referred to as *cylinders* for tracks located at the same radius on different platters on the same spindle. The tracks are in turn subdivided into

"sectors" that contain a fixed number of bytes comprising a disk *block*. The most common block size is 512 bytes. Consequently, disks are typically referred to as *block storage devices* due to this aspect of their organization.

Because the circumference of tracks on the inner diameter of the disk unit are much smaller than for the outer-tracks, *zonal* recording packs more sectors on the outer tracks of the disk unit in order to maintain uniform bit-density on all tracks. Zonal recording maximizes the storage density of the device, but the sustained bandwidth of data transfers from outer tracks is much higher than that of the inner tracks of the device. Consequently, algorithms for disk filesystems preferentially pack data on the outermost tracks in order to sustain maximum performance. Read and write operations that are smaller than the native block-size of the devices will waste bandwidth because the device works with data only at the granularity of blocks. For write operations, it is necessary to read an entire block, update a subset of that block, and then write the complete aligned block back to the disk subsystem. Consequently, unaligned accesses can suffer, and write operations consume both the read and the write bandwidth of the disk devices. Composite devices, such as RAID, also have a much larger logical block size, requiring much larger transaction sizes to maintain full performance.

Because disks are mechanical devices, there are a number of latencies involved in access. These latencies include: rotational latency, head-switch-latency, track-to-track latency, and seek latency. The disk heads are mounted on mechanical arms that position them over the disk platters. Most disk units operate the heads on all platters in tandem. After the heads are positioned over the correct cylinder or track, the disk must wait for the platters to rotate so that the correct sector is located under the read heads. The latency of the platter rotation is the rotational latency of the disk, and affects the rate at which the sector can be read or written to the cylinder. Read and write operations can only engage one head at a time, so the head-switch time (closely related to the rotational latency of the disk) refers to the amount of time required to switch between heads on different platters. The latency of repositioning the head differs greatly depending on whether it is between neighboring tracks or if it involves random accesses. To speed write operations, many disks employ buffer caches to hide these latencies, but only hide latencies for small transactions. Such buffers cannot mask the reduced throughput that results from the time spent moving the heads. Consequently, sequential (append-only writes or streaming reads) offer the best performance for disk devices. Random access (seeking) presents the worst case read and write performance due to the mechanical latencies involved in repositioning the disk heads. Solid state non-volatile storage devices may offer considerable advantages for such access patterns.

A common misconception about disk performance is that the device interface performance defines the expectations for the disk's performance. In fact, the sustained media transfer performance is often an order of magnitude less than the performance of the device interface. For example disks that adhere to the ATA-100 standard boast 100MB/s transfer rates, but in practice one is limited by the transfer rate of the raw device, which is typically on the order of 20MB/s or less internally. Composite devices such as RAID can share a common bus (such as SCSI or Fiberchannel) to saturate its available bandwidth, and are termed "spindle limited" if the performance of the underlying disk media is less than the available bandwidth of the device interface.

1.3.2 Performance Characteristics

Although there is significant work at the filesystem and operating system level to hide some of the performance characteristics of block devices, most of the characteristics of the underlying devices end up percolating up through the entire software stack and must become first-order design issues

for high performance scientific I/O applications.

The performance of disk devices are substantially affected by the size of the I/O transaction. The behavior is partly due to the disk device's organization, but also due to the POSIX I/O sequential consistency semantics for large scale systems, which will be described in much more detail in chapter 2 of this book. Figure 1.4(a) compares a cluster filesystem (GPFS) to the local filesystem performance (XFS) for varying transaction sizes for contiguous, append-only streaming writes using POSIX I/O. Given the huge performance disparity between best and worst performance shown on this plot, even a single small transaction interspersed with large transactions, can substantially lower the average delivered performance. Small I/O transactions, even for linear operations, can be extremely costly because the entire disk-block must be transfered even when a small subset of the data is updated.

Furthermore, I/O transactions that are aligned to the native block size of the disk offer the best performance whereas unaligned accesses can severely impact the effective performance delivered to applications as shown in figure 1.4(b). The block alignment issues manifest themselves at all layers of the storage hierarchy, and even the filesystem. GPFS is used in this example, but this is common to many different filesystem implementations. In addition, streaming write operations tend to be slower than reads by up to 50% because the need to read the disk block into the system memory, modify its contents, and then write it back to disk. Since read and write operations are mutually exclusive for the mechanical device (disk heads are in-effect half-duplex), then the read-modify-write can impact delivered bandwidth. However, the impact can be modest (<10%) because it is amortized by the track-to-track seek latencies.

1.3.3 Enterprise and Commodity Disk Technology

The disk storage market is typically divided into three market segments – Enterprise, Consumer, and Handheld. The Enterprise class storage market emphasizes performance (reduced failure rates, reduced seek times and increased transfer bandwidth) above all other considerations. Enterprise-class disk interfaces tend to employ host interfaces such as SCSI and Fiberchannel that emphasize reliability, support for multiple outstanding transactions, more robust error checking and correction, as well as support for addressing a larger number of devices on the same loop or channel interface. The spindle speeds tend to be higher than consumer-grade devices, leading to higher power consumption. A typical enterprise class device will have a lower storage capacity per unit than the consumer-class devices, but will spin at up to 15,000 RPMs, offer seek-times of less than 5 milliseconds, and consume 12-20Watts.

The consumer class devices emphasize low cost and high capacity for volume-driven consumer-electronics markets. A typical consumer-grade disk will offer double the storage capacity of the Enterprise-class storage devices, but spin at a lower spindle rate of 5,400-7,200 RPMs. The drive interface for these devices tends to rely more on the host computer interface to control the disks than the enterprise-class devices. Consumer class devices tend to support fewer devices on the same interface, and are less tolerant of faults. For example, IDE (Integrated Device Electronics) devices perform parity checking on the data bus, but not on the control interface. Interface technology is typically derived from IDE, ATA and SATA (Serial Advanced Technology Attachment) standards. The power consumed by such a device is typically limited by applications to less than 10Watts.

The handheld devices favor a design point that emphasizes reduced size and power consumption for devices such as MP3 players, ultraportable computers, and handheld movie cameras. The typical capacity for such devices is an order of magnitude smaller than for enterprise storage devices, but

power consumption is typically on the order of 0.5-2Watts. These devices are also being explored as a component for ultra-efficient composite (RAID) storage devices such as Massive Arrays of Idle Disks (MAID) [3].

While enterprise class storage devices have enjoyed a niche market in large-scale storage solutions, they have been gradually supplanted over time by consumer-grade devices that depend on assembly into hierarchical RAID devices to match the reliability and performance of the enterprise-class devices. Recent studies [5] [6] have cast doubt on claims that enterprise-class disks offer lower failure rates than their consumer-grade counterparts. Consequently, enterprise-class RAID storage subsystems composed of consumer-grade SATA disks are rapidly overtaking the enterprise market segment.

1.3.4 Future Trends

Magnetic disk technologies are faced with the theoretical limitation of superparamagnetism which proposes that as bit density is increased so must power to the device in order to prevent spontaneous changes of data (e.g. a bit flip) from occurring. Because this is a theoretical limitation, the actual limit is not known and continues to change each time a manufacturer produces a higher capacity disk. Several manufacturers are developing solutions to the problem, but the main concern with the limitation is that solutions will come at a significant increase in cost. This would have ramifications for the high performance computing sites that rely on squeezing ever increasing amounts of storage into smaller spaces at a fixed cost.

