Introduction to Hard Disk Drives

(What goes on inside your hard drive)

Larry Wittig

Lexington Computer and Technology Group Meeting

7 October 2009

Outline of Presentation

Background – Historic Overview

- What is a Hard Disk Drive (HDD)
  - Understanding the basics by looking at early HDD
- Amazing increase in storage capacity over 50 years
  - Through shrinkage of critical parts and advancements in technology

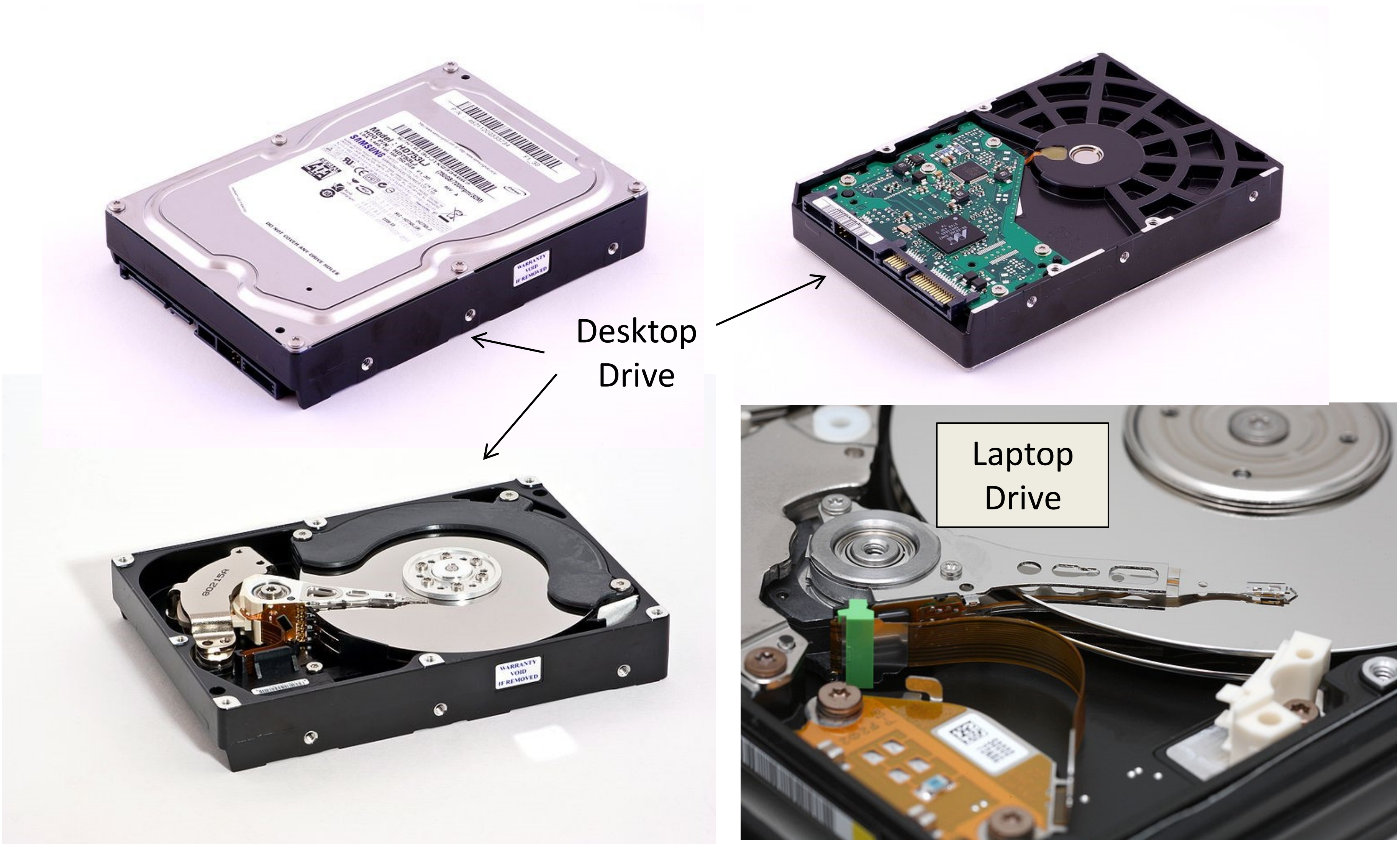
Modern HDDs

- Mechanical overview
- Read/Write Recording Heads
- Disks and Servo
- Actuator and Spindle motor
- Electronics and interfaces

