1.HARD DISKS

When the power to a PC is switched off, the contents of memory are lost. It is the PC's hard disk that serves as a non-volatile, bulk storage medium and as the repository for a user's documents, files and applications. It's astonishing to recall that back in 1954, when IBM first invented the hard disk, capacity was a mere 5MB stored across fifty 24in platters. 25 years later Seagate Technology introduced the first hard disk drive for personal computers, boasting a capacity of up to 40MB and data transfer rate of 625 KBps using the MFM encoding method. A later version of the company's ST506 interface increased both capacity and speed and switched to the RLL encoding method. It's equally hard to believe that as recently as the late 1980s 100MB of hard disk space was considered generous. Today, this would be totally inadequate, hardly enough to install the operating system alone, let alone a huge application such as Microsoft Office.

The PC's upgradeability has led software companies to believe that it doesn't matter how large their applications are. As a result, the average size of the hard disk rose from 100MB to 1.2GB in just a few years and by the start of the new millennium a typical desktop hard drive stored 18GB across three 3.5in platters. Thankfully, as capacity has gone up prices have come down, improved areal density levels being the dominant reason for the reduction in price per megabyte.

It's not just the size of hard disks that has increased. The performance of fixed disk media has also evolved considerably. When the Intel Triton chipset arrived, EIDE PIO mode 4 was born and hard disk performance soared to new heights, allowing users to experience high-performance and high-capacity data storage without having to pay a premium for a SCSI-based system.

Hard Disk :- A type of storage medium that retains data as magnetic patterns on a rigid disk, usually made of a magnetic thin film deposited on an aluminium or glass platter. Magnetic read/write heads are mounted on an actuator that resembles a record needle pickup arm.

Non-Volatile Memory:- Types of memory that retain their contents when power is turned off. ROMs, PROMs, EPROMs and flash memory are examples. Sometimes the term refers to memory that is inherently volatile, but maintains its content because it is connected to a battery at all times, such as CMOS memory and to storage systems, such as hard disks.

MFM :- Modified Frequency Modulation: the data storage system used by floppy disk drives and older early hard disk drives. Had twice the capacity of the earlier FM method but was slower than the competing RLL scheme.

ST506 :- Introduced in 1979, Seagate's ST506 was the first hard disk drive for personal computers. Supporting 5.25in full-height drives with a capacity of between 5MB and 40MB, the ST506 interface became an industry standard for the IBM PC and its successors, eventually being superseded by the IDE interface.

RLL :- Run Length Limited: a method used on some hard disks to encode data into magnetic pulses. RLL requires more processing, but stores almost 50 percent more data per disk than the older MFM (modified frequency modulation) method. The "run length" is the maximum number of consecutive 0s before a 1 bit is recorded.

Areal Density :- The amount of data that's stored on a hard disk per square inch, and equal to the tracks per inch multiplied by the bits per inch along each track. In the context of tape storage, the number of flux transitions per square unit of recordable area, or bits per square inch.

EIDE :- Enhanced Integrated Device Electronics or Enhanced Intelligent Drive Electronics: an enhanced version of the IDE drive interface that expands the maximum disk size from 504Mb to 8.4Gb, more than doubles the maximum data transfer rate, and supports up to four drives per PC (as opposed to two in IDE systems). EIDE's primary competitor is SCSI-2, which also supports large hard disks and high transfer rates.

PIO :- Mode Programmed Input Output Mode: a method of transferring data to and from a storage device (hard disk or CD device) controller to memory via the computer's I/O ports, where the CPU plays a pivotal role in managing the throughput. For optimal performance a controller should support the drive's highest PIO mode (usually PIO mode 4).

SCSI :- Small Computer System Interface: an American National Standards Institute (ANSI) interface between the computer and peripheral controllers. SCSI excels at handling large hard disks and permits up to eight devices to be connected along a single bus provided by a SCSI connection. The original 1986 SCSI-1 standard is now obsolete and references to "SCSI" generally refer to the "SCSI-2" variant. Also features in Narrow, Wide and UltraWide flavours. See also IDE.

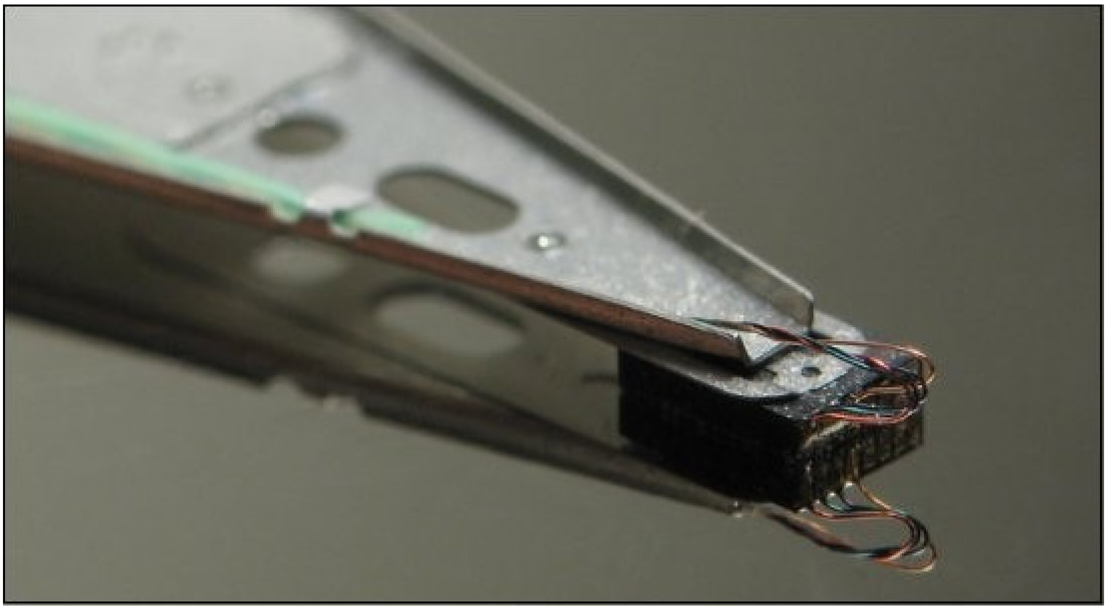
2.Construction

Hard disks are rigid platters, composed of a substrate and a magnetic medium. The substrate - the platter's base material - must be non-magnetic and capable of being machined to a smooth finish. It is made either of aluminium alloy or a mixture of glass and ceramic. To allow data storage, both sides of each platter are coated with a magnetic medium - formerly magnetic oxide, but now, almost exclusively, a layer of metal called a thin-film medium. This stores data in magnetic patterns, with each platter capable of storing a billion or so bits per square inch (bpsi) of platter surface.

Platters vary in size and hard disk drives come in two form factors, 5.25in or 3.5in. The trend is towards glass technology since this has the better heat resistance properties and allows platters to be made thinner than aluminium ones. The inside of a hard disk drive must be kept as dust-free as the factory where it was built. To eliminate internal contamination, air pressure is equalised via special filters and the platters are hermetically sealed in a case with the interior kept in a partial vacuum. This sealed chamber is often referred to as the head disk assembly (HDA).

Typically two or three or more platters are stacked on top of each other with a common spindle that turns the whole assembly at several thousand revolutions per minute. There's a gap between the platters, making room for magnetic read/write head, mounted on the end of an actuator arm. This is so close to the platters that it's only the rush of air pulled round by the rotation of the platters that keeps the head away from the surface of the disk - it flies a fraction of a millimetre above the disk. On early hard disk drives this distance was around 0.2mm. In modern-day drives this has been reduced to 0.07mm or less. A small particle of dirt could cause a head to "crash", touching the disk and scraping off the magnetic coating. On IDE and SCSI drives the disk controller is part of the drive itself.

There's a read/write head for each side of each platter, mounted on arms which can move them towards the central spindle or towards the edge. The arms are moved by the head actuator, which contains a voice-coil - an electromagnetic coil that can move a magnet very rapidly. Loudspeaker cones are vibrated using a similar mechanism.



The heads are designed to touch the platters when the disk stops spinning - that is, when the drive is powered off. During the spin-down period, the airflow diminishes until it stops completely, when the head lands gently on the platter surface - to a dedicated spot called the landing zone (LZ). The LZ is dedicated to providing a parking spot for the read/write heads, and never contains data.

Platter :- A disk made of metal (or other rigid material) that is mounted inside a fixed disk drive. Most drives use more than one platter mounted on a single spindle (shaft) to provide more data storage surfaces in a smaller area.

Servo Platter :- A separate surface containing only positioning and disk timing information but no data. Used only in a dedicated servo system.

HDA :- Head Disk Assembly: The mechanical components of a disk drive (minus the electronics), which includes the actuators, access arms, read/write heads and platters. Typically housed in a sealed unit.

Landing Zone:- The non-data area set-aside on a hard drive platter for the heads to rest when the system powers down.

Form Factor :- The physical size and shape of a device. It is often used to describe the size of circuit boards. The physical size of a device as measured by outside dimensions. With regard to a disk drive, the form factor is the overall diameter of the platters and case, such as 3.5in or 5.25in, not the size in terms of storage capacity. If the drive is a 5.25in form factor it means that the drive is the same size as a 5.25in diskette drive and uses the same fixing points.

Spindle :- The drive's centre shaft, on which the hard disk platters are mounted.

Spindle Speed :- Velocity at which the disk media spins within a hard disk, measured in rpm (revolutions per minute). By the late 1990s EIDE hard disks generally features a 5,400rpm or 7,200 mechanism, while SCSI drives were usually either 7,200rpm or 10,000rpm.

Fly Height :- The distance between the read/write head and the disk surface, made up of a cushion of air that keeps the head from contacting the media. Smaller flying heights permit denser data storage but require more precise mechanical designs.

Head Crash :- Damage to a read/ write head and magnetic media, usually caused by sudden contact of the heads with the disk surface. Head crash also can be caused by dust and other contamination inside the HDA.

Cache Controller :- The circuit in control of the interface between the CPU, cache and DR AM (main memory).

Controller :- The chip or circuit that translates computer data and commands into a form suitable for use by the hard drive. Also known as the disk controller.

Controller Card :- An expansion card that interprets the commands between the processor and the disk drive.

Memory Controller :- An essential component in any computer. Its function is to oversee the movement of data into and out of main memory. It also determines what type of data integrity checking, if any, is supported.