1.4 Tape Storage

Tape has been the primary medium for offline storage since the 1980's. Magnetic disks or hard disk drives (HDD) are its primary competitor in terms of capacity and data transfer capabilities. Through the early 1990's tape remained slow and well behind disk drives in terms of performance, but well ahead of disk in terms of capacity. Disk storage was still in the MBs of space and disk drives were a premium in terms of cost. In the early 1990's disk storage became commonplace and extremely affordable as the number of personal computers soared.

The key principles or advantages of tape storage that will continue to make it a viable mass storage solution if not the primary mass storage solution well into the future are that they are removable, extremely power efficient, maximize GB/sqft, and continue to offer a competitive price/GB. Like optical storage, it is removable media. This fact allows tape to be an ideal solution for offsite data requirements. Tape is also primarily used for system backups as data can easily be exchanged between systems or stored separate from the system to prevent risk from co-location with the primary data source. Tapes do not require power or a conditioned computer room environment to retain data or be prepared for use. Tape libraries requiring very little power continue to keep tape access times reasonable for sites with requirements to keep vast amounts of data accessible to users. Archival life of tape is very good with most tapes capable of retaining data anywhere from 15-30 years.

1.4.1 Fundamentals

Tape media has been made for about a decade and is composed of a durable substrate, mylar in most media, with single or dual layer magnetic alloys. Tape cartridges have tracks with data blocks or sections written to the tracks. Depending on the tape drive or application using the tape, file marks are written to the tape to serve as markers for moving to different points or data blocks on the tape. The primary downside of tape is its sequential access limitations. Tape does not lend itself to random access as tape is a continuous medium that must be moved forward and backward until the desired position is reached. The further down the tape the desired data resides, the longer it takes to reach the data.

Cartridges can be single or multi-reel. Single reel tapes are fed through the drive with rollers where the tape passes over the recording head of the drive. Dual-reel tapes are loaded into the drive, but keep the tape internal to the cartridge with the recording head of the tape drive positioning itself over the tape between the two reels. Tape drive speed, and tape recording head capabilities determine the tape's data transfer rate. Tape can be damaged given the high velocity and mechanical nature of reading and writing data. However, data loss is normally limited to a small part of a single tape cartridge.

There are several types of tape technology in use in the high performance computing industry. Tape drives utilize SCSI and Fiber Channel protocols. New tape drive models usually demand new tape media formats. Commodity tape drives typically don't plan for media reusability with multiple versions of drives, whereas enterprise tape drives plan for media reuse and backwards compatibility. Most tape drives are capable of and automatically compress data being written to tape. The largest tape formats allow nearly 1 TB of data uncompressed to fit on a single tape. The cost of storing data on tape is typically an order of magnitude less than the cost of storing the same data on disk, irrespective of upkeep, maintenance or power costs.

Access times for tape depend on the tape model used, but for the highest performing tape libraries, they typically range from 30 seconds to several minutes to mount the tape and deliver the first byte of data. The highest performing tape drives are capable of transferring uncompressed data at 180 MB/sec and achieve about 380 MB/sec with compression. Access time is highly dependent on the tape drive's seek capabilities. Tape is fundamentally sequential media and tape marks are essential to finding the exact beginning and end of user data. Tape marks are a unique pattern written to tape by the drive when requested by the application in order to separate user data on the tape. This separation is application specific, some use tape marks to designate the end of a file while others use them to separate blocks of a file. Tape mark processing is expensive and defeats the drive's ability to stream data. Enterprise tape drives include features to quickly seek to a particular place on the tape by avoiding the use of tape marks. One enterprise tape drive utilizes a special track on the tape called the servo track and a special index at the beginning of the tape to quickly position the tape to a particular point. The slowest tape seek times occur when applications move down the length of the tape counting tape marks to locate the data. Some applications, such as the High Performance Storage System (HPSS) software, are careful to allow placement of small files on tape with better access speed and larger files on slower access high capacity tape.

In terms of durability, tape is at least as durable as other data storage technologies. Depending on the amount of reuse, format and density of the tape, and the speed and handling of the drive, tape can be used for several decades to store data.

1.4.2 Enterprise Class Tape Technology

Enterprise tape drives are designed and manufactured under the strictest specifications to ensure extra reliability and durability of user data. Features include redundant components internal to the tape drives to eliminate any single point of failure to the greatest extent possible. At least one tape drive available in the market has dual tape heads, the most expensive component of the drive, to ensure that tape head failure does not interrupt tape drive data transfer capabilities.

Enterprise class drives provide ruggedized components to handle extra wear and tear they receive from high utilization. Tape media wear is also a primary consideration in designing enterprise quality tape drives. Enterprise class tape drives are typically capable of adjusting tolerances, such as tension on the tape or recording head positions, and are an order of magnitude more accurate than commodity tape drives.

Another consideration in differentiating between enterprise and commodity tape drives is the bit error rate capabilities of the drive. Commodity tape drives have bit error rate on the order of 10E-17. Currently, enterprise tape drive manufacturers are able to design the recording head and the pattern of data written to media to reduce bit error rate by up to two orders of magnitude compared to commodity tape drives. One other feature that enterprise class drives provide are multiple control and data path connections to the drive.

The main purpose of tape libraries is to provide storage for tape cartridges not in use and to load or unload tapes in or out of tape drives to the tape library. Enterprise class tape libraries provide redundant components such as robotic arms, grippers, visioning systems, and power feeds to prevent a single point of failure. When failures occur, hot swappable capabilities (such as replacing robotics while the library remains available for use) become another important feature of the industry leading tape libraries. There is currently only one tape library manufacturer that provides an interchange mechanism between libraries to allow cartridges to be mounted in tape drives existing in separate libraries connected via the passthrough port. This is an extremely useful feature in a mixed media multi-tape library environment in that it prevents the need to directly manage the usage of tapes in drives.

Experiences at most High Performance Computing centers with both commodity and enterprise class tape drives support the claim that enterprise class tape drives do actually exhibit less problems in terms of placing user data at risk than the commodity drives.

1.4.3 Commodity Tape Technology

For as long as computer technology has existed in the home, there has existed some form of commodity tape intended for general consumer use. There are several types of tape used in electronics with the most common being magnetic tape.

In the 1980's and 1990's commodity magnetic tape was primarily built around 8mm tape that was mostly used to handle backups of the machine to which they were attached. In the mid to late 1990's the commodity tape drive market opened up with the definition of a new kind of tape drive called LTO that defined a new market sector called mid-range tape.

The linear tape-open (LTO) tape technology has a strong hold in the marketplace today primarily because of its low cost and relatively good performance capabilities and large capacities. It was primarily intended for use with backup applications where absolute assurance of no single point of failure or data loss was not critical. This is due to the fact that backup data is a copy of the original data and in general loses value over time. However, because of its low cost and relatively

good performance and capacity, the technology has started to emerge as a viable option for the most data intensive high performance computing centers; especially for consideration in dual-copy or the lowest tiers of storage in a hierarchical system. If used in a mass storage system with archival storage requirements, most centers simply make multiple copies of data retained on this technology.