Appendix

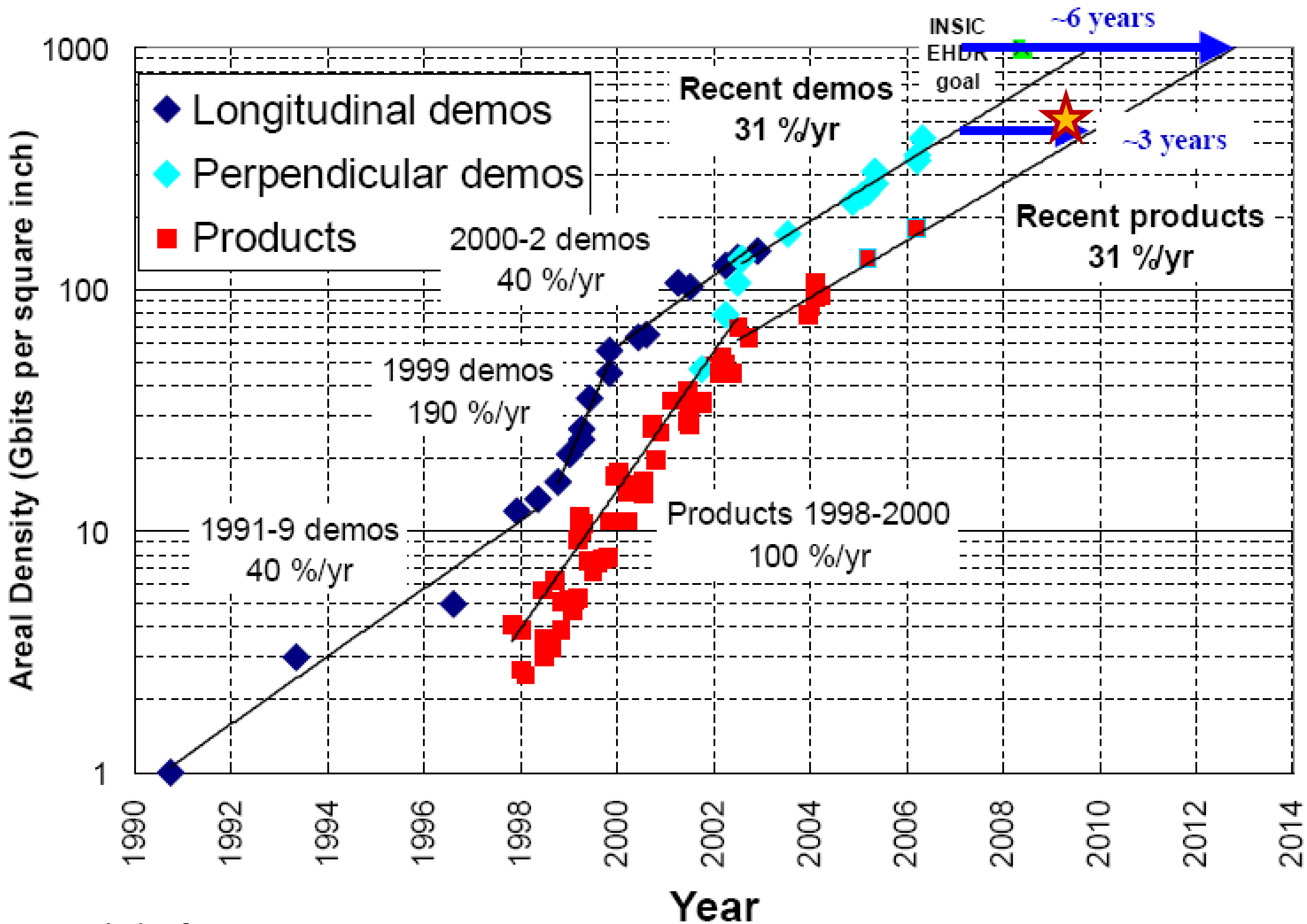
- More HHD advances in the works
- Solid State Drives (SSD)
- A Few Comments on RAID
- Some slides of a more philosophical nature on how much storage is enough

A Hard Disk Drive is a somewhat of a cross between a tape recorder and a record player
(but it's digital as opposed to analog)



