

DATA STORAGE TECHNOLOGIES FOR LHC

Les Robertson

CERN, Geneva, Switzerland

Abstract

The paper introduces some of the technologies that could be used for storing and managing the many PetaBytes of data that will be collected and processed at the Large Hadron Collider (LHC) accelerator, which is scheduled to begin operation at CERN in 2005. The state of the art of the current mainline hardware technologies is described, with a discussion of the likely evolution during the next five years. This paper is a summary of the material provided in the lecture notes, which include more detailed descriptions of some of the fundamental technologies and costs issues. The notes also introduce some more exotic techniques, and discuss the storage capacity and performance requirements of the experiments that will use the LHC accelerator.

1 MAGNETIC HARD DISKS

The principal technology used for permanent storage of digital computer data for the last thirty years is the magnetic hard disk. The state of the art hard disk product today can store 72 GigaBytes of data in a container 10 cm wide and 4 cm high, the format of the standard disk slots in a personal computer. The data is stored on twelve rigid disk platters mounted on a spindle that is rotated by an electric motor at around 10,000 rpm. There is one recording head for each disk surface, the heads being mounted on a set of arms (the *accessor*) that can move the heads across the recording surface of the disk. The data is recorded in concentric *tracks* as the disk platter rotates beneath the head.

Magnetic disk platters must provide a magnetic layer suitable for high-density recording and a surface layer that is smooth and durable to support very low head-flying heights. The platter is usually made by sputtering a thin layer of material with high magnetisation and coercivity¹ characteristics on to a rigid aluminium or glass substrate. This is followed by applying a very thin protective layer (diamond like carbon) and a lubricant.

The magnetic layer is usually a cobalt alloy (CoCrTa, CoPtCr, ...). The aluminium substrate is generally first coated with a chromium layer before the magnetic layer is applied. When the disk is powered off, the heads “land” on an area near the centre of the disk, which is laser textured to avoid the heads sticking to the otherwise perfectly smooth disk surface. Current production disks are recorded at a density of 3-4 Gbits per square inch (Gb/in²), while laboratory prototypes have been demonstrated at over 35 Gb/in². The areal density of delivered products is increasing at around 60% per year, requiring continuous improvements in materials, manufacturing techniques and head technology.

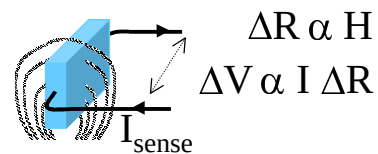
The area used to store individual bits decreases as the storage density increases. These smaller bit sizes imply smaller grains in the magnetic material, and higher coercivity (to sharpen the transitions between the bits). But there are limits beyond which the magnetic polarisation becomes unstable, when the fluctuations in thermodynamic energy at operating temperatures have a moderate probability of causing magnetic state changes. This is called the *super-paramagnetic* limit. This is not

¹ coercivity: The coercive force is the magnetic field which must be applied to neutralise the magnetic orientation of a material. Coercivity is a measure of how stable the magnetic orientation is in the material.

a fixed limit, as it depends on the magnetic properties of the materials, the recording method, shape of the bit pattern, etc. The limit with today's materials and heads is in the region of 40 Gb/in², but it is expected that this will be pushed progressively up as new materials, heads and techniques are applied to the problem. Some industry experts think that 100 Gb/in² will be reached before magnetic recording in the current sense will have to be replaced.

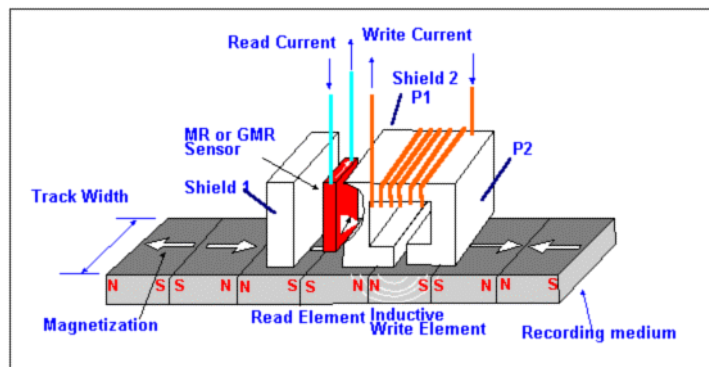
Data is recorded by passing a current through a tiny coil in the record head to generate a magnetic field of sufficient strength to overcome the coercivity of the recording material. Until a few years ago an inductive head was used to read back the information: the fluctuations in the magnetic field as the magnetised bits pass beneath the read head induce a current in the sensing coil. With reducing bit sizes, progressively weaker magnetic fields are available for the read head to sense and during the 1990s inductive read heads were abandoned in favour of magneto-resistive (MR) technology. MR heads use a material such as an NiFe alloy, which has the property that, placed in a magnetic field, its resistance changes with the strength of the magnetic flux. In an MR head such a material is used to form a conductor placed perpendicular to the plane of the recorded medium. A sense current is passed through the conductor, and the signal appears as a voltage drop proportional to the strength of the magnetic field.