The LTO specification has plans to release six generations or form factors of tape. For LTO-6, the final version plans to provide 3.2 TB of data and a speed of 270 MB/sec. Currently, the LTO-4 drive is available and delivers 800 GB of data on a single cartridge and up to 120 MB/sec data transfer speed.

In the last few years, several mid-range tape libraries have emerged in the market that are capable of handling the full spectrum of drives and tape technologies available. Several of these mid-range libraries are starting to adopt features of the enterprise class libraries, specifically redundant components, ruggedized robotics, and expansion capabilities. They are excellent for applications that don't require scalability beyond the limits of the library.

1.4.4 Future Trends

Current tape media technology has a physical limitation relative to the particular process or magnetic coating in use for the last 10 years or so. It is believed that a new media formulation will be required to allow tape to exceed 16 TB per cartridge. Tape capacity is increasing at a rate of about 40% per year.

1.5 Optical Storage

Optical storage looked very promising in the 1980's and 1990's. The new method of storage promised very large capacities and fast access methods in its early years. After two decades, it is apparent that the technology has a definite niche with its most popular formats, the compact disc (CD) and digital versatile disc (DVD), primarily suited for storing music and videos. The major benefit of optical storage today is its wide acceptance in the commercial marketplace and low cost. Optical drives are found on nearly every new computer and function particularly well as read-only media. Unfortunately, the technology was never able to achieve the write access rates or capacities that would allow it to compete with magnetic storage. Optical storage media is an order of magnitude behind magnetic storage in available capacity and transfer rates [11].

1.5.1 CDs

The smallest unit of data on compact discs is called a frame and discs capable of storing 650MB have tracks with pitches (distance between tracks) of 1.6μ m. As the track pitch gets closer together, the CD can hold more data but the ability of drives to read the CD decreases. The highest capacity CDs today hold 700MB or slightly more and have track pitches of 1.5μ m. Data rate in writing to CDs is determined by the speed of the drive used, which is represented by the factor of improvement over a baseline (1x) performance required for digital audio playback at 44KHz. Data rates of modern CD's have leveled off around 48x-52x on writing (48-52 times faster than the baseline rate), which yields a sustained data transfer rate of around 7-9 MB/sec. The archival life of CDs is not

comparable to magnetic storage either as each CD lasts typically 5-7 years before data is at risk. Due to the relatively short archive life, slow transfer speeds, and small capacity in comparison with magnetic storage the technology is not in use in high performance computing or mass storage system environments.

1.5.2 DVDs

The digital versatile discs can hold anywhere from 4-16GB of data currently depending on whether they are single or double sided and single or double layered. Pitch between tracks for DVD discs is 0.73 μ m. Access rates are determined by the speed of the drive used, but can achieve over 20 MB/sec on writes for a 16x drive with double layered discs. Archival life of the media is similar to CD in that a disc is expected to last 5-7 years before data may become unreadable. Speed, capacity and archival life are still only achieving what the oldest and least capable magnetic storage devices can deliver.

1.5.3 Blu-Ray and HD-DVD

Blu-ray Disk (BD) and high definition DVD (HD-DVD) are the newest multi-layer optical storage mediums and are currently capable of holding 25 or 50GB of data per Blu-ray disc depending on whether they are single or dual layer, and 15 GB per HD-DVD disc. Blu-ray has been locked in a multiyear battle with HD-DVD for the digital video market, but the battle has largely been won Blu-ray at this point in time. The much higher storage capacity of this technology is largely enabled by the availability of solid-state blue lasers (hence the Blu-ray name). The much shorter wavelength of blue light can be focused to a much smaller point size on the disk medium. The specification and design for this technology includes plans to expand a single disk to 100-200GB by simply defining new tracks. Current drives are capable of 1x-16x speeds providing anywhere from 6 MB/sec to 16 MB/sec. The specification allows for up to 35 MB/sec. These are substantial improvements over previous generation optical storage, but still not competitive with magnetic storage.

1.5.4 Future Trends

Traditional optical storage technologies have fundamental limits imposed by the physics of far-field diffraction. This means that the limits of the amount of data stored on a two-dimensional disc are determined by the wavelength of light used in the optical drive. In order to increase the amount of data on the disc, new optics with smaller wavelengths (blue and ultraviolet light) are necessary and there are physical limitations on how small optics can get. Holographic storage promises solutions that reduce or eliminate the far-field diffraction limitations by using several different techniques capable of using the entire volume of media available. However, holographic storage continues to struggle with bringing a product to market so it is not clear that the physical limitations of optical media will be overcome.

1.6 Composite Devices

Scaling up the performance of disk technology for server-class systems is increasingly expensive. Composite technologies embody the same spirit as commodity clusters to achieve the performance and reliability goals of specialized proprietary hardware using clusters of much simpler consumer-grade building blocks.

Composite devices also play a role in ensuring protection against media. In particular, the probability of a single-bit-error in typical disk storage media is 1 error in 10^{14} to 10^{15} bits, and this rate has been relatively consistent over time. However, with continued exponential increases in the capacity of storage devices, the likelihood of encountering uncorrectable errors is dramatically increased. Composite devices employ various forms of redundancy and error correcting codes (ECC) to improve media integrity.

1.6.1 RAID

The nomenclature for RAID configurations was formalized as RAID "levels" by Chen, et al, in their 1994 paper entitled "RAID: High-Performance Reliable Secondary Storage" [7]. RAID was originally introduced with five levels, but over time new configurations have been introduced that include nested configurations as well as some proprietary configurations. RAID employs some combination of striping for performance with various approaches to redundancy for fault resiliance. The standard RAID levels are as follows:

RAID 0: This configuration, which was not part of the original 5 RAID levels offers improved bandwidth at the expense of reliability by "striping" data across parallel disks comprising the volume.

RAID 1: This configuration implements mirroring to improve reliability by maintaining an exact copy of data on each disk comprising the RAID volume, but offers no improvement in performance over a single disk.

RAID 2: Implements hamming codes to support Error Correction Code (ECC) using a dedicated parity disk to store bit-interleaved parity information, much as ECC-corrected Dynamic Random Access Memory (DRAM) works. The approach achieves the reliability of mirroring with fewer redundant disks per protected bit, but suffers from poor performance for small transactions relative to mirroring. This form is rarely seen in practice.

RAID 3: Implements byte-level striping with a dedicated parity disk, and relies on the disk controller to identify the failed disk rather than strictly depending on the parity information as is the case for level-2. Consequently, it offers the same performance characteristics as level-2, but employs fewer disks to achieve fault resilience.

RAID 4: Implements block-level striping with a dedicated parity disk, which achieves the same fault-resilience as level-3, but matches the transfer unit sizes to the native sector-size for the device. This organization greatly improves performance by ensuring that the entire disk block is utilized on each read. This also greatly improves transfer performance for small transactions, in comparison to the lower levels of RAID.

RAID 5: Uses block-level striping like RAID-4, but distributes the parity across the disks that comprise the raid set. When a disk fails, the RAID5 can continue to operate at full performance. Some RAID systems allow the failed disk to be replaced and rebuilt while the unit is running. However, the rebuild process is typically very costly and can greatly reduce the performance of the array while the rebuild is in progress. A second failure while the RAID-5 is in degraded state will render the volume unusable, thereby motivating more robust approaches to encoding the parity information used to detect disk failures.