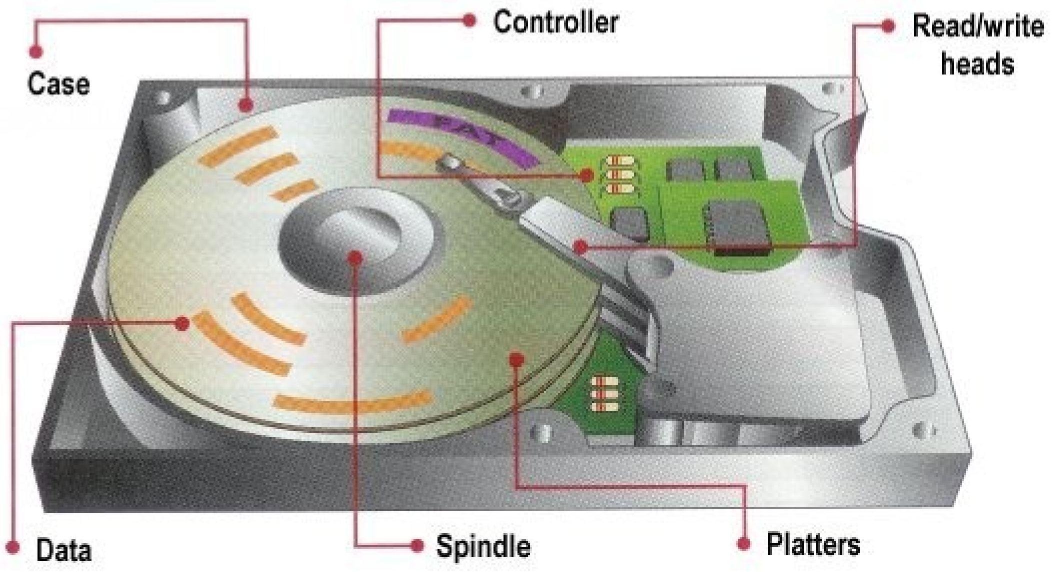
Actuator :- The internal mechanism that moves the read/write head to the proper track. Typically consists of a rotary voice coil and the head mounting arms. One end of each head mounting arm attaches to the rotor with the read/write heads attached at the opposite end of each arm. As current is applied to the rotor, it rotates, positioning the heads over the desired cylinder on the media. Also known as the rotary actuator or positioner.

Voice Coil :- A fast and reliable actuator motor that works like a loudspeaker, with the force of a magnetic coil causing a proportionate movement of the head. Voice coil actuators are more durable than their stepper counterparts, since fewer parts are subject to daily stress and wear and also provide higher performance.

3.Operation

The disc platters are mounted on a single spindle that spins at a typical 10,000rpm. On EIDE and SCSI drives the disk controller is part of the drive itself. It controls the drive's servo-motors and translates the fluctuating voltages from the head into digital data for the CPU.

Data is recorded onto the magnetic surface of a platter in exactly the same way as it is on floppies or digital tapes. Essentially, the surface is treated as an array of dot positions, with each "domain' of magnetic polarisation being set to a binary "1" or "0". The position of each array element is not identifiable in an "absolute" sense, and so a scheme of guidance marks helps the read/write head find positions on the disk. The need for these guidance markings explains why disks must be formatted before they can be used.



When it comes to accessing data already stored, the disk spins round very fast so that any part of its circumference can be quickly identified. The drive translates a read request from the computer into reality. There was a time when the cylinder/head/sector location that the computer worked out really was the data's location, but today's drives are more complicated than the BIOS can handle, and they translate BIOS requests by using their own mapping.

In the past it was also the case that a disk's controller did not have sufficient processing capacity to be able to read physically adjacent sectors quickly enough, thus requiring that the platter complete another full revolution before the next logical sector could be read. To combat this problem, older drives would stagger the way in which sectors were physically arranged, so as to reduce this waiting time. With an interleave factor of 3, for instance, two sectors would be skipped after each sector read. An interleave factor was expressed as a ratio, "N:1", where "N" represented the distance between one logical sector and the next. The speed of a modern hard disk drive with an integrated controller and its own data buffer renders the technique obsolete.

The rate at which hard disk capacities have increased over the years has given rise to a situation in which allocating and tracking individual data sectors on even a typical drive would require a huge amount of overhead, causing file handling efficiency to plummet. Therefore, to improve performance, data sectors have for some time been allocated in groups called clusters. The number of sectors in a cluster depends on the cluster size, which in turn depends on the partition size.

When the computer wants to read data, the operating system works out where the data is on the disk. To do this it first reads the FAT (File Allocation Table) at the beginning of the partition. This tells the operating system in which sector on which track to find the data. With this information, the head can then read the requested data. The disk controller controls the drive's servo-motors and translates the fluctuating voltages from the head into digital data for the CPU.

More often than not, the next set of data to be read is sequentially located on the disk. For this reason, hard drives contain between 256KB and 8MB of cache buffer in which to store all the information in a sector or cylinder in case it's needed. This is very effective in speeding up both throughput and access times. A hard drive also requires servo information, which provides a continuous update on the location of the heads. This can be stored on a separate platter, or it can be intermingled with the actual data on all the platters. A separate servo platter is more expensive, but it speeds up access times, since the data heads won't need to waste any time sending servo information.

However, the servo and data platters can get out of alignment due to changes in temperature. To prevent this, the drive constantly rechecks itself in a process called thermal recalibration. During multimedia playback this can cause sudden pauses in data transfer, resulting in stuttered audio and dropped video frames. Where the servo information is stored on the data platters, thermal recalibration isn't required. For this reason the majority of drives embed the servo information with the data.

Format:-

- **DIB File Format :-** Device-Independent Bitmap Format: a common bitmap format for Windows applications.
- **Format :-** A preparatory process that is necessary before data can be recorded to some storage devices. Formatting erases any previously stored data.
- **High Sierra Format :-** The standard logical file format for CD-ROM originally proposed by the High Sierra Group, revised and adopted by the International Standards Organisation as ISO 9660.
- **High-Level Formatting :-** Formatting performed by the operating system's format program (for example, the DOS FORMAT pro-gram). Among other things, the formatting program creates the root directory, file allocation tables, and other basic configurations. See also Low-Level Formatting.
- **ISO 9660 Format :-** An international standard specifying the logical format for files and directories on a CD-ROM. It provides a cross-platform format for storing filenames and directories which restricts the characters used to ensure all CD-ROM drives of all ages can read a data disc.The ISO 9660 data starts at track time 00:02:16 or sector 166 (logical sector 16) of track one. For a multisession disc the ISO 9660 data will be present in the first data track of each session containing CD-ROM tracks.
- **Low-Level Formatting :-** The process of creating sectors on the disk surface so that the operating system can access the required areas for generating the file structure. Also known as initialisation.
- **Physical Format :-** The actual physical layout of cylinders, tracks, and sectors on a disk drive.
- **Recording Format :-** The definition of how data is written to the tape. It defines such things as the number and position of tracks, bits per inch and the recording code to be used.

BIOS :- Basic Input Output System: a set of low-level routines in a computer's ROM that application programs (and operating systems) can use to read characters from the keyboard, output characters to printers, and interact with the hardware in other ways. It also provides the initial instructions for POST (Power On Self-Test) and booting the system files.

Interleave Factor :- Refers to a technique used by older hard disk drives to arrange sectors in a non-contiguous way so as to reduce rotational latency and thereby increase read/write performance. The interleave factor specifies the physical spacing between consecutive logical sectors.

Cluster :- A group of sectors on a hard disk drive that is addressed as one logical unit by the operating system.

Partition :- A portion of a hard disk accessible as a single logical volume, perhaps dedicated to a particular operating system or application.

FAT :- File Allocation Table: the file system used by DOS and Windows to manage files stored on hard disks, floppy disks, and other disk media. The file system takes its name from an on-disk data structure known as the file allocation table, which records where individual portions of each file are located on the disk. Earlier versions of Windows used the 16-bit version known as FAT16. Windows 98 has the option of using FAT32, which supports larger partition sizes and smaller cluster sizes, thereby improving disk performance and increasing available disk space. See also VFAT.

OS :- Operating System: the software controlling the overall operation of a multipurpose computer system, including such tasks as memory allocation, input and output distribution, interrupt processing, and job scheduling.

Servo Motor :- A closed-loop control system used to adjust head position and/or tape speed.

Look Ahead :- The technique of buffering data into cache RAM by reading subsequent blocks in advance to anticipate the next request for data. The look ahead technique speeds up disk access of sequential blocks of data.

BUFFER:-

- Back Buffer :- A hidden drawing buffer used in double-buffering. Graphics are drawn into the back buffer so that the rendering process cannot be seen by the user. When the drawing is complete, the front and back buffers are swapped.

- Buffer :- A storage location used for temporary storage of data read from or waiting to be sent to some device. Use of a memory buffer - often referred to as a "cache" - is used to speed up access to many devices, such as a hard disk, CD-ROM or tape drive.

- Cache Buffer :- An intermediate storage capacity between the processor and the disk drive used to store data likely to be requested next. Also known as Data Buffer. See also Look Ahead.

- Frame Buffer :- Display memory that temporarily stores (buffers) a full frame of picture data at one time.Frame buffers are composed of arrays of bit values that correspond to the display's pixels. The number of bits per pixel in the frame buffer determines the complexity of images that can be displayed.

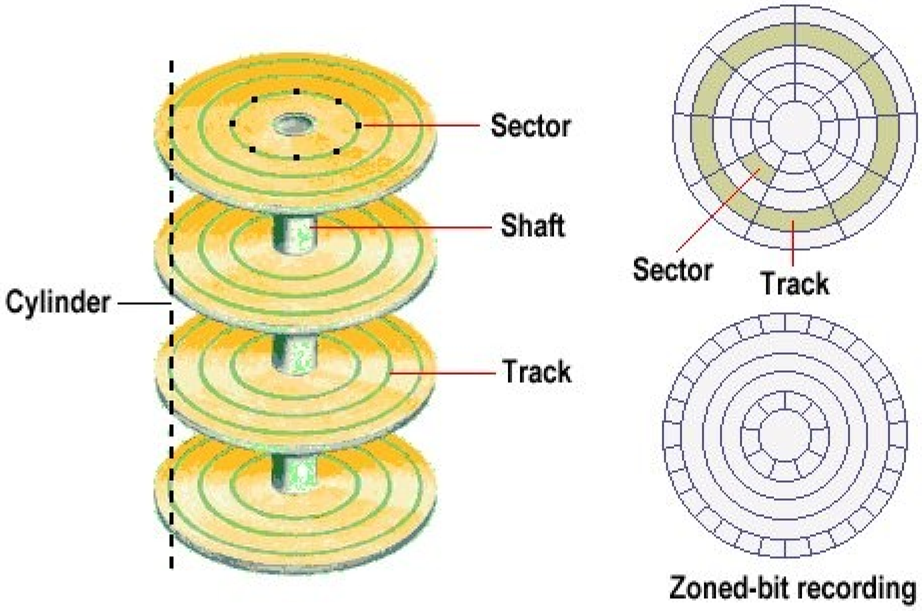
Embedded Servo :- The method most disks use to help the head locate tracks accurately; servo fields are interspersed with the real data, acting like runway lights for the head to line up on.

Servo Platter :- A separate surface containing only positioning and disk timing information but no data. Used only in a dedicated servo system.

Servo Data :- Magnetic markings written on the media that guide the read/write heads to the proper position.

4. Format

When a disk undergoes a low-level format, it is divided it into tracks and sectors. The tracks are concentric circles around the central spindle on either side of each platter. Tracks physically above each other on the platters are grouped together into cylinders which are then further subdivided into sectors of 512 bytes apiece.



The concept of cylinders is important, since cross-platter information in the same cylinder can be accessed without having to move the heads. The sector is a disk's smallest accessible unit. Drives use a technique called zoned-bit recording in which tracks on the outside of the disk contain more sectors than those on the inside.

TRACK :-

- **Bad Track Table:- A label affixed to the casing of a hard disk drive that tells which tracks are flawed and cannot hold data. The list is typed into the low-level formatting program when the drive is being installed.**
- **Off Track Retry :- Method of improving data recovery under error conditions. It is often possible to recover a data error by moving the head slightly off the track centre and re-reading the block.**
- **Track :- Sub-division of the recording area of storage media, such as magnetic disks, optical discs and magnetic tape.**

- **Track At Once:-** A writing mode that allows a session to be written in a number of discrete write events, called tracks. The mode mandates a minimum track length of 300 blocks (4 seconds), which equates to around 700KB, and a maximum of 99 tracks per disc. The disc may be removed from the writer and read in another writer before the session is fixated.