magneto-resistive read head



While inductive heads detect the rate of change in the magnetic field, MR heads sense directly the field strength. The output signal strength is proportional to the sense current, giving several advantages in overcoming noise in high density, high data rate systems. MR heads have an even greater advantage over inductive sensors in low-velocity recording applications such as low power disk, and magnetic tape.

An MR read head still needs an inductive head for writing and erasing. The MR element is sandwiched between two magnetic elements that shield it from the field of neighbouring recorded bits - a bit like the way the tube of a telescope keeps out stray light from the surroundings. To reduce the weight on the accessor the dual head is often manufactured as a single thin film composite head.



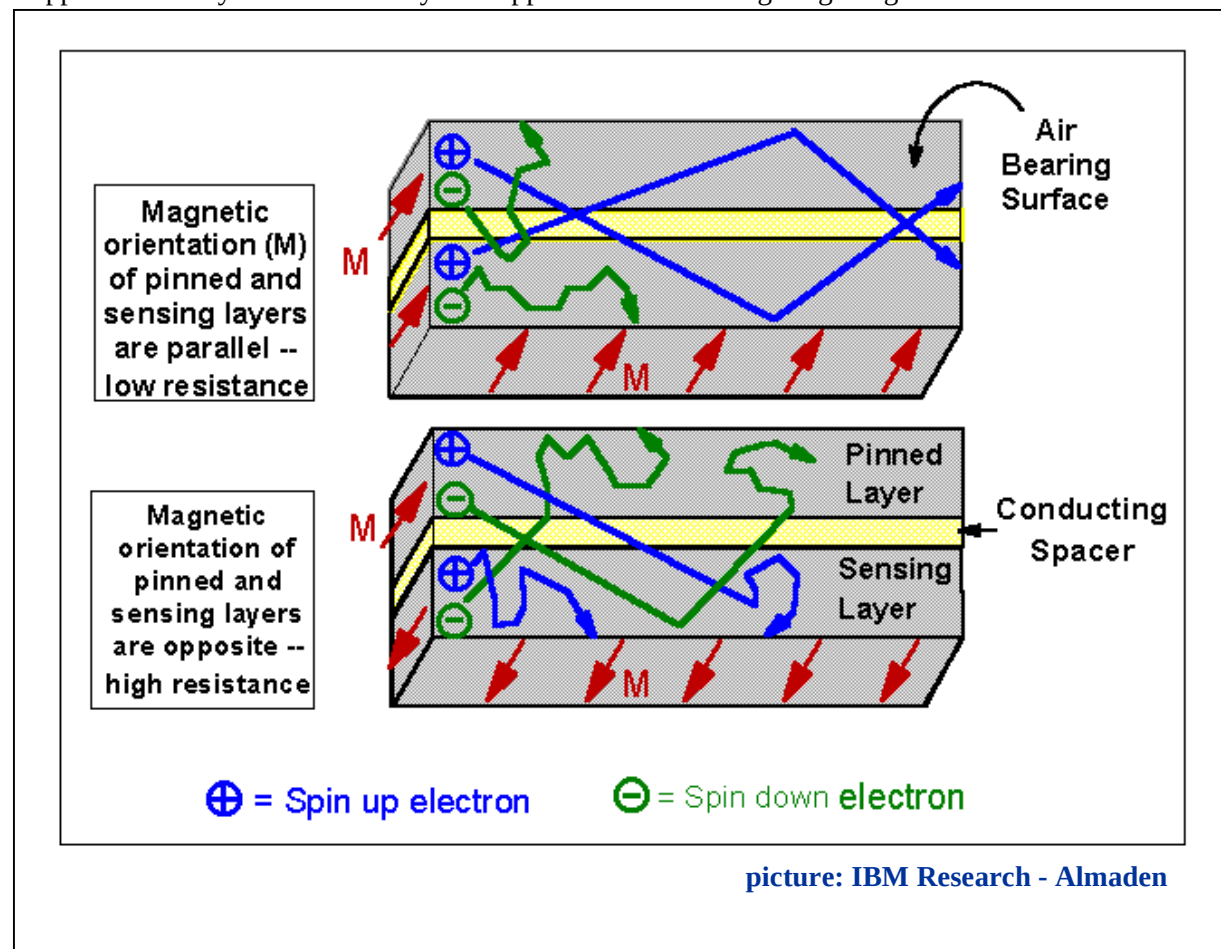
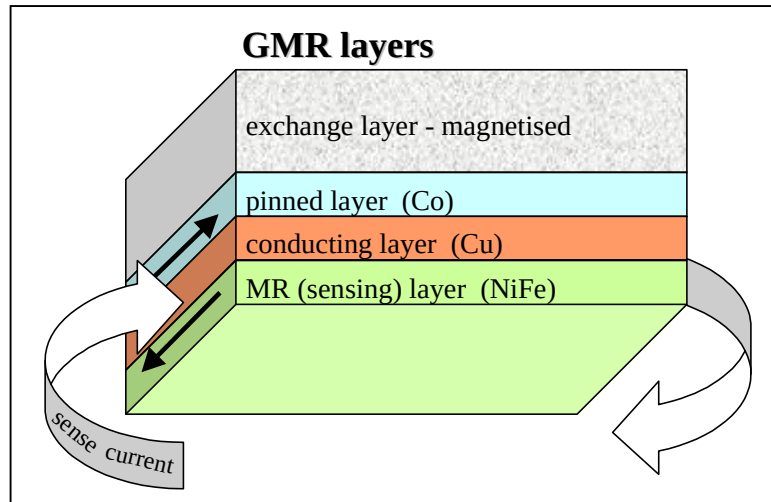
picture: IBM Research - Almaden