RAID 6: RAID level 6 was not part of the original RAID configurations, but is now commonly considered a standard RAID configuration. RAID 6 introduces an addtional parity block to handle the increased failure rates that are anticipated with extremely large disk configurations. In contrast, RAID-5 uses the simplest case of Reed-Solomon error correction codes, which enables it to handle loss of a single disk. However, technology trends and large data centers have increased the probability of seeing multiple disk failures within the same RAID block. RAID-6 extends the Reed-Solomon error correction field so that it can accommodate multiple simultaneous disk failures. The extended error correction enables RAID-6 to continue to operate in the presence of more than one simultaneous disk failure.

RAID can be implemented in either hardware or software. However, the parity checking and automatic volume rebuild process for RAID-2 and higher typically benefit from dedicated hardware. This has led to a robust hardware controller market for RAID systems. Given that RAID-0 and RAID-1 have less-demanding requirements for fault detection and correction, many operating systems incorporate logical volume managers that support concatenation of volumes, mirroring, or striping of volumes in software.

As large deployments are becoming more common, hierarchical implementations of RAID technology have become more common. The terminology for this hierarchical structure is RAID M+N or "RAID MN" where N is the baseline building block that is composed together in RAID-M fashion. So, for instance, a RAID 05 (also known as RAID 0+5) is a RAID 5 array comprised of a number of striped RAID 0 arrays, whereas a RAID 50 is a RAID 0 array striped across RAID 5 elements. RAID 50 and RAID60 arrangements are typically the most commonly employed hierarchical RAID implementations.

1.6.2 Virtual Tape Libraries(VTLs)

Virtual tape libraries are a somewhat new concept in storage and primarily came about due to the cheap cost and prevalence of disk. The concept of a VTL is to use disk to mimic tape such that the application using the VTL doesn't know that its manipulating disk rather than tape. These are primarily used in backup applications to eliminate dependence on streaming to tape for good performance and to increase the speed of recovery. Testing with at least one VTL in an archive application that writes small files or blocks of data to tape showed only a 2x improvement over a high capacity tape drive. The thought was that the VTL would do much better given that the tape drive used did not handle small files with many tape marks well. In testing, the site discovered that disk isn't that much faster or more efficient than tape at processing tape marks. One major problem with VTLs is that they have a cost somewhere between tape and high-end disk arrays. For mass storage systems, there seems to be little need or use for VTLs when most are already hierarchical storage systems with both disk and tape being used for both performance and cost reasons.

There have been other attempts at placing magnetic disk drives in tape cartridges so that they could fit in tape robotic libraries and mimic tape cartridges. However, there are few applications that can make use of removable disk and the disk drives must be ruggedized for extra wear and tear. None of these solutions have come to market.

1.6.3 Redundant Arrays of Independent Tape (RAIT)

Many high performance computing sites could take advantage of redundant tape systems or redundant array of inexpensive tape (RAIT). There are several hardware or software solutions on the market, but none that satisfy the needs of providing extreme bandwidths and parity protection. The hardware solution from Ultera provides mirroring or parity protection with a SCSI hardware controller card that connects SCSI tape drives together to perform RAIT. However, the solution is limited in bandwidth for what high performance storage would require. The software solutions available, primarily HPSS, Veritas or Legato, generally meet high performance bandwidth requirements but do not provide parity protection to reconstruct data in the event that a tape within the stripe is lost. At the time of this writing, HPSS is currently working to design and develop a RAIT solution for its customers.

One of the most promising RAIT systems, developed by StorageTek, made it to the prototype stage demonstrating the loss of two tape drives during a read while continuing to stream data to the application [10]. During the final phase of development of turning the prototype into a production system, the company halted the project as the market demand for the product was believed to be too limited.

1.6.4 Data Integrity

Media integrity is often confused as being equivalent to data integrity, but the important and consequential differences are often underappreciated. Composite devices only protect against data integrity problems with the disk media, but that does not ensure the integrity of data produced by a scientific application. From the application standpoint, any intermediate step along the complex pathway to storage can compromise data integrity before it arrives at the storage device, which will dutifully record the incorrect values to permanent storage. Unfortunately, commonly deployed storage hierarchies (for example, memory to disk to tape) provide no such end-to-end data integrity checking, so it is left to the application developer to employ various forms of checksums and other tests to verify the integrity of their data.

CERN (the European Organization for Nuclear Research) recently performed a study of data integrity issues on their own systems [12] by writing 3,000 special 2GB files of a predefined pattern every 2 hours, and then reading them back to check for errors every 5 weeks. What they found was that even with RAID to protect against media errors, they observed 300 uncorrected and unreported errors in the 2.4 petabytes of data that were stored on the data volumes. In all, after examining 8.7 TB of user data for corruption (33,700 files), they found 22 corrupted files that had not been detected by any part of the storage infrastructure. In some cases, the corruption remained undetected because it was caused by errors in various tiers of the software infrastructure rather than random bit-flips. Ultimately, CERN found an overall byte error rate of 1 in $3*10^7$ bytes, which is considerably higher than the media error rate of 1 byte in 10^{14}. Without some form of end-to-end monitoring of data integrity it is clear that the larger disk based storage systems will observe data inconsistencies. Focusing on media alone for data integrity protection is insufficient.

There are emerging storage technologies such as ZFS [13] (Zettabyte File System) that are capable of ensuring end-to-end data integrity. However, until such systems are widely deployed, it is important for scientists to incorporate their own checksum checks (such as Cyclic Redundancy Check/CRC or Message Digest/MD5) and error detection mechanisms into their data storage practices to protect against silent data corruption.

1.7 Emerging Technologies

This section describes technology that is poised to compete with mainstream tape and disk technology, but for various reasons has not yet overtaken these technologies. The reasons include manufacturing economics in a volume market and technologies that are still in development. Many of these technologies are already available, but have had a limited impact due to market economics or a narrow band of applications where the technology offers superior price/performance. Others are still in development; it remains to be seen if an effective volume manufacturing process can be found to bring them to market.

1.7.1 Massive Array of Idle Disks (MAID)

Given current technology trends for rotating disk storage, data centers are increasingly reliant on larger numbers of disk spindles to maintain I/O bandwidth growth rates that match performance improvements of the computing infrastructure. However, as commensurate power consumption of such system grows, there is increasing concern for reigning in the power consumed by the storage subsystem. The concept of a Massive Array of Idle Disks (MAID) is to power down disks that are not in use so that power savings may be realized. A secondary benefit of powering down the disks is that they have higher reliability for lack of use. MAID technologies normally use SATA disks which allow the overall system to present an extremely dense array of disks in a small footprint for a very competitive price.

There are several MAID technologies on the market today. And new products that are not explicitly MAID systems continue to adopt similar power management features. MAID systems are typically close to the cost of enterprise tape drives with the reduced benefit of not being removable or scalable as tape technology provides. There is also concern over the increased wear-and-tear of the disks due to powering them on and off. In many of the mass storage systems in production, the disk cache normally has high utilization and taking advantage of the power down feature would not be possible.