- **Track MultiSession :-** This write mode is very similar to Track At Once. In the Multisession environment, each "session" must contain at least one track. Again, the size of the track must be at least 300 blocks. Track Multisession mode allows tracks to be incrementally added to a disc (this should not to be confused with Incremental Writing). Each session will take up about 13.5MB of disc space in overhead; the so-called Lead-in and Lead-out areas.

Sector :-

- **Bad Sectors :-** Areas on the tape that cannot reliably retain data. This information is held in the tape header block.

- **Boot Sector :-** Reserved sectors on disk that are used to load the operating system. On start-up, the computer looks for the master book record (MBR), which is typically the first sector in the first partition of the disk. The MBR contains pointers to the first sector of the partition that contains the operating system, and that sector contains the instructions that cause the computer to "boot" the operating system (from the phrase "pulling yourself up from your bootstraps").

- **Sector :-** Describes the minimum segment of track length that can be assigned to store data. Magnetic disks are typically divided into tracks, each which contains a number of sectors. A sector contains a predetermined amount of data, such as 512 bytes. CDs can contain [(75 sectors per second) x (60 seconds per minute) x (number of minutes on disc)] sectors, the capacity of a sector depending on what physical format and mode is used for recording.

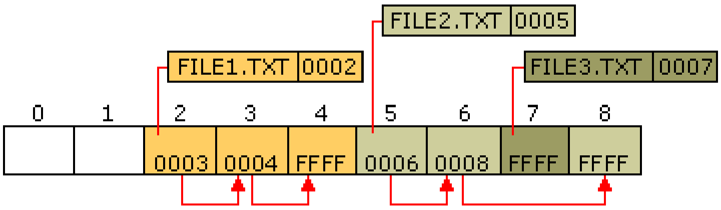
Cylinder :- When disks are placed directly above one another along the shaft, the circular, vertical "slice" consisting of all the tracks located in a particular position.

5. File systems

The precise manner in which data is organised on a hard disk drive is determined by the file system used. File systems are generally operating system dependent. However, since it is the most widely used PC operating system, most other operating systems' file systems are at least read-compatible with Microsoft Windows.

The FAT file system was first introduced in the days of MS-DOS way back in 1981. The purpose of the File Allocation Table is to provide the mapping between clusters - the basic unit of logical storage on a disk at the operating system level - and the physical location of data in terms of cylinders, tracks and sectors - the form of addressing used by the drive's hardware controller.

The FAT contains an entry for every file stored on the volume that contains the address of the file's starting cluster. Each cluster contains a pointer to the next cluster in the file, or an end-of-file indicator at (0xFFFF), which indicates that this cluster is the end of the file. The diagram shows three files: File1.txt uses three clusters, File2.txt is a fragmented file that requires three clusters and File3.txt fits in one cluster. In each case, the file allocation table entry points to the first cluster of the file.



The first incarnation of FAT was known as FAT12, which supported a maximum partition size of 8MB. This was superseded in 1984 by FAT16, which increased the maximum partition size to 2GB. FAT16 has undergone a number of minor modifications over the years, for example, enabling it to handle file names longer than the original limitation of 8.3 characters. FAT16's principal limitation is that it imposes a fixed maximum number of clusters per partition, meaning that the bigger the hard disk, the bigger the cluster size and the more unusable space on the drive. The biggest advantage of FAT16 is that it is compatible across a wide variety of operating systems, including Windows 95/98/Me, OS/2, Linux and some versions of UNIX.

Dating from the Windows 95 OEM Service Release 2 (OSR2), Windows has supported both FAT16 and FAT32. The latter is little more than an extension of the original FAT16 file system that provides for a much larger number of clusters per partition. As such, it offers greatly improved disk utilisation over FAT16. However, FAT32 shares all of the other limitations of FAT16 plus the additional one that many non-Windows operating systems that are FAT16-compatible will not work with FAT32. This makes FAT32 inappropriate for dual-boot environments, although while other operating systems such as Windows NT can't directly read a FAT32 partition, they can read it across the network. It's no problem, therefore, to share information stored on a FAT32 partition with other computers on a network that are running older versions of Windows.

With the advent of Windows XP in October 2001, support was extended to include the NTFS. NTFS is a completely different file system from FAT that was introduced with first version of Windows NT in 1993. Designed to address many of FAT's deficiencies, it provides for greatly increased privacy and security. The Home edition of Windows XP allows users to keep their information private to themselves, while the Professional version supports access control and encryption of individual files and folders. The file system is inherently more resilient than FAT, being less likely to suffer damage in the event of a system crash and it being more likely that any damage is recoverable via the chkdsk.exe utility. NTFS also journalises all file changes, so as to allow the system to be rolled back to an earlier, working state in the event of some catastrophic problem rendering the system inoperable.

FAT16, FAT32 and NTFS each use different cluster sizes depending on the size of the volume, and each file system has a maximum number of clusters it can support. The smaller the cluster size, the more efficiently a disk stores information because unused space within a cluster cannot be used by other files; the more clusters supported, the larger the volumes or partitions that can be created.

The table below provides a comparison of volume and default cluster sizes for the different Windows file systems still commonly in use:

Volume Size	FAT16 Cluster Size	FAT32 Cluster Size	NTFS Cluster Size
7MB - 16MB	2KB	Not supported	512 bytes
17MB - 32MB	512 bytes	Not supported	512 bytes
33MB - 64MB	1KB	512 bytes	512 bytes
65MB - 128MB	2KB	1KB	512 bytes
129MB - 256MB	4KB	2KB	512 bytes
257MB - 512MB	8KB	4KB	512 bytes
513MB - 1GB	16KB	4KB	1KB
1GB - 2GB	32KB	4KB	2KB
2GB - 4GB	64KB	4KB	4KB
4GB - 8GB	Not supported	4KB	4KB
8GB - 16GB	Not supported	8KB	4KB
16GB - 32GB	Not supported	16KB	4KB
32GB - 2TB	Not supported	Not supported	4KB

Volume:-

- **Volume :- A logical division of data, comprising of a number of files. In the context of hard disk drives, a volume is formatted by using a file system - such as FAT or NTFS - and has a drive letter assigned to it. A single hard disk can have multiple volumes and, unlike partitions, volumes can span multiple disks. Under the ISO 9660 standard, a "volume" refers to a single CD-ROM disc.**
- **Volume Descriptors :- In ISO 9660, a set of information on the disc containing vital information about the CD and how the computer should read it.**
- **Volume Label :- Data block written at the front of a volume to identify it.**

NTFS :- NT File System: the file system that is native to Microsoft Windows NT. NTFS is probably the most advanced file system available for personal computers, featuring superior performance, excellent security and crash protection, and the ability to handle large volumes of data.

6. Performance

The performance of a hard disk is very important to the overall speed of the system - a slow hard disk having the potential to hinder a fast processor like no other system component - and the effective speed of a hard disk is determined by a number of factors.

Chief among them is the rotational speed of the platters. Disk RPM is a critical component of hard drive performance because it directly impacts the latency and the disk transfer rate. The faster the disk spins, the more data passes under the magnetic heads that read the data; the slower the RPM, the higher the mechanical latencies. Hard drives only spin at one constant speed, and for some time most fast EIDE hard disks span at 5,400rpm, while a fast SCSI drive was capable of 7,200rpm. In 1997 Seagate pushed spin speed to a staggering 10,033rpm with the launch of its UltraSCSI Cheetah drive and, in mid 1998, was also the first manufacturer to release an EIDE hard disk with a spin rate of 7,200rpm.

In 1999 Hitachi broke the 10,000rpm barrier with the introduction of its Pegasus II SCSI drive. This spins at an amazing 12,000rpm - which translates into an average latency of 2.49ms. Hitachi has used an ingenious design to reduce the excessive heat produced by such a high spin rate. In a standard 3.5in hard disk, the physical disk platters have a 3in diameter. However, in the Pegasus II, the platter size has been reduced to 2.5in. The smaller platters cause less air friction and therefore reduce the amount of heat generated by the drive. In addition, the actual drive chassis is one big heat fin, which also helps dissipate the heat. The downside is that since the platters are smaller and have less data capacity, there are more of them and consequently the height of the drive is increased.

Mechanical latencies, measured in milliseconds, include both seek time and rotational latency. "Seek Time" is measured defines the amount of time it takes a hard drive's read/write head to find the physical location of a piece of data on the disk. "Latency" is the average time for the sector being accessed to rotate into position under a head, after a completed seek. It is easily calculated from the spindle speed, being the time for half a rotation. A drive's "average access time" is the interval between the time a request for data is made by the system and the time the data is available from the drive. Access time includes the actual seek time, rotational latency, and command processing overhead time.

The "disk transfer rate" (sometimes called media rate) is the speed at which data is transferred to and from the disk media (actual disk platter) and is a function of the recording frequency. It is generally described in megabytes per second (MBps). Modern hard disks have an increasing range of disk transfer rates from the inner diameter to the outer diameter of the disk. This is called a "zoned" recording technique. The key media recording parameters relating to density per platter are Tracks Per Inch (TPI) and Bits Per Inch (BPI). A track is a circular ring around the disk. TPI is the number of these tracks that can fit in a given area (inch). BPI defines how many bits can be written onto one inch of a track on a disk surface.

The "host transfer rate" is the speed at which the host computer can transfer data across the IDE/EIDE or SCSI interface to the CPU. It is more generally referred to as the data transfer rate, or DTR, and can be the source of some confusion. Some vendors list the internal transfer rate, the rate at which the disk moves data from the head to its internal buffers. Others cite the burst data transfer rate, the maximum transfer rate the disk can attain under ideal circumstances and for a short duration. More important for the real world is the external data transfer rate, or how fast the hard disk actually transfers data to a PC's main memory.

By late 2001 the fastest high-performance drives were capable of an average latency of less than 3ms, an average seek time of between 4 and 7ms and maximum data transfer rates in the region of 50 and 60MBps for EIDE and SCSI-based drives respectively. Note the degree to which these maximum DTRs are below the bandwidths of the current versions of the drive's interfaces - Ultra ATA/100 and UltraSCSI 160 - which are rated at 100MBps and 160MBps respectively.

RPM :- Revolutions Per Minute.

Average Seek Time:- he average time it takes for the read/write head to move to a specific location. To compute the average seek time, divide the time it takes to complete a large number of random seeks by the number of seeks performed

Seek Time :- The time taken for the actuator to move the heads to the correct cylinder in order to access data.

Latency :- The time between initiating a request for data and the beginning of the actual data transfer. For example, the average latency of a hard disk drive is easily calculated from the spindle speed, as the time for half a rotation. In communications, network latency is the delay introduced when a packet is momentarily stored, analysed and then forwarded.

Access Time :- Time interval between the instant that a piece of information is requested from a memory or peripheral device and the instant the information is supplied by the device. Access time includes the actual seek time, rotational latency, and command processing overhead time

MBps :- Megabytes per second: a performance measure used for mass storage devices and memory systems.