Composite head: inductive write, MR read

Today's most advanced heads use the Giant Magneto-Resistive (GMR) effect. GMR sensors are made of four layers. The *sensing*, or *magneto-resistive* layer changes magnetic orientation according to the field of the recorded material. This is covered by a *conducting* layer, separating it from the *pinned* layer, which has a fixed magnetic orientation maintained by the permanently magnetised *exchange* layer.

The sensing current passes through the first three layers, electrons moving freely within these layers. The physical mechanism used depends on the electron spin. Electrons with a direction of spin parallel to the magnetic field encounter low resistance, while electrons with a spin opposite to the magnetic orientation of the material encounter higher resistance.

Because a GMR head exploits the different behaviour of conduction electrons with spin parallel to or opposed to the magnetic orientation of the field it is also referred to as a *Spin Valve*. The following diagram explains the operation in more detail. When both pinned and sensing layers have the same magnetic orientation, some electrons are stopped in both materials, while electrons of opposite spin flow freely in both materials. On the other hand, when the pinned and sensing layers have opposite orientations, electrons of both spins tend to be stopped when they flow into the layer of opposite orientation - giving a higher resistance.



The effective density of recording is measured as the number of bits per square inch of recording surface. Today the bits are much wider than they are long because the narrowness of the track is limited by various mechanical factors concerned with positioning of the heads and track-

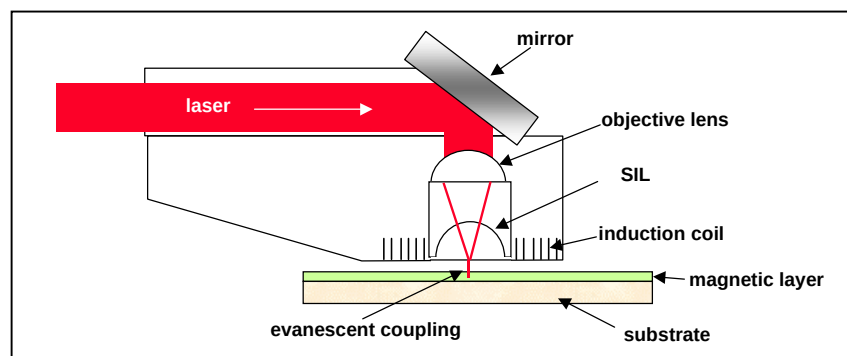
following. The length of the bit is less constrained, limited only by the electro-magnetic properties mentioned above, and the size and sensitivity of the head. At an areal recording density of 10 Gb/in² the bit is about 1 micron wide and 0.06 microns long.