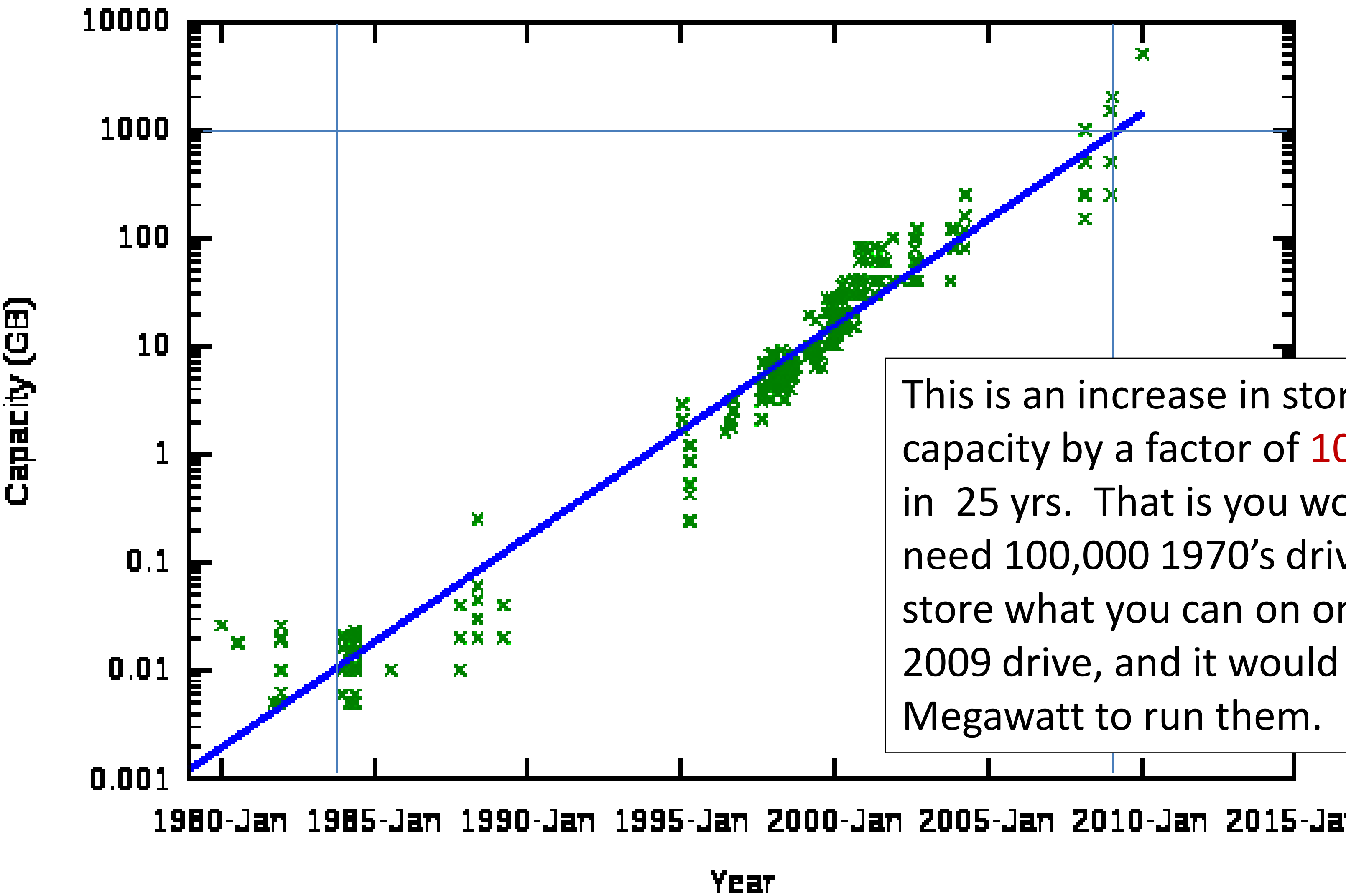
Mid 2009 ~ 450 Gbits/in^2. Approx. 300k Tracks/in x 1.5 Mbits/in. That's about 1000 tracks in the width of a human hair, and 500GB/3.5" disk.