TPI :- Tracks Per Inch: the number of tracks written within each inch of a storage medium's recording surface. In the context of hard disk drives, EIDE drives generally have a higher TPI than SCSI drives. Also referred to as Track Density.

BPI :- Bits Per Inch: a measure of how densely information is packed on a storage medium. See also Flux Density.

Bus Master IDE :- Capability of the PIIX element of Triton chipset to effect data transfers from disk to memory with minimum intervention by the CPU, saving its horsepower for other tasks.

IDE :- Integrated Device Electronics or Intelligent Drive Electronics: a drive-interface specification for small to medium-size hard disks (disks with capacities up to 504Mb) in which all the drive's control electronics are part of the drive itself, rather than on a separate adapter connecting the drive to the expansion bus. This high level of integration shortens the signal paths between drives and controllers, permitting higher data transfer rates and simplifying adapter cards. See also EIDE and SCSI.

DTR :- Data Transfer Rate: the speed at which data is transferred between a host and a data recording device. Usually noted in KBps or MBps, and sometimes in MB/minute. Can mean a "peak" rather than a "sustained" transfer rate.

7. AV capability

Audio-visual applications require different performance characteristics than are required of a hard disk drive used for regular, everyday computer use. Typical computer usage involves many requests for relatively small amounts of data. By contrast, AV applications - digital audio recording, video editing and streaming, CD writing, etc. - involve large block transfers of sequentially stored data. Their prime requirement is for a steady, uninterrupted stream of data, so that any "dropout" in the analogue output is avoided.

In the past this meant the need for specially designed, or at the very least suitably optimised, hard disk drives. However, with the progressive increase in the bandwidth of both the EIDE and SCSI interfaces over the years, the need for special AV rated drives has become less and less. Indeed, Micropolis - a company that specialised in AV drives - went out of business as long ago as 1997.

The principal characteristic of an " AV drive" centred on the way that it handled thermal recalibration. As a hard drive operates, the temperature inside the drive rises causing the disk platters to expand (as most materials do when they heat up). In order to compensate for this phenomenon, hard drives would periodically recalibrate themselves to ensure the read and write heads remain perfectly aligned over the data tracks. Thermal recalibration (also known as "T-cal") is a method of re-aligning the read/write heads, and whilst it is happening, no data can be read from or written to the drive.

In the past, non-AV drives entered a calibration cycle on a regular schedule regardless of what the computer and the drive happened to be doing. Drives rated as "AV" have employed a number of different techniques to address the problem. Many handled T-cal by rescheduling or postponing it until such time that the drive is not actively capturing data. Some additionally used particularly large cache buffers or caching schemes that were optimised specifically and exclusively for AV applications, incurring a significant performance loss in non-AV applications.

By the start of the new millennium the universal adoption of embedded servo technology by hard disk manufacturers meant that thermal recalibration was no longer an issue. This effectively weaves head-positioning information amongst the data on discs, enabling drive heads to continuously monitor and adjust their position relative to the embedded reference points. The disruptive need for a drive to briefly pause data transfer to correctly position its heads during thermal recalibration routines is thereby completely eliminated.

AV Drive :- Audio Video drive: a hard disk drive that is optimised for audio and video applications. Transferring analogue high-fidelity audio and video signals onto a digital disk and playing them back at high performance levels requires a drive that can sustain continuous reads and writes without interruption. AV drives are designed to avoid thermal recalibration during reading and writing so that lengthy transfers digital video data will not be interrupted, and frames will not be lost.

Thermal Recalibration :- The periodic sensing of the temperature in hard disk drives so as to make minor adjustments to the alignment servo and data platters. In an AV drive, this process is performed only in idle periods so that there is no interruption in reading and writing long streams of digital video data.

8. Capacity

Since its advent in 1955, the magnetic recording industry has constantly and dramatically increased the performance and capacity of hard disk drives to meet the computer industry's insatiable demand for more and better storage. The areal density storage capacity of hard drives has increased at a historic rate of roughly 27% per year - peaking in the 1990s to as much as 60% per year - with the result that by the end of the millennium disk drives were capable of storing information in the 600-700 Mbits/in^2 range.

The read-write head technology that has sustained the hard disk drive industry through much of this period is based on the inductive voltage produced when a permanent magnet (the disk) moves past a wire-wrapped magnetic core (the head). Early recording heads were fabricated by wrapping wire around a laminated iron core analogous to the horseshoe-shaped electromagnets found in elementary school physics classes. Market acceptance of hard drives, coupled with increasing areal density requirements, fuelled a steady progression of inductive recording head advances. This progression culminated in advanced thin-film inductive (TFI) read-write heads capable of being fabricated in the necessary high volumes using semiconductor-style processors.

Although it was conceived in the 1960s, it was not until the late 1970s that TFI technology was actually deployed in commercially available product. The TFI read/write head - which essentially consists of wired, wrapped magnetic cores which produce a voltage when moved past a magnetic hard disk platter - went on to become the industry standard until the mid-1990s. By this time it became impractical to increase areal density in the conventional way - by increasing the sensitivity of the head to magnetic flux changes by adding turns to the TFI head's coil - because this increased the head's inductance to levels that limited its ability to write data.

The solution lay in the phenomenon discovered by Lord Kelvin in 1857 - that the resistance of ferromagnetic alloy changes as a function of an applied magnetic field - known as the anisotropic magnetoresistance (AMR) effect.

TFI :- Thin film inductive heads use a minute coil deposited onto a thin film using the photo-etching techniques employed to create integrated circuits; as the magnetic flux of the platter cuts the coil it induces a detectable current.

9. Capacity barriers

Whilst Bill Gates' assertion that "640KB ought to be enough for anyone" is the most famous example of lack of foresight when it comes to predicting capacity requirements, it is merely symptomatic of a trait that has afflicted the PC industry since its beginnings in the early 1980s. In the field of hard disk technology at least 10 different capacity barriers occurred in the space of 15 years. Several have been the result of BIOS or operating system issues, a consequence of either short-sighted design, restrictions imposed by file systems of the day or simply as a result of bugs in hardware or software implementations. Others have been caused by limitations in the associated hard disk drive standards themselves.

IDE hard drives identify themselves to the system BIOS by the number of cylinders, heads and sectors per track. This information is then stored in the CMOS. Sectors are always 512 bytes in size. Therefore, the capacity of a drive can be determined by multiplying the number of cylinders by the number of sectors by 512. The BIOS interface allows for a maximum of 1024 cylinders, 255 heads and 63 sectors. This calculates out at 504MiB. The IEC's binary megabyte notation was intended to address the confusion caused by the fact that this capacity is referred to as 528MB by drive manufacturers, who consider a megabyte to be 1,000,000 bytes instead of the binary programming standard of 1,048,576 bytes.

The 528MB barrier was the most infamous of all the hard disk capacity restrictions and primarily affected PCs with BIOSes created before mid-1994. It arose because of the restriction of the number of addressable cylinders to 1,024. Its removal - which led to the "E" (for Enhanced) being added to the IDE specification - by abandoning the cylinders, heads and sectors (CHS) addressing technique in favour of logical block addressing, or LBA. This is also referred to as the BIOS Int13h extensions. With this system the BIOS translates the cylinder, head and sector (CHS) information into a 28-bit logical block address, allowing operating systems and applications to access much larger drives.

Unfortunately, the designers of the system BIOS and the ATA interface did not set up the total bytes used for addressing in the same manner, nor did they define the same number of bytes for the cylinder, head, and sector addressing. The differences in the CHS configurations required that there be a translation of the address when data was sent from the system (using the system BIOS) and the ATA interface. The result was that the introduction of LBA did not immediately solve the problem of the 528MB barrier and also gave rise to a further restriction at 8.4GB.

The 8.4GB barrier involved the total addressing space that was defined for the system BIOS. Prior to 1997 most PC systems were limited to accessing drives with a capacity of 8.4GB or less. The reason for this was that although the ATA interface used 28-bit addressing which supported drive capacities up to 2**28 x 512 bytes or 137GB, the BIOS Int13h standard imposed a restriction of 24-bit addressing, thereby limiting access to a maximum of only 2**24 x 512 bytes or 8.4GB.

The solution to the 8.4GB barrier was an enhancement of the Int13h standard by what is referred to as Int13h extensions. This allows for a quad-word or 64 bits of addressing, which is equal to 2**64 x 512 bytes or 9.4 x 10**21 bytes. That is 9.4 Tera Gigabytes or over a trillion times as large as an 8.4GB drive. It was not until after mid-1998 that systems were being built that properly supported the BIOS Int13h extensions.

By the beginning of the new millennium, and much to the embarrassment of the drive and BIOS manufacturers, the 137GB limit imposed by the ATA interface's 28-bit addressing was itself beginning to look rather restrictive. However - better late than never - it appears as though the standards bodies may have finally learnt from their previous mistakes. The next version of the EIDE protocol (ATA-6) - being reviewed by the ANSI committee in the autumn of 2001 - allows for 48 bits of address space, giving a maximum addressable limit of 144PB (Petabytes). That's 100,000 times higher than the current barrier and, on previous form, sufficient for the next 20 years at least!.

CMOS RAM :- Complementary Metal Oxide Semiconductor Random Access Memory: a bank of memory that stores a PC's permanent configuration information, including type identifiers for the drives installed in the PC, and the amount of RAM present. It also maintains the correct date, time and hard drive information for the system

CMOS :- Complementary Metal Oxide Semiconductor: a process that uses both N- and P-channel devices in a complimentary fashion to achieve small geometries and low power consumption.

MiB :- Mebibyte: a unit of measure consisting of 1024KiB.

IEC :- nternational Electrotechnical Commission: the body that attempted to resolve the confusion surrounding the use of "MB" to mean either binary megabytes and decimal megabytes - depending on context - by their approval, in late 1998, of names and symbols for prefixes for binary multiples for use in the fields of data processing and data transmission. The acceptance of the symbols - "Ki", "Mi", "Gi" etc. - by the PC industry has been somewhat disappointing.

ATA :- AT Attachment: the specification, formulated in the 1980s by a consortium of hardware and software manufacturers, that defines the IDE drive

interface. AT refers to the IBM PC/AT personal computer and its bus architecture. IDE drives are sometimes referred to as ATA drives or AT bus drives. The newer ATA-2 specification defines the EIDE interface, which improves upon the IDE standard. See also IDE and EIDE.

ANSI :- American National Standards Institute. A standards-setting, non-government organisation which develops and publishes standards for voluntary use in the United States.

10. MR technology

In 1991, IBM's work on AMR technology led to the development of MR (magnetoresistive) heads capable of the areal densities required to sustain the disk drive industry's continued growth in capacity and performance. These circumvented the fundamental limitation of TFI heads - fact that their recording had alternately to perform conflicting task writing data on as well retrieving previously-written by adopting a design which read write elements were separate, allowing each be optimised for its specific function.

In an MR head, the write element is a conventional TFI head, while the read element is composed of a thin stripe of magnetic material. The stripe's resistance changes in the presence of a magnetic field, producing a strong signal with low noise amplification and permitting significant increases in areal densities. As the disk passes by the read element, the disk drive circuitry senses and decodes changes in electrical resistance caused by the reversing magnetic polarities. The MR read element's greater sensitivity provides a higher signal output per unit of recording track width on the disk surface. Not only does magnetoresistive technology permit more data to be placed on disks, but it also uses fewer components than other head technologies to achieve a given capacity point.