The speed with which data can be recorded and read back from a hard disk is proportional to the linear recording density (number of bits per inch measured along the track) and the rotational speed of the disk. The linear density is increasing at about the square root of the annual rate of increase of the areal density, or some 26% per year. However, the rotational speed has increased by barely 50% in the past ten years. We shall therefore see a growing discrepancy between data rate and disk capacity. It has not been found practical or cost-effective to use parallel heads to increase the data transfer rate, and this problem is more likely to be alleviated by the continued decrease in the overall size (capacity) of the disk unit.

Before beginning to transfer data from the disk the head must be positioned over the correct track (*seek* time), and then we must wait for the beginning of the data segment to come under the head (called the *latency*). The *seek* time has hardly improved by 50% over the past ten years. The *latency*, the time taken for half a revolution of the disk, has also not improved very much in that time. Without a data cache in the disk controller this would lead to very poor performance for small data transfers. The data cache, several MBytes in size in current disks, uses techniques like read-ahead buffering, and speed matching between the disk and the data channel to improve performance.

There is some way to go in further developing the current hard disk technology, before the physical limits are reached, but several alternative technologies are being explored which could ultimately lead to much higher recording densities. One such line of development uses materials developed for magneto-optical devices. These are composites of rare earths and transition metals (such as GdFeCo, TbFeCo), which have very stable magnetisation at normal operating temperatures, and a low Curie point². The material is heated above its Curie point by a laser, and then magnetised using an induction coil. The data is read back using the conventional magneto-optic technique - the magnetic field changes the polarisation of the reflected laser beam (Kerr effect).

However, the recording density of laser-based systems is limited by the resolution of the optics to about one half of the wavelength, which is 0.35 microns for red light. A novel technique used to reduce the laser spot size combines a *Solid Immersion Lens* (SIL) with the *near field* effect, or *evanescent coupling*. The SIL, based on work performed at Stanford University, focuses the laser beam internally on the surface of the lens, which is made from a material with a very high refractive index. The energy of the laser is then transferred between the lens surface and the recording layer using the near field effect which operates within one wavelength of the radiating surface. The TeraStor Corporation has demonstrated a disk using this technique.



² *Curie point*: The temperature above which a material loses its magnetic orientation

2 MAGNETIC TAPE

Magnetic tape has traditionally been used as an inexpensive storage medium for backup of disk data, and for archiving old versions of files. It used to be a clear two orders of magnitude cheaper than disk, and had excellent shelf storage characteristics in terms of volumetric storage density, and ruggedness. More recently, for the past fifteen years or so, large reliable tape robots have enabled gigantic quantities of data to be available cheaply with an access time of a few minutes. StorageTek, the leader in the development of high volume tape robots, uses the name Nearline to describe this class of storage. However, in recent years, hard disk and optical storage have begun to provide strong competition for magnetic tape in terms of both volumetric storage density and cost.

Magnetic tape is made from a thin flexible substrate, about 10 microns thick, made from Polyethylene³. The recording layer used by most high quality magnetic tapes today is microscopic metal particles held in a binder which coats the substrate - MP tape. Metal Evaporated (ME) tape is manufactured using a process in which the substrate is passed through cobalt or cobalt-nickel vapour, which condenses in a thin layer on the surface. ME tape has better recording characteristics at very high frequencies.

The tape is usually stored in a single-reel cartridge. When mounted on a drive the end of the tape is pulled out of the cartridge and attached to a take-up spool. The drive then moves the tape back and forth past the recording heads by controlling the cartridge reel and the take-up spool. The tape path is carefully designed to minimise friction and tape tension. Some systems use cassettes where both reels are mounted in the container. This avoids extracting the tape and is usually much faster in loading. However, the cassette hold less tape than can be packed into a single reel cartridge.