MAID systems are novel and timely now that power management is becoming a larger and larger part of storage and compute centers planning and design.

1.7.2 FLASH

FLASH memory is the evolution of Electronically Erasable Programmable Read Only Memory (EEPROM) technology. A FLASH memory cell, like its EEPROM predecessors, is based on a Field Effect Transistor (FET). Normally, a FET's on/off state is controlled by charging the gate with electrons to apply an electric field to the transistor channel. In the case of FLASH memory

cells, the FET contains a floating gate that is completely surrounded by insulating silicon oxide. The gate is programmed by applying a high (>12V) voltage to the electrically connected FET gate, which causes electrons to tunnel through the insulator and into the floating gate in a process known as *hot electron injection* or *avalanche injection.* The residual electric charge in the floating gate provides enough energy to maintain the gate's programmed state. The gate can be deprogrammed by exposing it to UV light in the case of earlier Electronically Erasable Programmable Read Only Memories (EPROM), or applying a large inverse voltage in the case of EEPROMs and FLASH memory cells. Whereas EEPROM were designed to be programmed and erased as a whole, FLASH extended the technology to allow finer-grained block-level reprogramming of storage cells.

Most digital FLASH storage devices use Single Layer Chips (SLC) which store a binary 1 or 0 in each cell of the chip. Another FLASH technology referred to as Multi-Layer Chip (MLC) can store multiple values per cell as different voltage levels. This can greatly increase storage density, but is more susceptible to failure than the SLC device, so it is typically not used for storing digital data.

Initially, FLASH technology was very low density and high cost, which relegated it to niche applications that were extremely power-constrained. However, with improvements to the technology and silicon chip lithography, FLASH prices have made it increasingly popular for storage in consumer-electronics devices such as MP3 players, and digital cameras. The enormous volumes supported by these applications has brought FLASH memory prices down to the point that they are price-competitive with high-end disk solutions and may drop further still. Already, storage devices in laptops and other portable computers are poised to be replaced by FLASH as an alternative.

From the standpoint of scientific applications, FLASH memory can be read in random access fashion with little performance impact. The typical read access latencies are less than 0.1 ms, which makes them considerable higher performance than mechanical disk units, which offer latencies on the order of 7 ms. However, writing data to FLASH takes considerably longer than the read operation due to the much longer latencies required to program the cells. Whereas read rates can be achieved that approach 200MB/s, the write performance is typically more on the order of tens of megabytes per second or less. Prior to 2008, state-of-the-art NAND-FLASH based storage devices are typically limited to 1 Megabyte/second peak write performance. New high-performance Double Data Rate (DDR) interfaces, and improvements in the cell organization to reduce effective cell size are enabling FLASH to push performance past 100MB/s write and 200MB/s read.

One of the main problems with FLASH memory is that the cells wear out after a limited number of writes. For a typical NAND-FLASH, 98% of the blocks can be reprogrammed at least 100,000 cycles before they fail. As FLASH densities increase through improvements in chip lithography, the problem of preventing cell wear-out becomes more challenging. Solid state disks attempt to mitigate the cell-wear-out problem by using load-leveling algorithms. The load-levelers attempt to spread the write operations evenly across the device so as to reduce the chance of cell wear-out. As a result, given the practical bandwidths available for accessing the device would require 5 years of continuous access before the device will encounter cell wear-out – which is on-par with the mean time between failures (MTBF) of mechanical disk storage devices. However, occasionally the load-leveling algorithm encounters degenerate cases that result in unexpected access delays as data blocks are remapped to maintain even distribution of the writes. The cell wear-out issues with FLASH leave the door open for competing technologies such as Magnetoresistive Random Access Memory (MRAM) to step in.

1.7.3 MRAM

MRAM [8] is the first commercially viable solid-state device to use the principles of "spintronics" that emerged from the discovery of the Giant Magneto-Resistive (GMR) effect (see section 1.2.4 above). MRAM uses electron spin to store information in its cell array. It offers many of the benefits of FLASH memory, such as non-volatile storage and low power, without suffering from the cell wear-out that is inherent to NAND-FLASH technology. In addition to non-volatile storage, it also promises to match the performance of SRAM (typically used for CPU cache memory). The bit densities of MRAM are still an order of magnitude below that of leading-edge FLASH memory implementations, but the technology is maturing rapidly. It may prove to be a strong competitor, and possibly the heir-apparent to FLASH memory in portable consumer electronics devices. It is likely to compete with FLASH densities and cost-competitiveness in the 2010 timeframe given current trends in the improvement of this technnology. In the short term, MRAM competes against FLASH for lower-density applications that require DRAM-like write bandwidths.

1.7.4 Phase-Change Technology

Phase-change memory is another form of non-volatile storage that is still in development stages. Phase-change storage devices rely on using current-induced heating to reversibly change the chemical composition of Chalcogenide (GeSbTe) material between the cross-points of a wire mesh that forms the storage array [9] shown in figure 1.5(a). The technology is further from commercial introduction than MRAM at this time, but it has the potential to scale much faster to high bit densities using commercial manufacturing processes.

1.7.5 Holographic Storage

Holographic storage promises to resolve the limitations that current optical storage contends with in terms of increasing the capacity of a single disc. Currently, optical storage is focused on decreasing the pitch of tracks or the size of the data format in order to fit more data in the same form factor or limit of a two-dimensional space. Holographic storage uses a volumetric approach to storing data, such that data is not strictly limited by the two-dimensional size of the disc. Holographic storage has been in development since at least the 1990's. No products are available in the market today, but the two companies working on this technology have roadmaps that plan to deliver media roughly the same as the current DVD that will hold 300 GB of data and be capable of transferring data at a rate of 20 MB/sec. The roadmap extends the media to a planned maximum capacity of 2 TB. Initial products are expected to be write-once read-many with future plans to handle re-writable media.

1.7.6 Direct Molecular Manipulation

Continued improvements in cost-effectiveness of devices is closely related to storage densities of the medium. As we press towards areal densities that approach the atomic scale, there has been increased interest in technologies that can encode data by directly manipulating atomic structures. Direct mechanical manipulation attempts to make novel technology cost-competitive with current storage devices by leapfrogging their areal densities. Some examples include direct manipulation of atoms using Scanning Tunneling Microscope (STM), nanotube devices that depend on van-der-

wals interaction between crossed tubes, and IBM's Millipede device (shown in figure 1.5) that uses many parallel micromechanical (MEMS) styli to encode data into a polymer medium. Each of these devices promises storage densities that exceed the current magnetic limit of 1Tbit/mm2. However, such devices are still in their infancy.

1.8 Summary and Conclusions

Mechanical magnetic storage devices such as disk and tape, have been the dominant technology for secondary storage for the past 30 years. Although solid state devices, such as FLASH have been gaining ground over the past few years, areal density and cost trends ensure that disk will remain competitive for at least the next decade.

One important trend that may complicate future storage for the highest-end scientific computing system is the growing gap between disk capacity and the delivered bandwidth of these devices. Storage trends continue to show 40-60% per year compounded growth rate for storage capacity thanks to continued improvements in areal density. However, the bandwidth delivered by these same disk subsystems is only growing by 17-20% per year. The performance of such systems for random access has become nearly stagnant, which favors linear streaming read or append-only write operations. If these trends continue unchanged, HPC systems will be forced to purchase larger numbers of disk spindles over time that are accessed in parallel in order to maintain existing balance ratios between the HPC system performance and storage subsystem bandwidth. As such, the disk subsystem will likely consume a larger fraction of the area and power budget for future systems without some technology change.