The MR read element is smaller than the TFI write element. In fact, the MR read element can be made smaller than the data track so that if the head were slightly off-track or misaligned, it would still remain over the track and able to read the written data on the track. Its small element size also precludes the MR read element from picking up interference from outside the data track, which accounts for the MR head's desirable high signal-to-noise ratio.

Manufacturing MR heads can present difficulties. MR thin film elements are extremely sensitive to electrostatic discharge, which means special care and precautions must be taken when handling these heads. They are also sensitive to contamination and, because of the materials used in its design, subject to corrosion.

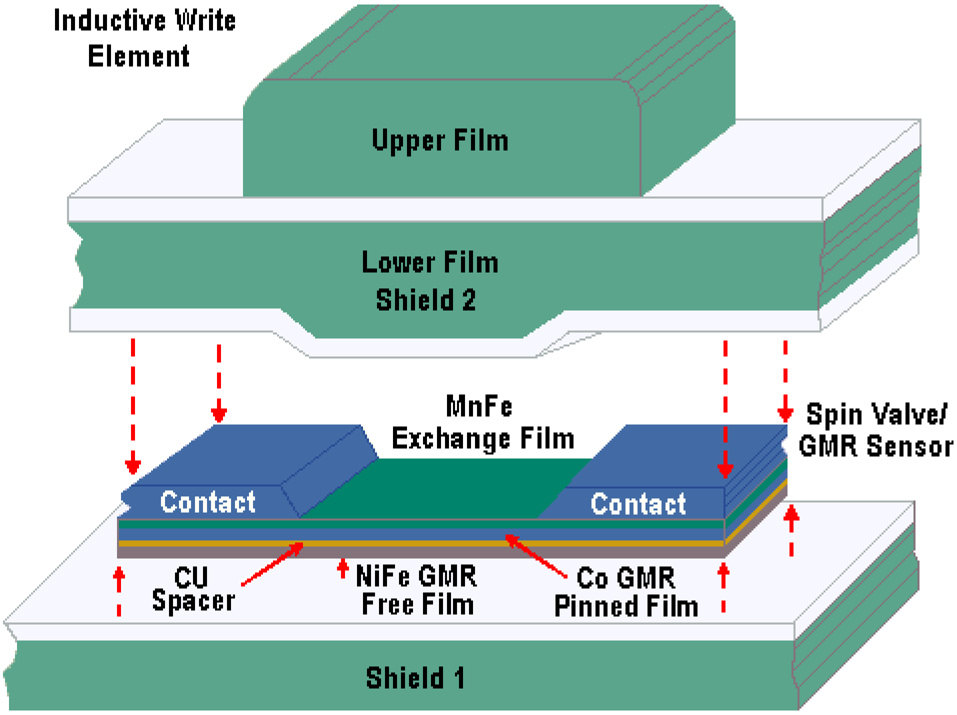
MR heads also introduced a new challenge not present with TFI heads: thermal asperities, the instantaneous temperature rise that causes the data signal to spike and momentarily disrupt the recovery of data from the drive. Thermal asperities are transient electrical events, usually associated with a particle, and normally do not result in mechanical damage to the head. Although they can lead to misreading data in a large portion of a sector, new design features can detect these events. A thermal asperity detector determines when the read input signal exceeds a predetermined threshold, discounts that data value and signals the controller to re-read the sector.

The various improvements offered by MR technology amount to an ability to read from areal densities about four times denser than TFI heads at higher flying heights. In practice this means that the technology is capable of supporting areal densities of at least 3 Gbits/in^2. The technology's sensitivity limitations stem from the fact that the degree of change in resistance in an MR head's magnetic film is itself limited. It wasn't long before a logical progression from MR technology was under development, in the shape of Giant Magneto-Resistive (GMR) technology.

11. GMR technology

Giant Magneto-Resistive (GMR) head technology builds on existing read/write technology found in TFI and anisotropic MR, producing heads that exhibit a higher sensitivity to changing magnetisation on the disc and work on spin-dependent electron scattering. The technology is capable of providing the unprecedented data densities and transfer rates necessary to keep up with the advances in processor clock speeds, combining quantum mechanics and precision manufacturing to give areal densities that are expected to reach 10Gbits/in^2 and 40Gbits/in^2 by the years 2001 and 2004 respectively.

In MR material, e.g. nickel-iron alloys, conduction electrons move less freely (more frequent collisions with atoms) when their direction of movement is parallel to the magnetic orientation in the material. This is the "MR effect", discovered in 1988. When electrons move less freely in a material, the material's resistance is higher. GMR sensors exploit the quantum nature of electrons, which have two spin directions-spin up and spin down. Conduction electrons with spin direction parallel to a film's magnetic orientation move freely, producing low electrical resistance. Conversely, the movement of electrons of opposite spin direction is hampered by more frequent collisions with atoms in the film, producing higher resistance. IBM has developed structures, identified as spin valves, in which one magnetic film is pinned. This means its magnetic orientation is fixed. The second magnetic film, or sensor film, has a free, variable magnetic orientation. These films are very thin and very close together, allowing electrons of either spin direction to move back and forth between these films. Changes in the magnetic field originating from the disk cause a rotation of the sensor film's magnetic orientation, which in turn, increases or decreases resistance of the entire structure. Low resistance occurs when the sensor and pinned films are magnetically oriented in the same direction, since electrons with parallel spin direction move freely in both films.



Higher resistance occurs when the magnetic orientations of the sensor and pinned films oppose each other, because the movement of electrons of either spin direction is hampered by one or the other of these magnetic films. GMR sensors can operate at significantly higher areal densities than MR sensors, because their percent change in resistance is greater, making them more sensitive to magnetic fields from the disk.

Current GMR hard disks have storage densities of 4.1Gbit/in^2, although experimental GMR heads are already working at densities of 10Gbit/in^2. These heads have a sensor thickness of 0.04 microns, and IBM claims that halving the sensor thickness to 0.02 microns - with new sensor designs - will allow possible densities of 40Gbit/in^2. The advantage of higher recording densities is that disks can be reduced in physical size and power consumption, which in turn increases data transfer rates. With smaller disks for a given capacity, combined with lighter read/write heads, the spindle speed can be increased further and the mechanical delays caused by necessary head movement can be minimised.

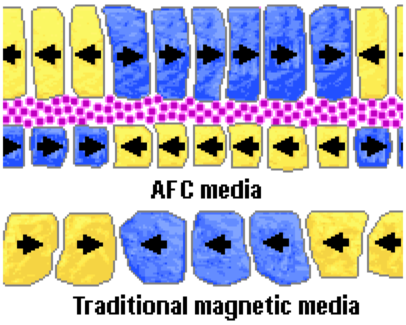
IBM has been manufacturing merged read/write heads which implement GMR technology since 1992. These comprise a thin film inductive write element and a read element. The read element consists of an MR or GMR sensor between two magnetic shields. The magnetic shields greatly reduce unwanted magnetic fields coming from the disk; the MR or GMR sensor essentially "sees" only the magnetic field from the recorded data bit to be read. In a merged head the second magnetic shield also functions as one pole of the inductive write head. The advantage of separate read and write elements is both elements can be individually optimised. A merged head has additional advantages. This head is less expensive to produce, because it requires fewer process steps; and, it performs better in a drive, because the distance between the read and write elements is less.

Areal Density :- The amount of data that's stored on a hard disk per square inch, and equal to the tracks per inch multiplied by the bits per inch along each track. In the context of tape storage, the number of flux transitions per square unit of recordable area, or bits per square inch.

12. "Pixie dust"

In the past decade, the data density for magnetic hard disk drives has increased at a phenomenal pace, doubling every 18 months and - since 1997 - doubling every year. The situation had left researchers fearing that hard drive capacities were getting close to their limit. When magnetic regions on the disk become too small, they cannot retain their magnetic orientations and thus their data - over the typical lifetime of the product. This is called the "superparamagnetic effect", and has long been predicted to appear when densities reached 20 to 40 billion bits (gigabits) per square inch, not a huge way away from the densities that had been reached by the start of the new millennium.

However, in the summer of 2001, IBM announced a breakthrough in storage technology that could prolong the life of the conventional hard disk drive for the foreseeable future. The key to the breakthrough is a three-atom-thick layer of the element ruthenium, a precious metal similar to platinum, sandwiched between two magnetic layers. That only a few atoms could have such a dramatic impact has resulted in scientists referring to the ruthenium layer informally as "pixie dust".

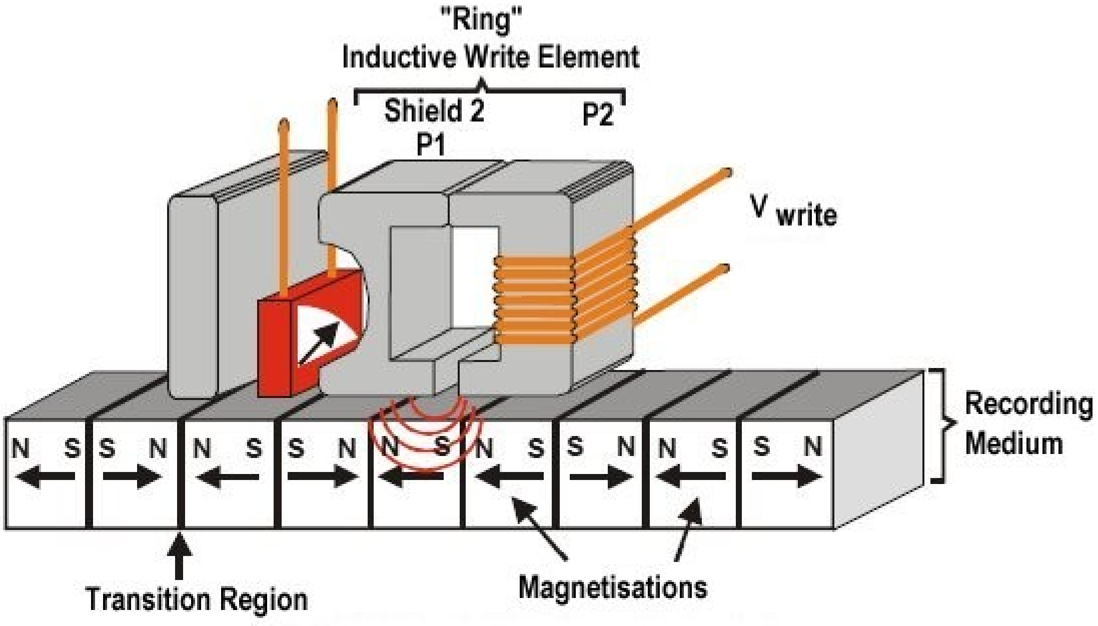


Known technically as "antiferromagnetically-coupled (AFC) media", the new multilayer coating is expected to permit hard disk drives to store 100 billion bits (gigabits) of data per square inch of disk area by 2003 and represents the first fundamental change in disk drive design made to avoid the high-density data decay due to the superparamagnetic effect. Conventional disk media stores data on a single magnetic layer. AFC media's two magnetic layers are separated by an ultra-thin layer of ruthenium. This forces the adjacent layers to orient themselves magnetically in opposite directions. The opposing magnetic orientations make the entire multilayer structure appear much thinner than it actually is. Thus small, high-density bits can be written easily on AFC media, but they will retain their magnetisation due to the media's overall thickness.