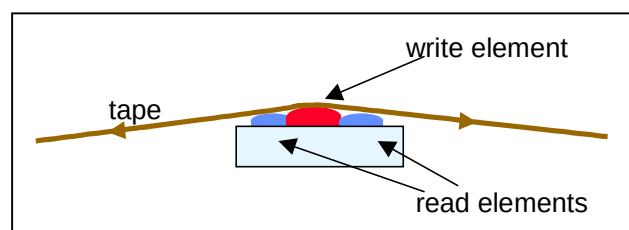
The recording and read technologies use developments from the higher performance and higher density hard disk industry, but there are major mechanical problems with tape which are not present in hard disks.

Magnetic tape is basically a sequential access medium. The tape is loaded at the beginning, and must then be positioned by spacing across intervening data until the desired position is found. This is not entirely true, as modern tapes are recorded in parallel bands recorded along the length of the tape (as we shall see later), but data written on a tape cannot be deleted, or re-written (there is no update possibility). New data must be appended to the current end of data. Most tapes maintain a directory that is written in a special area at the start of the tape, and updated before the tape is dismounted. This contains the addresses of the beginning of each file on the tape, and allows the drive to use positioning information in pre-recorded servo tracks to space relatively quickly to the start of a specific file.

There are two basic techniques for recording data on magnetic tape: *linear*, where the tracks are recorded parallel to the length of the tape; and *helical*, where the tracks are recorded at an angle across the width of the tape.

2.1 Linear Recording

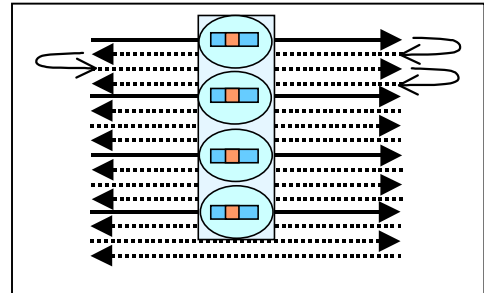
In linear tape recording, the tape passes over a fixed head, which usually contains several sets of recording elements to enable several tracks to be recorded or read back in parallel in order to increase the effective data transfer rate. Each set comprises a write and read element, so that recorded data can immediately be re-read and checked for errors. In order to support bi-directional recording, a third element (read) is required on



³ PET: polyethylene terephthalate, or PEN: polyethylene naphthalate

the other side of the write element. Erase elements may also be required. In linear recording the tape barely touches the head, reducing wear of the head elements and the risk of damage to the tape.

The multi-track head writes a series of parallel tracks along the tape until the end-of-tape marker is detected. Then the head is moved slightly across the tape, and recording proceeds with the tape running in the opposite direction. This continues until all of the data bands have been written. Since the head elements are generally wider than the track, track groups actually comprise a set of parallel tracks, separated by tracks that belong to another track group. This multi-track-group method of recording is called *serpentine* recording.

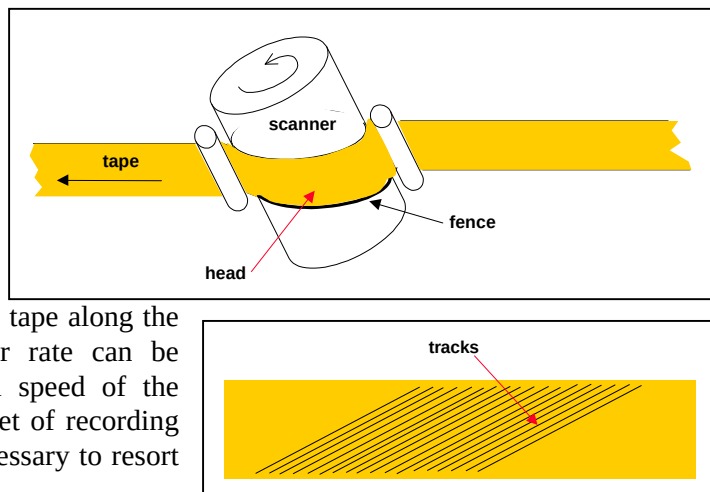


There are a number of practical problems that must be addressed by the designers and manufacturers of magnetic tape systems. Friction between the tape and head leads to head wear and the risk of tape damage. Dust and debris from the slitting process used to cut half-inch tapes from broad strips of medium can exacerbate this problem. The path followed by the tape as it moves from one reel to the other must be as simple as possible, to ensure that minimum tension is required to maintain the extremely thin medium in contact with the heads. The capstans in the path usually employ air bearings to reduce friction. As the temperature and humidity change, so the substrate expands and contracts. The lateral expansion/contraction of the tape poses a major problem for parallel track recording, as the distance between tracks may vary between the time the tracks are written and read. Finally, when the tape is stored on a shelf or in a robot slot gravity slowly distorts the reel of tape, making it more difficult for the tape guidance system to position it. Tape should always be stored vertically, and should be re-spooled regularly to minimise this problem.