HDD Areal Density Trends – Demos & Products



Slide from → ©2007 Information Storage Industry Consortium

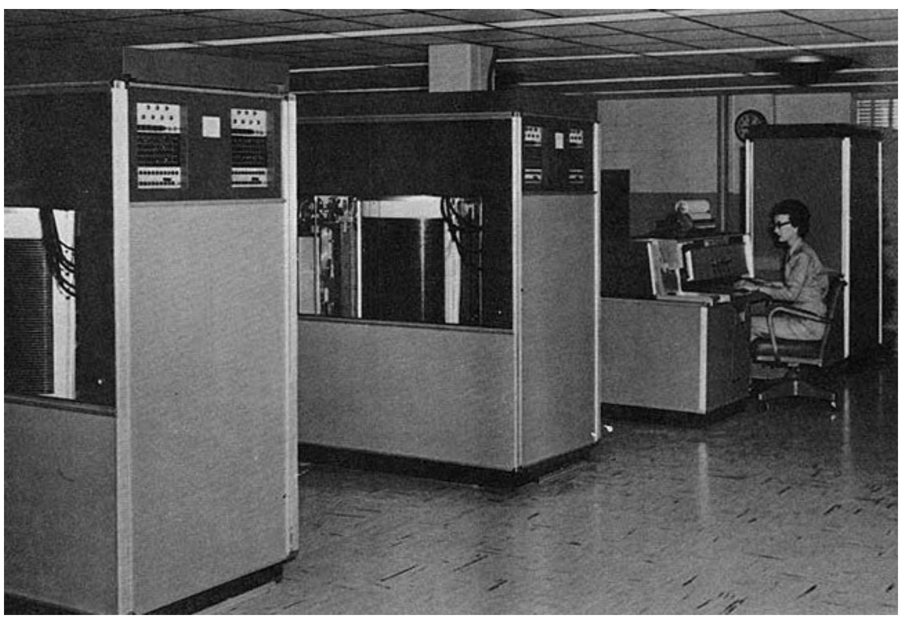
PC Hard Drive Capacity
(source Wikipedia)

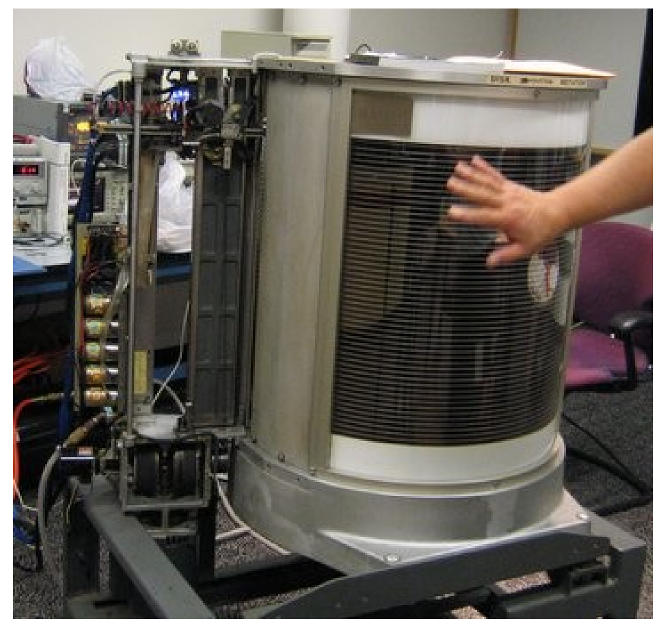


This is an increase in storage capacity by a factor of 100,000 in 25 yrs. That is you would need 100,000 1970's drives to store what you can on one 2009 drive, and it would take a Megawatt to run them.

First Hard Disk Drive

First commercial Hard Disk Drive (HDD) – IBM's RAMAC (Random Access Method of Accounting and Control) stored 5 MBs. A present 2 TB desktop HDD stores 400,000 times more data. RAMAC had fifty 24-inch diameter disks and was leased for $3,200 per month equivalent to a purchase price of about $160,000 in 1957 dollars (about $1.2M in 2009 dollars)





Size Progression of HDDs
From about 1980 to Present

Desktop ~4"x 6"

Laptop ~3"x 4"

PC slot, IPOD

Micro

5.25" size that really made the transition. Seagate came out with a 1-inch high variation of this which became the de facto desktop standard

Shown with
a thin (1 disk)
laptop drive.

Larger than desktop drives were killed by:

R Redundant
A Arrays of
I Inexpensive (Independent)
D Drives

Based on a U. Cal. Berkley paper published in 1988. It showed that you could get higher overall reliability and lower cost by using multiple redundant inexpensive drives instead of a smaller number of highly reliable more costly drives without redundancy. The trick is you have to quickly replace any failed drives and reconstruct the data before another drive fails. The article points out several ways to do this. The simplest way to do this is RAID1 (also called mirroring) where the same data is written to two HDDs Some levels of RAID can also improve performance.