As a consequence, the technology is expected to allow densities of 100 gigabits per square inch and beyond.

13. Longitudinal recording

For nearly 50 years, the disk drive industry has focused nearly exclusively on a method called longitudinal magnetic recording, in which the magnetisation of each data bit is aligned horizontally in relation to the drive's spinning platter. In longitudinal recording, the fields between two adjacent bits with opposing magnetisations are separated by a transition region.



© 2005, Hitachi Global Storage Technologies

While areal densities have historically pretty much doubled every year,more recently however, the rate of increase has slowed, and the limit of areal densities for hard disk technology using longitudinal recording - due to the superparamagnetic effect - is now generally reckoned to be in the region of 100 to 200 Gbits/in^2.

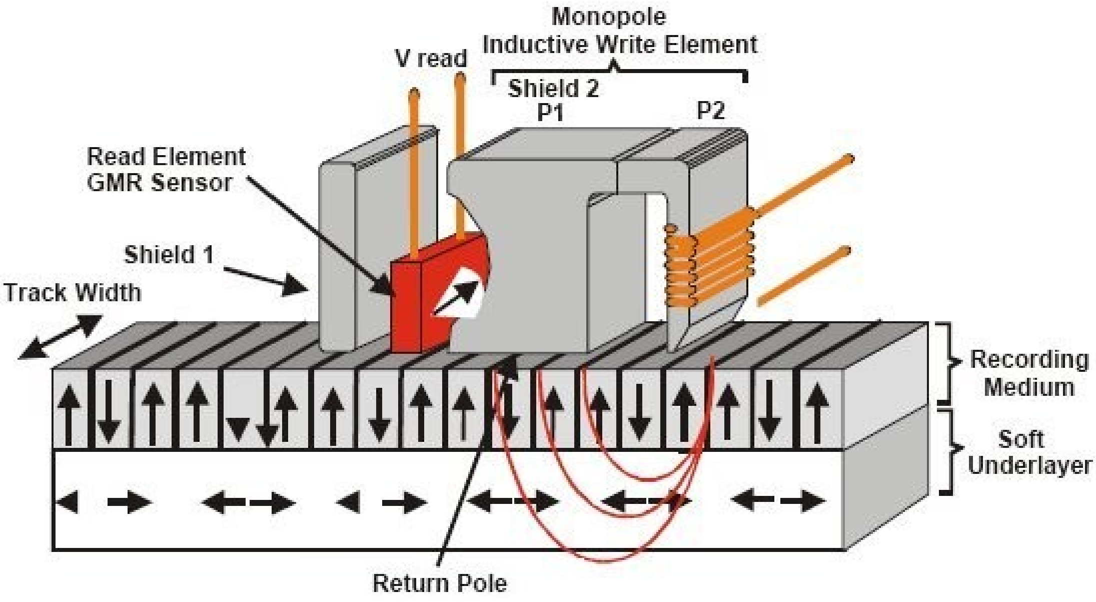
This has led to more aggressive approaches to delaying the superparamagnetic effect being sought, and the emergence of perpendicular magnetic recording (PMR) technology as a solution capable of doubling hard disk data densities in the relative short term and of ultimately enabling a more 10-fold increase in storage capacity over today's technology.

Superparamagnetic effect :- The point at which discrete magnetic areas of a disk's surface are so tiny that their magnetic orientation is unstable at room temperature.

14. Perpendicular recording

In perpendicular recording, the magnetisation of the bit is aligned vertically - or perpendicularly - in relation to the disk drive's platter. Since the bits do not directly oppose each other, the need for transition packing is significantly reduced. This allows bits to be more closely packed with sharper transition signals, facilitating easier bit detection and error correction. The potential for higher areal density results.

To help understand how perpendicular recording works, consider the bits as small bar magnets. In conventional longitudinal recording, the magnets representing the bits are lined up end-to-end along circular tracks in the plane of the disk. If you consider the highest-density bit pattern of alternating ones and zeros, then the adjacent magnets end up head-to-head (north-pole to north pole) and tail-to-tail (south-pole to south-pole). In this scenario, they want to repel each other, making them unstable against thermal fluctuations. In perpendicular recording, the tiny magnets are standing up and down. Adjacent alternating bits stand with north pole next to south pole; thus, they want to attract each other and are more stable and can be packed more closely.



© 2005, Hitachi Global Storage Technologies

Key to being able to make the bits smaller, without superparamagnetism causing them to lose their memory, is the use of a magnetically "stronger" (higher coercivity) material as the storage medium. This is possible due to the fact that in a perpendicular arrangement the magnetic flux is guided through a magnetically soft (and relatively thick) underlayer underneath the hard magnetic media films (considerably complicating and thickening the total disk structure). This magnetically soft underlayer can be effectively considered a part of the write head, making the write head more efficient, thus making it possible to produce a stronger write field gradient with essentially the same head materials as for longitudinal heads, and therefore allowing for the use of the higher coercivity - and therefore inherently more thermally stable - magnetic storage medium.

By 2005 all of the major hard disk manufacturers were working on PMR technology and the technology was expected to be broadly adopted in products by the end of 2007. Hitachi Global Storage Technologies demonstrated an areal density of 345 Gbits/in^2 in laboratory testing in the autumn of that year, expected to bring hard drive areal density half way to the 345 Gbits/in^2 mark with a 1 TB 3.5in product in the first half of 2007, and was predicting a 2 TB 3.5in desktop drive by 2009.

Extensions to PMR technology are expected to take hard drive advancements out beyond the next two decades, and eventually allow information densities of up to 100 Tbits/in^2.

TB :- Terabyte: a unit of measure for storage capacity 1,000,000,000,000 (1 trillion) bytes, or 1,000,000 (1 million) megabytes or 1,000 (1 thousand) gigabytes.

15. RAID

In the 1980s, hard-disk drive capacities were limited and large drives commanded a premium price. As an alternative to costly, high-capacity individual drives, storage system developers began experimenting with arrays of smaller, less expensive hard-disk drives. In a 1988 publication, "A Case for Redundant Arrays of Inexpensive Disks", three University of California-Berkeley researchers proposed guidelines for these arrays. They originated the term RAID - redundant array of inexpensive disks - to reflect the data accessibility and cost advantages that properly implemented arrays could provide. As storage technology has advanced and the cost per megabyte of storage has decreased, the term RAID has been redefined to refer to "independent" disks, emphasising the technique's potential data *availability* advantages relative to conventional disk storage systems.

The original concept was to cluster small inexpensive disk drives into an array such that the array could appear to the system as a single large expensive drive (SLED). Such an array was found to have better performance characteristics than a traditional individual hard drive. The initial problem, however, was that the Mean Time Before Failure (MTBF) of the array was reduced due to the probability of any one drive of the array failing. Subsequent development resulted in the specification of six standardised RAID levels to provide a balance of performance and data protection. In fact, the term "level" is somewhat misleading because these models do not represent a hierarchy; a RAID 5 array is not inherently better or worse than a RAID 1 array. The most commonly implemented RAID levels are 0, 3 and 5:

- Level 0 provides "data striping" (spreading out blocks of each file across multiple disks) but no redundancy. This improves performance but does not deliver fault tolerance. The collection of drives in a RAID Level 0 array has data laid down in such a way that it is organised in stripes across the multiple drives, enabling data to be accessed from multiple drives in parallel.
- Level 1 provides disk mirroring, a technique in which data is written to two duplicate disks simultaneously, so that if one of the disk drives fails the system can instantly switch to the other disk without any loss of data or service. RAID 1 enhances read performance, but the improved performance and fault tolerance are at the expense of available capacity in the drives used.
- Level 3 is the same as Level 0, but 0 sacrifices some capacity, for the same number of drives, to achieve a higher level of data

integrity or fault tolerance by reserving one dedicated disk for error correction data. This drive is used to store parity information that is used to maintain data integrity across all drives in the subsystem.

- Level 5 is probably the most frequently implemented. It provides data striping at the byte level and also stripe error correction information. This results in excellent performance coupled with the ability to recover any lost data should any single drive fail.

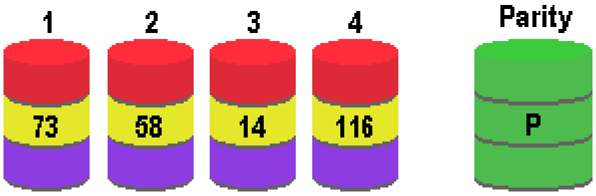
The data striping storage technique is fundamental to the concept and used by a majority of RAID levels. In fact, the most basic implementation of this technique, RAID 0, is not *true* RAID unless it is used in conjunction with other RAID levels since it has no inherent fault tolerance. Striping is a method of mapping data across the physical drives in an array to create a large virtual drive. The data is subdivided into consecutive segments or "stripes" that are written sequentially across the drives in the array, each stripe having a defined size or "depth" in blocks. A striped array of drives can offer improved performance compared to an individual drive if the stripe size is matched to the type of application program supported by the array:

- In an I/O-intensive or transactional environment where multiple concurrent requests for small data records occur, larger (block-level) stripes are preferable. If a stripe on an individual drive is large enough to contain an entire record, the drives in the array can respond independently to these simultaneous data requests.
- In a data-intensive environment where large data records are stored, smaller (byte-level) stripes are more appropriate. If a given data record extends across several drives in the array, the contents of the record can be read in parallel, improving the overall data transfer rate.

Extended Data Availability and Protection (EDAP) is another data storage concept, closely related to RAID. A storage system with EDAP capability can protect its data and provide on-line, immediate access to its data, despite failure occurrence within the disk system, within attached units or within its environment. The location, type and quantity of failure occurrences determine the degree of EDAP capability attributed to the disk system. Two types of RAID provide EDAP for disks: Mirroring and Parity RAID. Mirroring predated Parity RAID and was identified in the Berkeley Papers as RAID Level 1. Its disadvantage is that, unlike Parity RAID, Mirroring requires 100% redundancy. Its advantages, unlike Parity RAID, are that read performance is improved, the impact on write performance is generally modest and a higher percentage of disks in a Mirrored redundancy group may fail simultaneously as compared to a

Parity RAID redundancy group. Parity RAID is identified in the Berkeley Papers as RAID Levels 3, 4, 5 and 6. In these cases, overhead (redundant data in the form of Parity) as compared to Mirroring (redundant data in the form of a complete copy) is significantly reduced to a range of 10% to 33%.

Parity RAID levels combine striping and parity calculations to permit data recovery if a disk fails. The diagram illustrates the concepts of both data striping and Parity RAID, depicting how a block of data containing the values 73, 58, 14, and 126 may be striped across a RAID 3 array comprising four data drives and a parity drive, using the even-parity method.



Up until the late 1990s, the implementation of RAID had been almost exclusively in the server domain. By then, however, processor speeds had reached the point where the hard disk was often the bottleneck that prevented a system running at its full potential. Aided and abetted by the availability of motherboards that included a RAID controller - by 2000 the deployment of RAID's striping technique had emerged as a viable solution to this problem on high-end desktop systems.

RAID :- Redundant Array of Inexpensive Disks: when configured for performance a RAID writes and reads data in parallel from multiple drive simultaneously. In theory data can be moved at the speed of one drive multiplied by the number of drives working in parallel, although in practice management overheads reduce this significantly.

MTBF :- Mean Time Between Failure: the average time a specific component is expected to work without failure.