2.2 Helical scan recording

Helical scan recording was developed for the entertainment business, in order to obtain higher recording densities - both higher track densities and higher linear densities.

In these systems (used in the mass market VCR) the tape is partially wrapped around a scanner mounted at an angle to the direction of the tape. The tape moves slowly past the scanner, which contains a set of head elements that spin at a much higher speed. The effective relative speed of head and tape is therefore increased without incurring the mechanical difficulties of moving the tape along the tape path at high speed. Data transfer rate can be maintained by increasing the rotational speed of the scanner or by including more than one set of recording elements in the scanner. But it is not necessary to resort to parallel track recording.



There are some fundamental problems encountered with helical scan technology, which make it more difficult to achieve the level of data integrity that can be obtained with linear techniques. The main problems are:

- Head wear due to the increased contact pressure coming from the tape wrap on the scanner (capstan effect).

- The helical wrapping around the scanner is achieved by riding the tape on a fence or step that protrudes from the scanner. This requires that the tape edge has been formed very accurately during slitting, and there is a high risk of edge damage. Once the edge is damaged the recorded tracks develop a curvature which makes it difficult for the read head to follow. In contrast, linear tapes reserve a guard band at each edge of the tape to minimise problems arising from manufacturing or usage damage.

Helical scan tape systems achieve today substantially higher track densities than that possible with linear recording methods. For example, the StorageTek *Redwood* drive achieves 2,800 tracks per inch, enabling it to record 50 GBytes of user data on 1,200 feet of half inch tape. In contrast the 9840 linear tape system from the same company only stores 20 GBytes of user data on a 900 foot tape, at about 600 tracks per inch. New techniques adapted from hard disk technology will allow linear recording to narrow this gap.

2.3 Data compression

One significant advantage of sequential recording systems like magnetic tape is that data compression can be used by the recording system to improve the apparent recording density. This is not possible with a random access system like hard disk, where the physical address of any data block must be able to be computed directly from its logical position in the file. Early compression systems used algorithms that were effective for commercial data streams, but gave little advantage in the case of the random patterns encountered in scientific data or data where straightforward compression techniques had already been applied - such as raw physics events.

More recently, the availability of inexpensive but high performance processors has enabled more sophisticated compression algorithms to be used. In common use today is DLZ1. This is one of the set of Lempel-Ziv compression algorithms which map variable length input strings to variable length output symbols. During compression, the algorithm builds a dictionary of strings which is accessed by means of a hash table. Compression occurs when input data matches a string in the table and is replaced with the output symbol. Advanced DLZ1 algorithms perform well even on pre-compressed scientific data.