As a result of requiring more disks to achieve performance requirements, HPC centers are deploying file systems that are for the first time eclipsing the size of archival storage systems. Considering the sustainability paradox, presented in section 3.2.1, it is likely that users will need to more carefully consider how best to use archival storage systems to manage their most important data.

In addition, as the number of spindles continues to increase the likelihood of device or even RAID system failures occurring increases, as does the rebuild time for RAID systems that are able to recover. These are challenges that have recently been addressed with innovations in offering new levels of RAID. This problem is discussed in some detail in chapter 2. However, at some point storage devices will not be able to handle error detection or recovery on their own and will require innovation in other technologies to solve (i.e. file system checksums,...).

Several technologies emerged to keep power consumption under control and to fill the widening gap between primary storage (DRAM) performance and secondary storage (disk) performance, such as MAID and FLASH.

Solid state Non-Volatile Random Access Memory (NVRAM) technologies such as FLASH are becoming cost-competitive with the high-end of disk technology and may soon reach parity with consumer disk storage due to dramatic rise of mass-market applications. NVRAM technologies address issues of poor response to random accesses due to lack of mechanical "seek" cost, power dissipation, and bandwidth scalability of conventional disk devices. However, cell-wear-out and poor write performance of FLASH relative to the mechanical devices keep impact on high-end scientific computing storage marginal. While load leveling technology offers some protection against wear-out, the algorithms are still subject to edge-cases where intensive recopying of data is required.

In the interim, FLASH may offer advantages for read-intensive storage applications such as data-mining applications, but for write-intensive applications (such as data output from time-evolution simulation codes or checkpoint/restart), we may need to wait for commercialization of alternative NVRAM technology that doesn't suffer from cell wear-out such as MRAM, or phase-change devices.

Tape subsystems are also seeing new demands that run counter to their original performance characteristics. Formerly, there was emphasis on streaming performance of tape systems for moving small numbers of very large files. However, over time archival storage, even at scientific computing facilities, has been increasingly dominated by large numbers of small files. For a sequential access medium, managing many small files using current tape technology poses daunting technical challenges, especially as older smaller capacity devices are replaced by media that hold more and more data. Tape speed is stable, which is desirable as increased speed means increased wear-and-tear on the tape. User wait time to first byte of data will continue to increase linearly as tape media capacity increases. Current high performance archival storage management software, such as the High Performance Storage System (HPSS), are historically geared towards handling large files where data can be streamed to tape. HPSS is planning to deliver a feature to enable aggregation of small files when they are migrated to tape to address part of this problem. However, such systems will need to be refactored to better handle small files by providing policies that optimize aggregation based on the access patterns to the tape.

Every few years, there have been whitepapers questioning the long-term viability of tapes, and the likelihood of them being replaced by spinning disk storage. Despite this, the cost-performance and power-performance of tapes continue to maintain order of magnitude benefit over disks and even NVRAM devices. Due to their enormous surface area and the efficient storage layout offered by helical scan heads, tapes maintain high storage density despite being far below leading edge areal densities offered by leading-edge magnetic storage technology, and thus provide very little pressure on vendors to push the limits on the technology. Were a competing technology to emerge that put pressure on tape market, the tape technology vendors have significant headroom to improve densities and price-performance. However, such a competitor has yet to emerge and current power, density, and storage trends for disk make it unlikely to be the likely successor to tape.

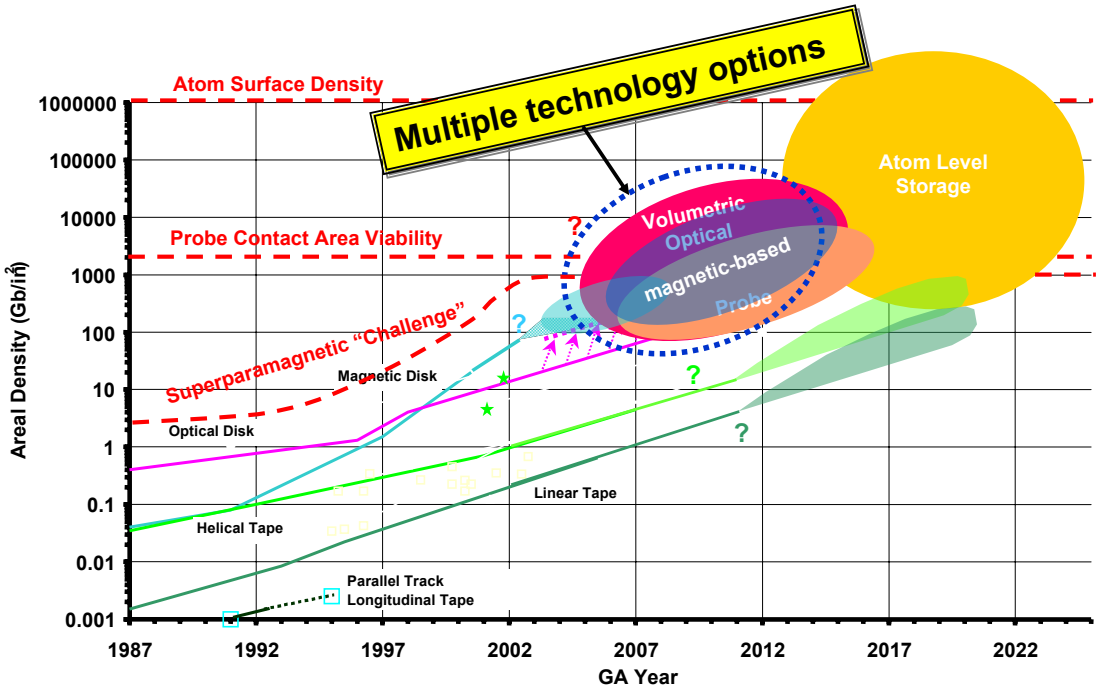
It is not clear that there is a viable competitor to the tape market for the lowest tiers of storage in mass storage systems primarily due to the cost and power savings that tape continues to provide. However, a potential future competitor to the tape market, namely holographic storage, continues to remain on the horizon. Holographic storage has been "a year away from production," for at least two decades, with issues that continue to challenge its ability to achieve commercial production. Holographic storage is much like optical storage in that it will likely have its niche applications and uses within the storage industry when it arrives, but will take a while to develop characteristics or features that would make it competitive with the demands that high performance computing centers require of disk or tape systems. Likewise, other passive technologies such as direct molecular manipulation (IBM's millipede) are unlikely to compete with tape on the basis of cost, density, power, or streaming performance alone – but they do offer higher performance for random accesses, which would be more appropriate for storage of large numbers of small files.