EDAP :- Extended Data Availability and Protection: Created by the RAID Advisory Board in 1997, EDAP introduces a classification system for the resilience of the entire storage system and that is not confined to disk-based storage alone. Availability of an EDAP-certified system is sustainable even in the event of failure, the degree of resiliency provided being reflected in the level of EDAP capability attributed to the system.

16. SMART

In 1992, IBM began shipping 3.5-inch hard disk drives that could actually predict their own failure - an industry first. These drives were equipped with Predictive Failure Analysis (PFA), an IBM-developed technology that periodically measures selected drive attributes - things like head-to-disk flying height - and sends a warning message when a predefined threshold is exceeded. Industry acceptance of PFA technology eventually led to SMART (Self-Monitoring, Analysis and Reporting Technology) becoming the industry-standard reliability prediction indicator for both IDE/ATA and SCSI hard disk drives.

There are two kinds of hard disk drive failures: unpredictable and predictable. Unpredictable failures happen quickly, without advance warning. These failures can be caused by static electricity, handling damage, or thermal-related solder problems, and there is nothing that can be done to predict or avoid them. In fact, 60% of drive failures are mechanical, often resulting from the gradual degradation of the drive's performance. The key vital areas include:

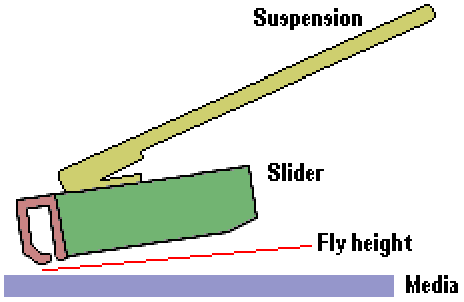
- Heads/head assembly: crack on head, broken head, head contamination, head resonance, bad connection to electronics module, handling damage
- Motors/bearings: motor failure, worn bearing, excessive run out, no spin, handling damage
- Electronic module: circuit/chip failure, bad connection to drive or bus, handling damage
- Media: scratch, defect, retries, bad servo, ECC corrections, handling damage.

These have been well explored over the years and have led to disk drive designers being able to not only develop more reliable products, but to also apply their knowledge to the prediction of device failures. Through research and monitoring of vital functions, performance thresholds which correlate to imminent failure have be determined, and it is these types of failure that SMART attempts to predict.

Just as hard disk drive architecture varies from one manufacturer to another, so SMART-capable drives use a variety of different techniques to monitor data availability. For example, a SMART drive might monitor the fly height of the head above the magnetic media. If the head starts to fly too high or too low, there's a good chance the drive could fail. Other drives may monitor additional or different conditions, such as ECC circuitry on the hard drive card or soft error rates. When impending failure is suspected the drives sends an alert through the operating system to an application that displays a warning message.

A head crash is one of the most catastrophic types of hard disk failure and - since the height at which a head flies above the surface of the media has decreased steadily over the years as one of the means to increase areal recording densities, and thereby disk storage capacities - it might reasonably be expected to be an increasingly likely form of failure. Fortunately, this is not the case, since flying height has always been one of the most critical parameters for disk drive reliability and as this has steadily decreased, so the techniques used to predict head crashes have become progressively more sophisticated. Not only are heads flying too low are in danger of crashing, but if the recording head flies higher than intended, even for a short period of time, the magnetic field available may be insufficient to reliably write to the media. This is referred to as a "high fly write". External shock, vibration, media defect or contamination may cause this. Soft errors caused by this phenomenon are recoverable, but hard errors are not.

The fly height is controlled by the suspension attached to the slider containing the magnetic recording head and the airbearing of the slider. This aerodynamic system controls the variation in fly height as the slider is positioned over the surface of the media. Traversing the head between the inner and outer radius of the disk causes a two-to-one change in velocity. Prior to current technology in airbearing designs, this change in velocity would have created a two-to-one change in nominal fly height. However, with current day airbearing designs, this variation can be reduced to a fraction of the nominal value and fly heights - the distance between the read/write elements and the magnetic surface - are typically of the order of a few millionths of an inch and as low as 1.2 micro-inches. There are several conditions - for example, altitude, temperature, and contamination - that can create disturbances between the airbearing and the disk surface and potentially change the fly height.



Thermal monitoring is a more recently introduced aspect of SMART, designed to alert the host to potential damage from the drive operating at too high a temperature. In a hard drive, both electronic and mechanical components - such as actuator bearings, spindle motor and voice coil motor - can be affected by excessive temperatures. Possible causes include a clogged cooling fan, a failed room air conditioner or a cooling system that is simply overextended by too many drives or other components. Many SMART implementations use a thermal sensor to detect the environmental conditions that affect drive reliability - including ambient temperature, rate of cooling airflow, voltage and vibration - and issue a user warning when the temperature exceeds a pre-defined threshold - typically in the range 60-65°C).

The table below identifies a number of other failure conditions, their typical symptoms and causes and the various factors whose monitoring can enable impending failure to be predicted:

Type of Failure	**Symptom/Cause**	**Predictor**
Excessive bad sectors	Growing defect list, media defects, handling damage	Number of defects, growth rate
Excessive run-out	Noisy bearings, motor, handling damage	Run-out, bias force diagnostics
Excessive soft errors	Crack/broken head, contamination	High retries, ECC involves
Motor failure, bearings	Drive not ready, no platter spin, handling damage	Spin-up retries, spin-up time
Drive not responding, no connect	Bad electronics module	None, typically catastrophic
Bad servo positioning	High servo errors, handling damage	Seek errors, calibration retries
Head failure, resonance	High soft errors, servo retries, handling damage	Read error rate, servo error rate

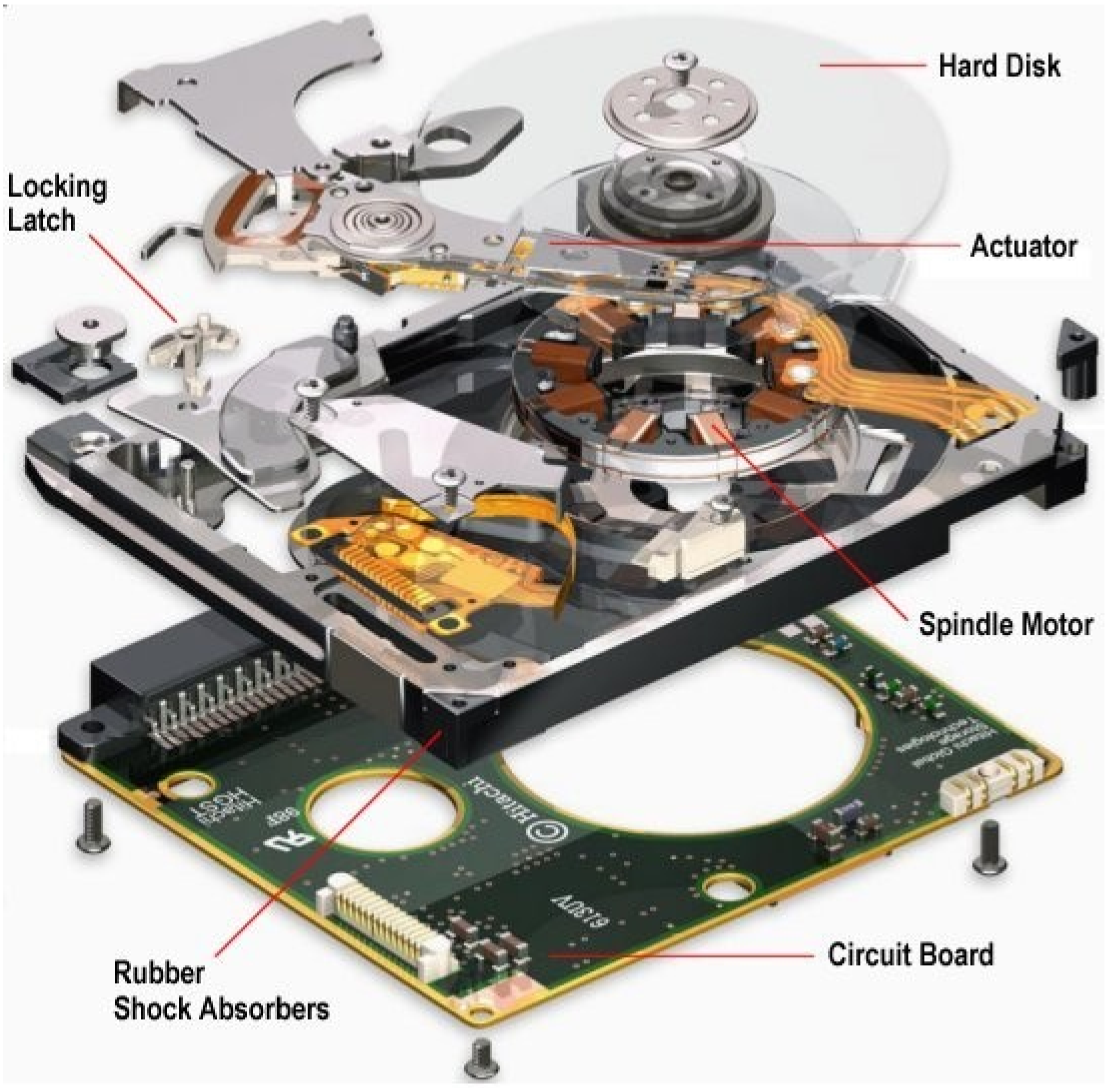
In its brief history, SMART technology has progressed through three distinct iterations. In its original incarnation SMART provided failure prediction by monitoring certain online hard drive activities. A subsequent version improved failure prediction by adding an automatic off-line read scan to monitor additional operations. The latest SMART technology not only monitors hard drive activities but adds failure prevention by attempting to detect and repair sector errors. Also, whilst earlier versions of the technology only monitored hard drive activity for data that was retrieved by the operating system, this latest SMART tests all data and all sectors of a drive by using "off-line data collection" to confirm the drive's health during periods of inactivity.

ECC :- Error Correction Code: a system of scrambling data and recording redundant data in stored data in order to enable the detection of errors that can be corrected by the device's controller when the data is read. ECC memory can detect up to 4-bit memory errors; only single-bit errors, however, can be corrected. See also CRC.

ECC optimized :- On a SIMM or DIMM, the use of a module addressing architecture that facilitates the use of the memory module by systems with ECC. ECC optimised memory modules do not have byte-write capability.

17. Microdrive

GMR technology was a key feature of IBM's (the company's hard disk drive business was subsequently acquired by Hitachi Global Storage Technologies in 2002) revolutionary Microdrive device, launched in mid-1999. The world’s smallest, lightest one-inch hard drive was invented at IBM’s San Jose Research Center by a team of engineers engaged in the study of micromechanical applications such as Si micro-machined motors and actuators for possible usage in very small disk drives.



© 2004 Hitachi Global Storage Technologies

The Microdrive uses a single one-inch diameter glass platter with a metal coating which is less than a thousandth of the thickness of a human hair. Powered by nine elecromagnets, the spindle motor spins the 16g disk at 4,500rpm, data being transmitted and received through the actuator's miscroscopic read-write heads as it sweeps across the disk. Rubber shock absorbers and an actuator locking latch prevent damage to the disk's surface, both under normal operating conditions and if the unit is jarred or dropped. The drive's circuit board acts as the drive's brain, controlling its functions from speed to dataflow.