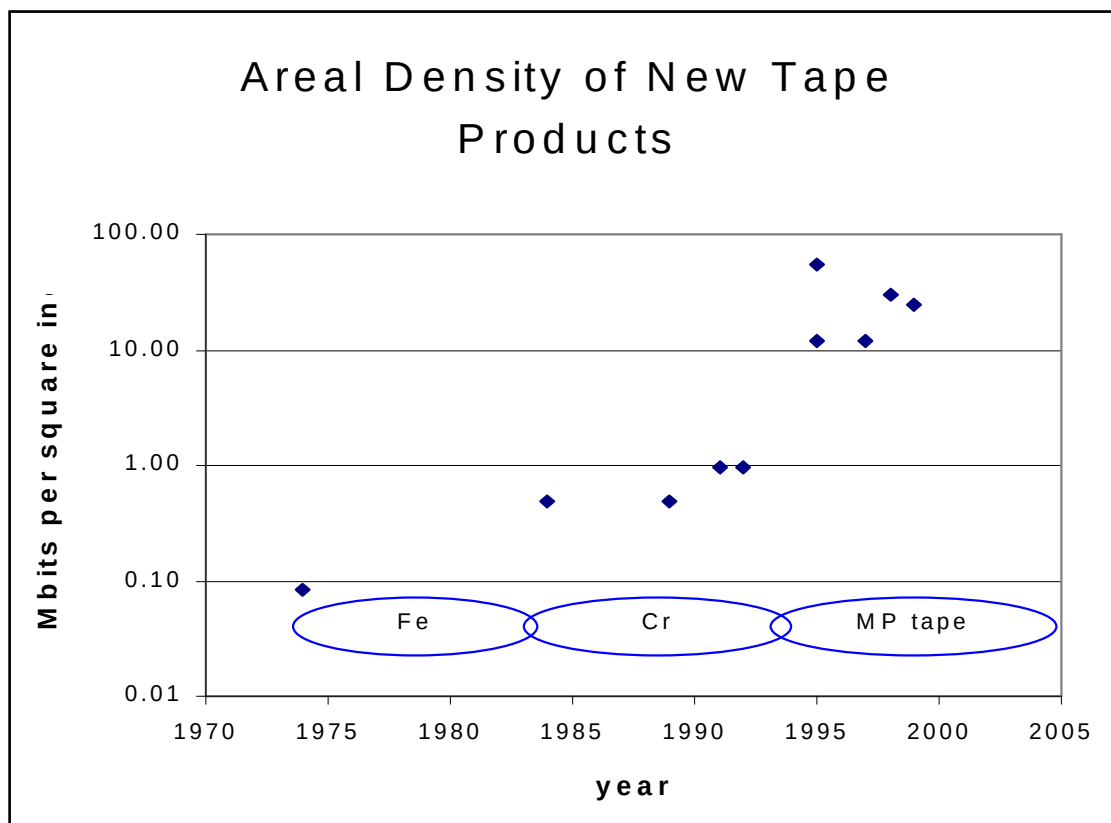
2.4 Technology trends

Unlike in the case of the hard disk market it is rather difficult to identify clear trends in magnetic tape developments. Historically there have been long periods in which a single standard and recording density has dominated the market. During the past 25 years, as can be seen from the diagram below, there have been three distinct generations of data centre quality tape systems.

- the 9-track “6250 bpi” technology using iron oxide which was introduced in 1974;
- the 3480/90 technology, using 18 or 36 tracks on a chrome dioxide medium, which was first introduced in 1984;
- in 1995 the first of a number of new products appeared, all incompatible, but using MP tape, and giving a leap from 800 MByte cartridges to multi-GByte systems

Helical scan has proven to be too difficult a technology for the quality required at the high end - data applications have a completely different definition of reliability from video applications, where helical scan is the standard. Products with acceptable reliability for digital applications are unacceptably expensive. We are not likely to see this change, as linear technology is steadily achieving higher areal densities.

The major applications for magnetic tape are backup and archiving - which require a different level of reliability from applications which keep the master copy of the current data on tape - like high energy physics applications. The model of using tapes as a storage medium for active data needs to be reconsidered.



Since it is difficult to generalise about industry trends, we have to look at specific manufacturers and products in order to have an idea about how things may develop over the next 5 years. The situation at the top end of the market is quite satisfactory, with three or four competitors using proven linear technology aimed at the data centre. The manufacturers have announced “road maps”, explaining how they can realistically develop their systems to improve the capacity and performance by a factor of 2-4 over the next 5 years, without requiring fundamental changes in technology.

At the low end of the market the major products at present use helical scan. As we have seen this is a more difficult technology than linear recording. More important however is the competition that these products, aimed at the mass market, will experience from optical systems.