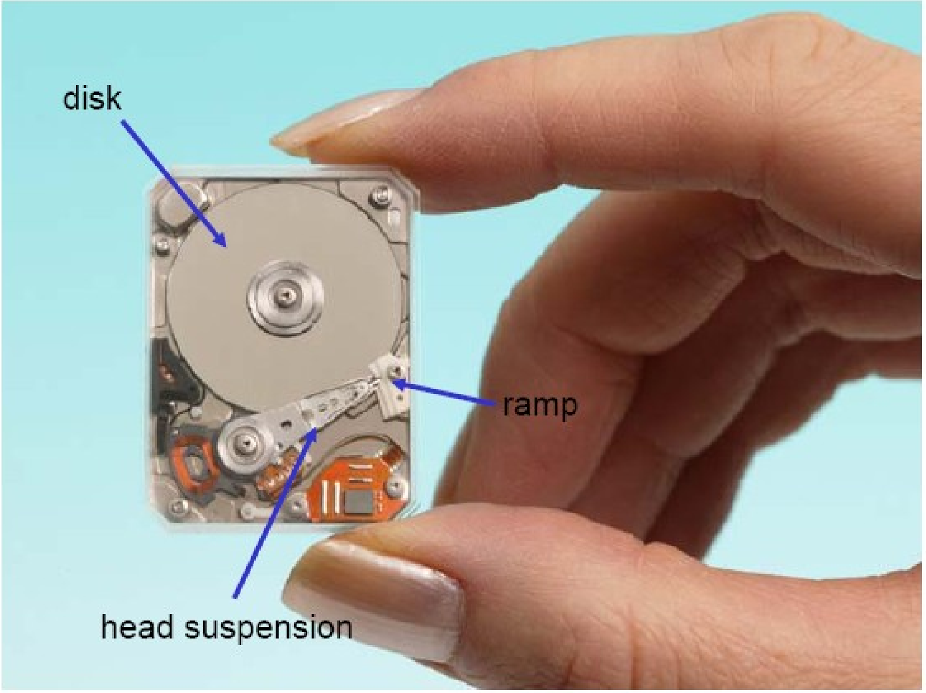
There are a few additional slides in the Appendix on RAID

Micro Drives Have Not Succeeded Very Well Against Recent Advances in USB Flash Drives

TOSHIBA **0.85" HDD Structure**

Even in the ultra small 0.85" HDD, the basic structure is the same as the other large size HDD. JUST down-scaling!

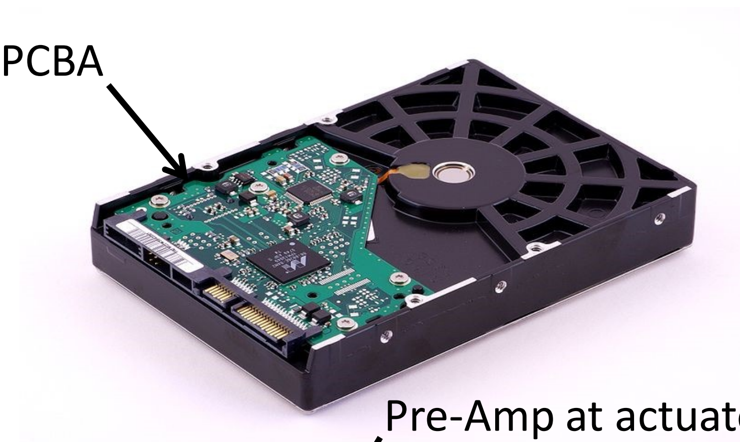
IBM Drive with a 1-inch disk

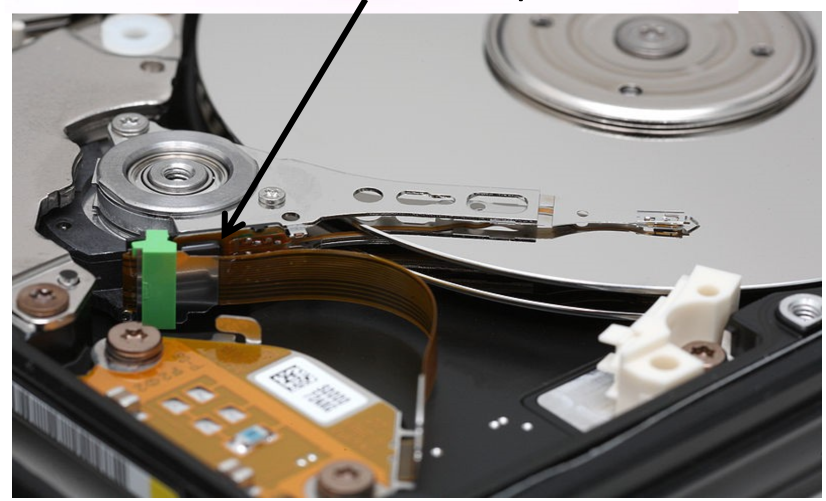


Micro drives peaked at a capacity of approximately 20GB, but they are no longer manufactured.