In summary, there exist many new and exciting emerging storage technologies with interesting and unusual storage characteristics. In the near term, none of these are likely to disrupt the current high performance computing industry trends of primarily using disk and tape to store data. As the petascale computing age begins, the storage industry is likely to see another significant increase in the amount of data stored and retrieved from these larger and more capable systems. The leading challenges to disk storage are power and reliability or data integrity. However, tape will

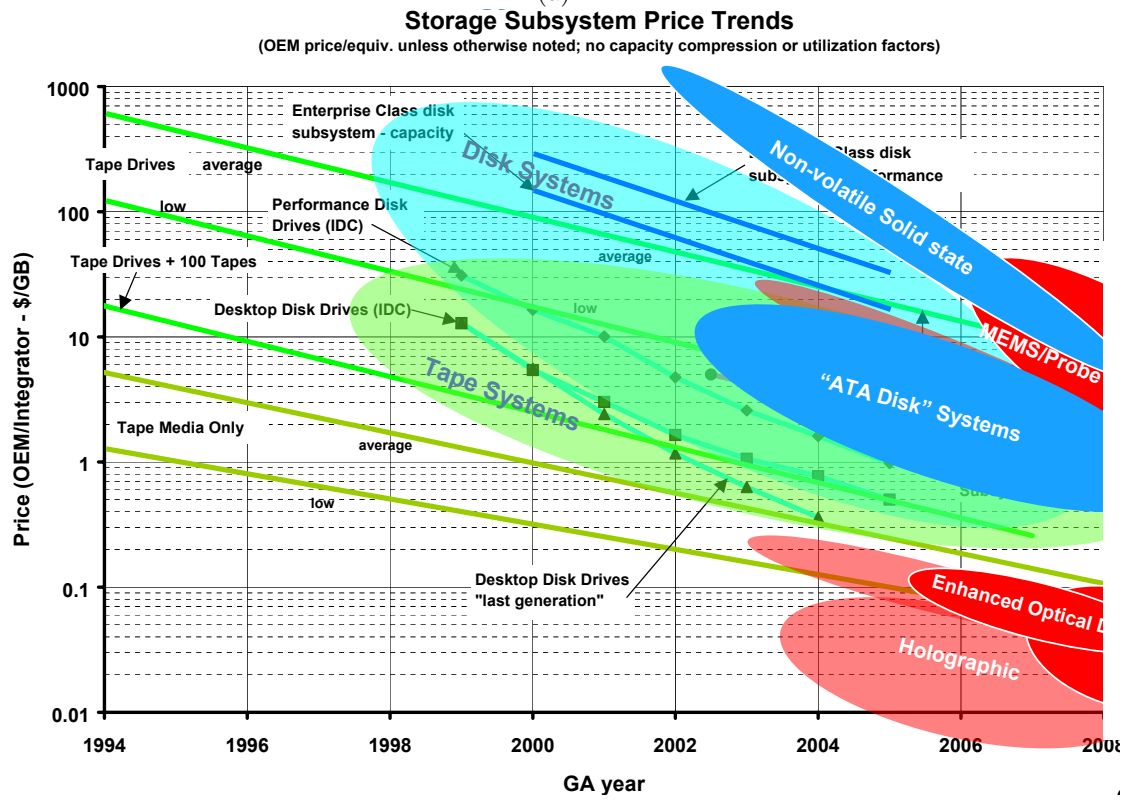
be challenged with increases in access times and the amount of data at risk as the size of a single cartridge increases. These are problems that the storage industry is working to solve, but at some point or scale will cease to be reasonable for any storage device to handle. This demands solutions, potentially by data management software to improve on the reliability and access to ever increasing amounts of data.

Acknowledgments

The authors were supported by the Office of Advanced Scientific Computing Research in the Department of Energy Office of Science under contract number DE-AC02-05CH11231. We also thank Stephen Cranage from Sun Microsystems for his contribution of technology roadmap diagrams to this document.

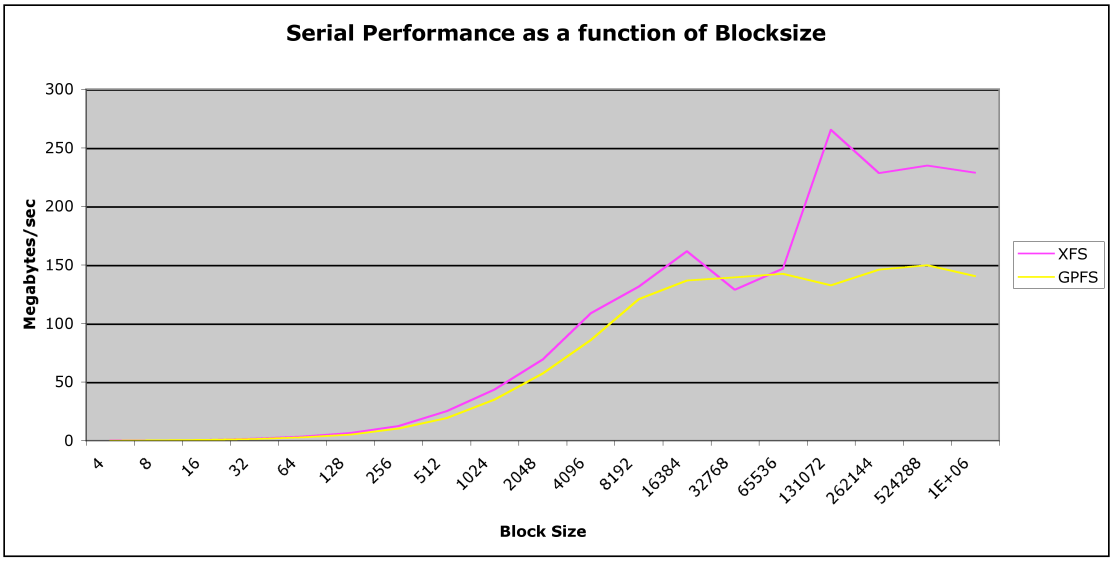


(a)

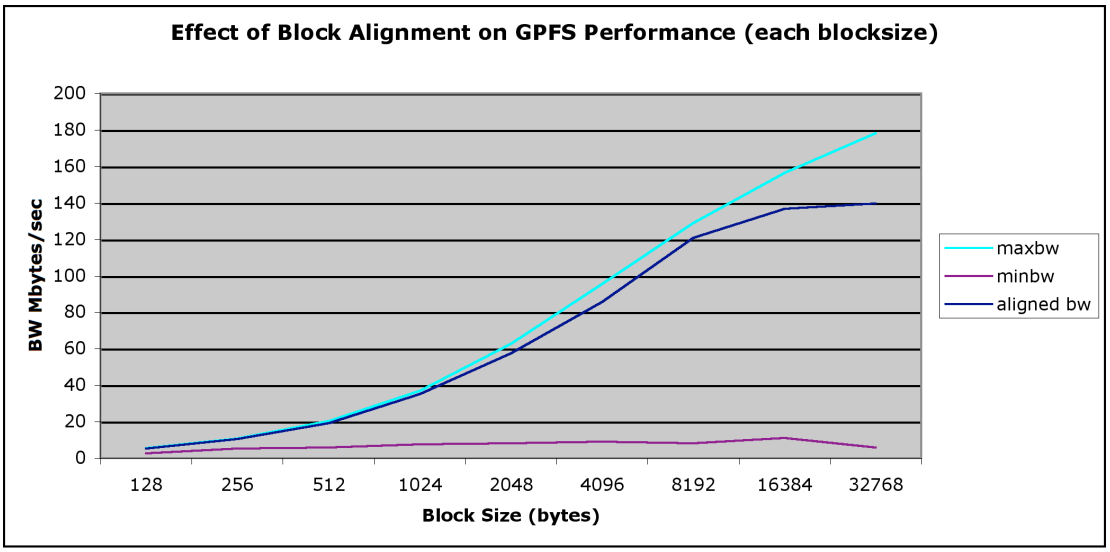


(b)

FIGURE 1.3: This figure summarizes the evolution of area density and price for storage devices (Courtesy of Sun Microsystems). Item (a) summarizes the trends in the areal density for different classes of storage media. Graph (b) shows the cost per bit for different classes of storage media. .