3 OPTICAL RECORDING

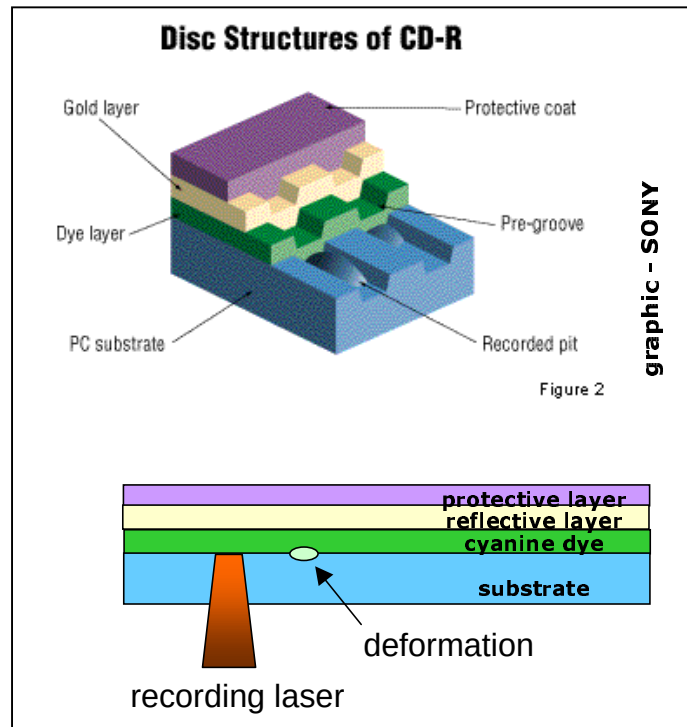
For over twenty years claims have been made that optical recording will displace magnetic recording for secondary or at least tertiary storage. The reason is (or was) that optical systems have the potential for much higher areal densities - up to 5 Gbits per square inch in the case of a red laser. There have been many products using optical techniques, almost all write-once systems, but the achieved density (user data bits per square inch) has been disappointing, performance modest, and costs for media and drives too high. Recordable optical disk was used only for special applications - such as those where the fact that it was write-once was an advantage. Meanwhile magnetically recorded hard disk has steadily improved in density, performance and cost, to the point where it has now exceeded the maximum recording density of purely optical systems.

Paradoxically, although optical recording has lost this fundamental advantage, for the first time we are beginning to see real competition, at least for magnetic tape, coming from optical products. The reason is that recordable CDs and DVDs, products emerging from the mass market of entertainment, have achieved reliability and price levels which place them in direct competition with

low end tape systems. In addition, as these are random access systems, they offer a substantially different (better) level of functionality.

3.1 Write Once - CD-R and DVD-R

Recordable (write once) CDs and DVDs use a disk made from a polycarbonate substrate. The recording layer uses heat-sensitive material, usually Cyanine dye. The heat of the recording laser chemically decomposes spots in the dye layer and physically deforms the adjacent substrate. During playback, the decomposed spot absorbs light from the playback laser, and the substrate deformation scatters the light, together making the spot look like a pit which can be read by a CD or DVD ROM drive. The disk is mechanically preformatted with a flat-bottomed helical groove. The ridge between adjacent grooves is called the land, which has a flat surface. Address information (sector headers, etc.) is embossed on the disk during manufacture.



The disks can be made double sided, and DVD-R disks can store up to 4.7 GBytes on each side. This technology does not support the dual layer per side technique used by mechanically formed DVD-ROM disks. The recording speed of current products is about 1.3 MBytes/second.

3.2 Erasable DVD-RAM

Re-writeable optical disks use phase change technology. The recording layer consists of an alloy of tellurium, germanium and antimony. A laser beam heats the material above the 600°C melting point. The layer cools to a crystalline or amorphous state depending on the power level of the laser. With high power the material melts quickly and cools quickly forming an amorphous spot. At lower power the process is slower and the surface assumes its crystalline state. When reading takes place the amorphous spots have much lower reflectivity than the crystalline spots - so they look like “pits”.

The DVD-RAM format uses a substrate with pre-formed groove and sector headers. Data is recorded in the groove and on the “land”, or ridge between the grooves, to increase the capacity. This involves re-focussing the laser. The original DVD-RAM format has a capacity of 2.6 GBytes per side, but a second version of the standard has been agreed by the DVD Forum which uses a reduced spot size to hold 4.7 GBytes per side.

BIBLIOGRAPHY

- The Complete Handbook of Magnetic Recording, Finn Jorgensen, Tab Books 1995, ISBN 0-07-033045-X
- Magnetic Storage Handbook - Second edition, C. Denis Mee, Eric D. Daniel, McGraw-Hill 1996, ISBN 0-07-041275-8
- Magnetic Recording Technology, C. Denis Mee, Eric D. Daniel, McGraw-Hill Publishing Company 1996, ISBN 0-07-041276-6
- Digital Magnetic Recording, Albert S. Hoagland, James E. Monson, John Wiley & Sons 1991, ISBN 0-47-140144-7
- DVD Demystified, Jim Taylor, McGraw-Hill 1997, ISBN 0-07-064841-7