The tiny elements inside the Microdrive confer some unique advantages. For example, since the actuator has 50 times less inertia than one in a larger drive, it can ramp up to full speed in half a second. Consequently, its possible to allow the drive to stop spinning when data is not being accessed - thereby improving the device's power conservation characteristics.

Use of the industry-standard CF+ Type II interface - allowing easy integration into a variety of handheld products and providing compatibility with PCMCIA Type II interface via the use of an adapter - took CompactFlash storage into completely new territory, enabling high-capacity yet cost-effective personal storage for a wide variety of electronic devices.

The Microdrive was initially released with capacities of 170MB and 340MB, with a claimed seek time of 15ms, average latency of 6.7ms and a data transfer rate of between 32 Mbit/s and 49 Mbit/s. In 2000, 512MB and 1GB versions were released and capacities continued to increase, reaching 8GB by 2005. By that time, other manufacturers had entered the market, with Seagate having launched an 8GB 1in drive in the same year and Sony licensing re-badged Hitachi-made models under the brand name "Sony Microdrive".

Hitachi expects to see the recent advancements in Perpendicular Magnetic Recording (PMR) technology translate into storage capacities of up to 20GB on a one-inch Microdrive by the year 2007.

Microdrive :- An ultra-miniature hard disk technology from IBM that uses a single one-inch diameter platter to provide either 170MB or 340MB storage capacity and either one or two GMR heads, the Microdrive is built into a Type II CompactFlash form factor.

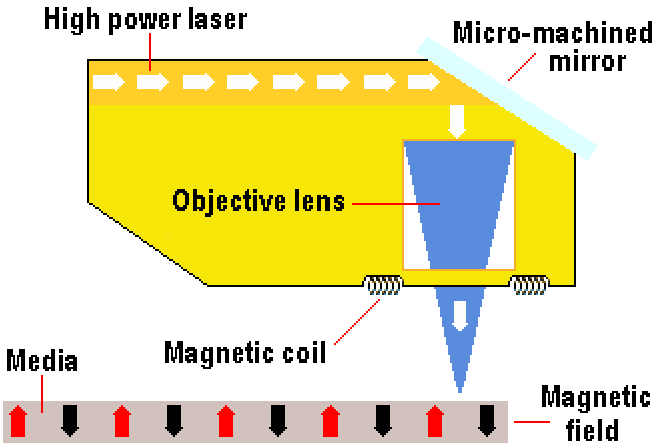
PCMCIA :- Personal Computer Memory Card International Association: a consortium of computer manufacturers that devised the standard for the credit card-size adapter cards used in many notebook computers. PCMCIA defines three card types: Type I cards can be up to 3.3mm thick and are generally used for RAM and ROM expansion cards; Type II cards can be as thick as 5.5mm and typically house modems and fax modems; Type III cards are the largest of the lot (up to 10.5mm thick) and are mostly used for solid state disks or miniature hard disks. PCMCIA cards are also known as "PC Cards".

PMR :- Perpendicular Magnetic Recording; a technique for writing bits to magnetic media which allows for more room in which to pack data, thus enabling higher recording or "areal" densities.

OAW technology

While GMR technology is looking to areal densities of up to 40Gbit/in^2 in the next few years, some hard disk manufacturers anticipate that the phenomenon of losing data due to data bits being packed too close together will begin to happen when drives begin holding around 20Gbit/in^2. The Seagate subsidiary, Quinta Corporation, is planning to combat this storage technology barrier, otherwise known as the Superparamagnetic Limit, with its Optically Assisted Winchester (OAW) technology, expected to appear sometime during 1999.

OAW has as much in common with magneto-optical technology as conventional magnetic hard drives, but uses a clever bundle of techniques to get around the size and performance drawbacks of MO formats. The technique uses laser light to select various heads within the drive, rather than the typical electromagnetic system in conventional drives. The principle is simple: a laser is focused on the surface of a platter and can be used for reading and writing. The former relies on the Kerr Effect, whereby polarised light bounced off a magnetised surface has its polarisation twisted. Put that through a further polarising filter and the intensity of the light corresponds to the alignment of the illuminated domain.



This method of reading needs less power from the laser, so the heating effect on the medium is minimal, preventing data corruption. The same laser and optics can be used for writing, using a magneto-optical technique. A tiny spot on the hard disk is heated with a higher-power output laser to beyond the temperature called the Curie Point, above which the surface material's magnetic properties can be flipped with a magnetic coil. This changes the light polarisation characteristics. Unlike regular magneto-optics, however, OAW heats the media and writes to it in one pass, rather than heating on one rotation and writing on the next.

OAW is able to do this as it uses a micromachined servo mirror and a minute objective lens to focus the laser very accurately on the smallest area possible. Adjacent areas aren't heated up and are therefore left unchanged.

Since the laser used to heat the media can be focused on a much smaller area than a magnet, higher data densities are possible. The media itself is made of an amorphous alloy that does not have a granular structure, with an atomic level density constraint. Unlike conventional magneto-optical disks, the laser light is carried to the head via an optical fibre rather than being directed via mirrors through air. As a result, the head and arm take up much less space, allowing multiple platter configurations to be fitted into the same form factor as a Winchester hard disk. Performance should be comparable to a regular hard disk too, but durability will be much higher, as the disk media is extremely non-volatile at room temperature.

MO Technology :- **Magneto-Optical Technology: a rewritable optical storage technology that uses a combination of magnetic and optical methods. Data is written on an MO disk by both a laser and a magnet. The laser heats the bit to the Curie point, which is the temperature at which molecules can be realigned when subjected to a magnetic field. A magnet then changes the bit's polarity. Writing takes two passes. Unlike with phase-change drives MO disks do not have to be "reformatted" when full. See also LIMDOW.**

Kerr Effect :- A change in rotation of light reflected off a magnetic field. The polarity of a magneto-optic bit causes the laser to shift one degree clockwise or counterclockwise.

Curie Point :- The temperature at which the molecules of a material can be altered when subjected to a magnetic field. In optical material, it is approximately 200 degrees centigrade.

Winchester Disk :- The term "Winchester" comes from an early type of disk drive developed by IBM that stored 30MB and had a 30-millisecond access time; so its inventors called it a Winchester in honour of the .30-calibre rifle of the same name. Although modern disk drives are faster and hold more data, the basic technology is the same, so "Winchester" has become synonymous with "hard".

18. PLEDM

Notwithstanding the various candidates for the hard disk technology for the next millennium, it could be that the future of hard disks doesn't lie in mechanical storage systems at all. Developments in memory technology could mean that solid-state storage becomes a serious alternative to the hard disk. Solid-state disk technology, with data stored in huge banks of high speed RAM, is already available at a high-end server level, but with current RAM it can't reach the capacities or price points to make it a mainstream proposition. However, researchers for Hitachi, Mitsubishi, Texas Instruments and others are already at work on low-cost, high-capacity RAM chips that could fit the bill. Hitachi and Cambridge University have been working on PLEDM (Phase-state Low Electron-number Drive Memory).

PLEDM uses tiny two-transistor cells instead of the capacitors used in regular DRAM and can be developed to retain memory even with power switched off. It's expected to become a commercial product around 2005, promising a read/write time of less than 10 ns and a large signal even at low voltage.

Nanosecond :- ns: one thousand-millionths of a second of a second (.000000001 sec.). Light travels approximately 8 inches in 1 nanosecond.

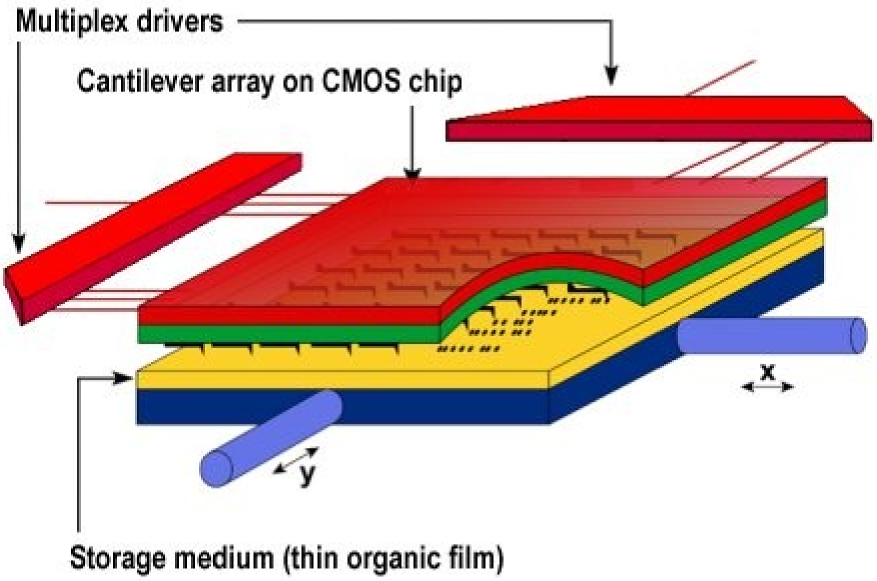
19. Millipede

In late 1999, IBM's Zurich Research Laboratory unveiled a concept which suggests that micro- and nanomechanic systems may be able to compete with electronic and magnetic devices in the mass storage arena. Instead of writing bits by magnetising regions of a disk's surface, "Millipede" - as the scientists have nicknamed their novel device - melts tiny indentations into the recording medium's surface.

Based on Atomic Force Microscopy, the technology uses tips mounted on the ends of tiny cantilevers etched in silicon to scan surfaces in minute detail. Millipede's tips are heated with electric pulses to 750 degrees F (400 degrees C), hot enough to melt the disk's polymer-film surface. The tips leave holes just 30 to 50 nm across. Each hole represents a bit. To read the data, Millipede detects whether a tip is in a hole by taking the cantilever's temperature.

The most recent array design consists of an array of 64 x 64 cantilevers (4096) on a 100 µm pitch. The 6.4 x 6.4 mm^2 array is fabricated on a 10 x 10 mm^2 silicon chip using a newly developed "transfer and join" technology that allows the direct interconnection of the cantilevers with CMOS electronics used to control the operation of the cantilevers. With this technology the cantilevers and CMOS electronics are fabricated on two separate wafers, allowing the processes used in the fabrication to be independently optimised. This is a critical feature, as many of the processes used to fabricate mechanical structures such as cantilevers are not compatible with the fabrication of CMOS electronics.

The cantilevers used in the array are of a three-terminal design, with separate heaters for reading and writing, and a capacitive platform for electrostatic actuation of the cantilevers in the z-direction. The cantilevers are approximately 70 µm long, with a 500-700 nm long tip integrated directly above the write heater. The apex of each tip has a radius on the scale of a few nanometers allowing data to be written at extremely high densities (greater than 1 Tbits/in^2). In addition to the cantilevers, the array chip also carries eight thermal sensors which are used to provide x/y positioning information for closed-loop operation of the micro-scanner.


Multiplex drivers
Cantilever array on CMOS chip
x
y
Storage medium (thin organic film)

High data rates can be achieved by parallel operation of a large number of tiny tips in a small area. IBM scientists believe their technique will eventually allow storage densities of up to 500 gigabits per square inch (80 Gbit/cm^2). That's five to ten times the presumed upper limit for magnetic storage.

Nanometre :- nm: one thousand millionth of a metre.

Micron :- µm: a unit of measure equivalent to one-millionth of a metre; synonymous with micrometre.