(a)



(b)

FIGURE 1.4: This figure (a) compares a cluster filesystem (GPFS) to the local filesystem performance (XFS) for varying transaction sizes for contiguous, append-only streaming writes. Figure (b) shows that the performance of disk devices, even for relatively large transaction sizes, is substantially impacted by the alignment of the data transfers relative to the native block boundaries of the disk device.

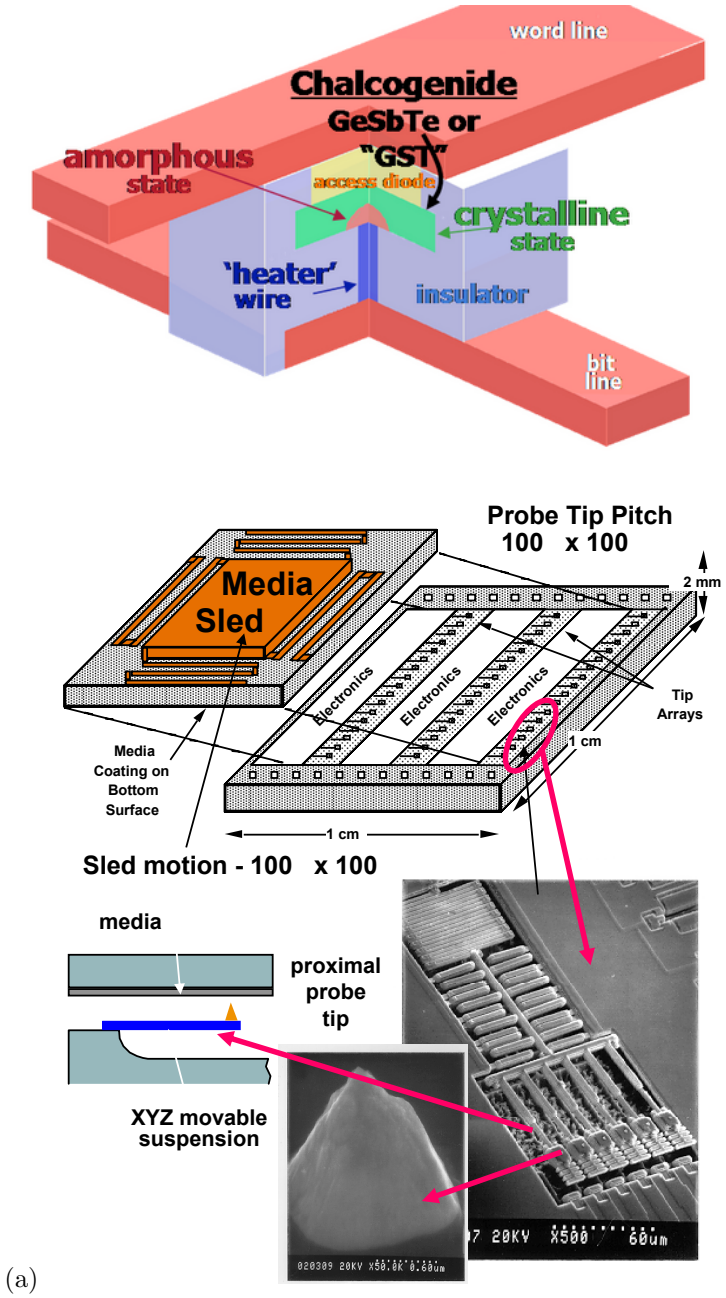


FIGURE 1.5: Part (a) shows a schematic of a single cell of a phase-change memory cell that uses electrically induced heating to induce phase changes in the embedded Chalcogenide (GeSbTe) material. Part (b) shows the architecture of the Micro-Electro-Mechanical system (MEMS)-based Millipede storage system that employs small STM heads to directly manipulate a polymer medium for storage.

References

[1] K. Asanovic, R. Bodik, B.C. Catanzaro, J.J. Gebis, P. Husbands, K. Keutzer, D.A. Patterson, W.L. Plishker, J. Shalf, S.W. Williams, K.A. Yelick. The Landscape of Parallel Computing Research: A View from Berkeley. *EECS Department, University of California: UCB/EECS-2006-183*, 2006.

[2] S. H. Charap, Pu Ling Lu, and Yanjun He. Thermal stability of recorded information at high densities. *In IEEE Transactions on Magnetics,* volume 33, pages 978-983, January 1997.

[3] D. Colarelli and D. Grunwald Massive arrays of idle disks for storage archives. in Supercomputing02: Proc. ACM/IEEE Conference on Supercomputing. Los Alamitos, CA, USA: IEEE Computer Society Press, 2002, pp. 1 - 11.

[4] S. P. Parkin. The Spin on Electronics. *ICMENS,* pages 88-89, 2004.

[5] E. Pinheir, W. D. Weber, L.A. Barroso. Failure Trend in a large Disk Drive Population at Google Inc. *In Proc. of the 5th USENIX conf. (FAST07)*, 2007.

[6] B. Schroeder, G. A. Gibson. Disk Failure in the Real world: What does an MTTF of 1,000,000 hours mean to you. In Proc. Of 5th USENIX conf. (FAST07), 2007.

[7] Peter M. Chen, Edward K. Lee, Garth A. Gibson, Randy H. Katz, David A. Patterson. RAID: high-performance, reliable secondary storage. *ACM Comput. Surv.*, 26(2), 1994.

[8] W. J. Gallagher, S. S. P. Parkin. Development of the magnetic tunnel junction MRAM at IBM: from first junctions to a 16-Mb MRAM demonstrator chip. *IBM J. Res. Dev.*, 50(1), pp5-23, 2006.

[9] Y.C. Chen et al. Ultra-Thin Phase-Change Bridge Memory Device Using GeSb. *International Electron Devices Meeting (IEDM) of the IEEE International Solid-State Circuits Conference.* San Francisco, February 2007.

[10] J. Hughes, C. Milligan, J. Debiez. High Performance RAIT. *Storage Technology Corporation funded by DOE ASCI Path Forward.*

[11] D. Saird, B.H. Schechtman. A Roadmap for Optical Data Storage Applications. *Optics and Photonics News.*, April 2007.

[12] Bernd Panzer-Steindel. Data Integrity *CERN Technical Report Draft 1.3*, CERN/IT, April 8, 2007.

[13] E. Kusarts ZFS: The Last Work on Filesystems. *http://www.opensolaris.org/os/community/zfs/*, 2007.

0-8493-0052-5/00/$0.00+$.50