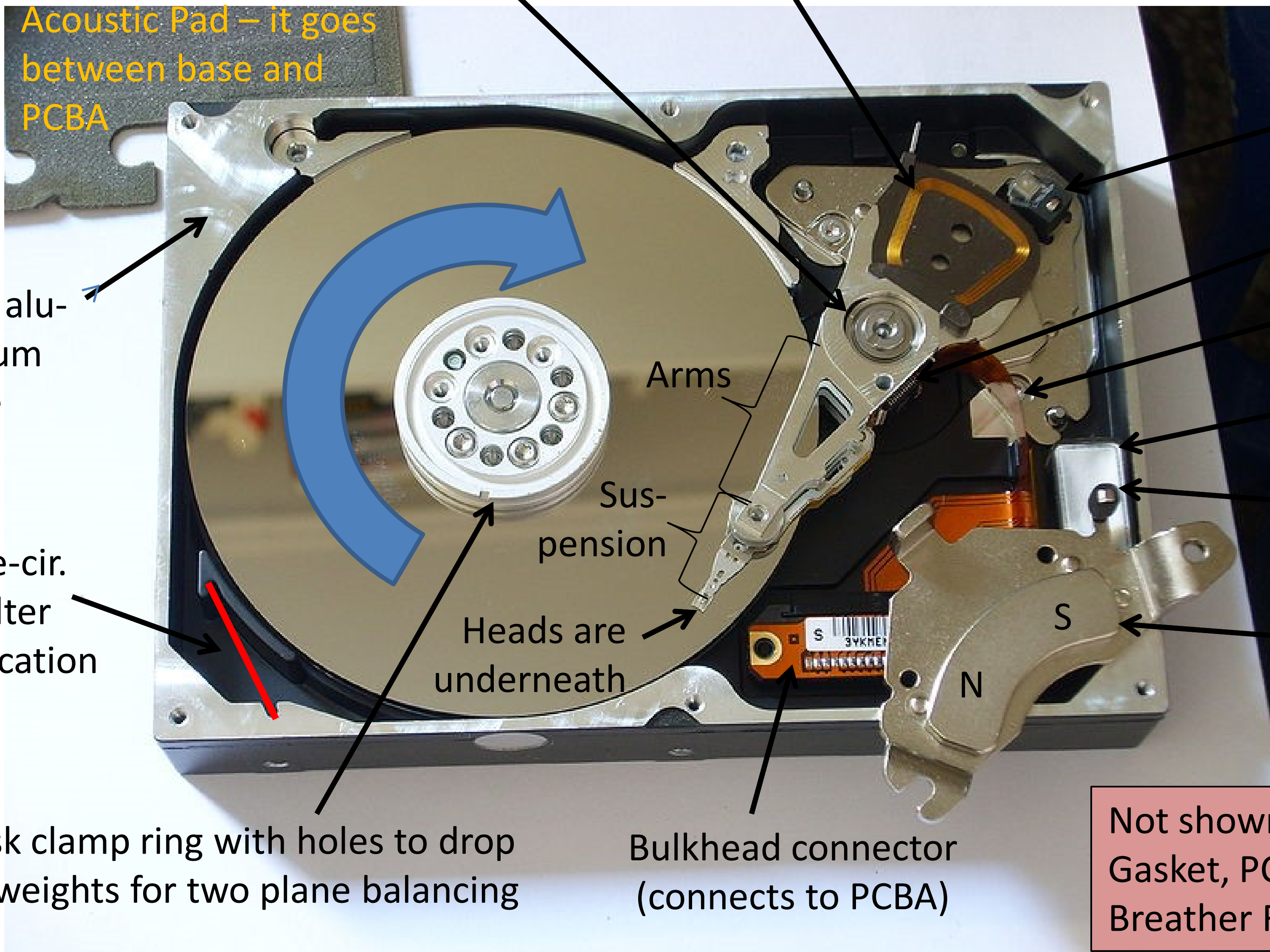
Back to what goes on inside your Hard Drive

It's called a **hard drive** because the disks are **hard** (usually aluminum but sometimes glass). The maximum number of disks in a desktop drive now days is usually four, and the max in a laptop drive is two. In most cases there are two heads per disk (one on each side) but only one head in the drive is working at any one time.





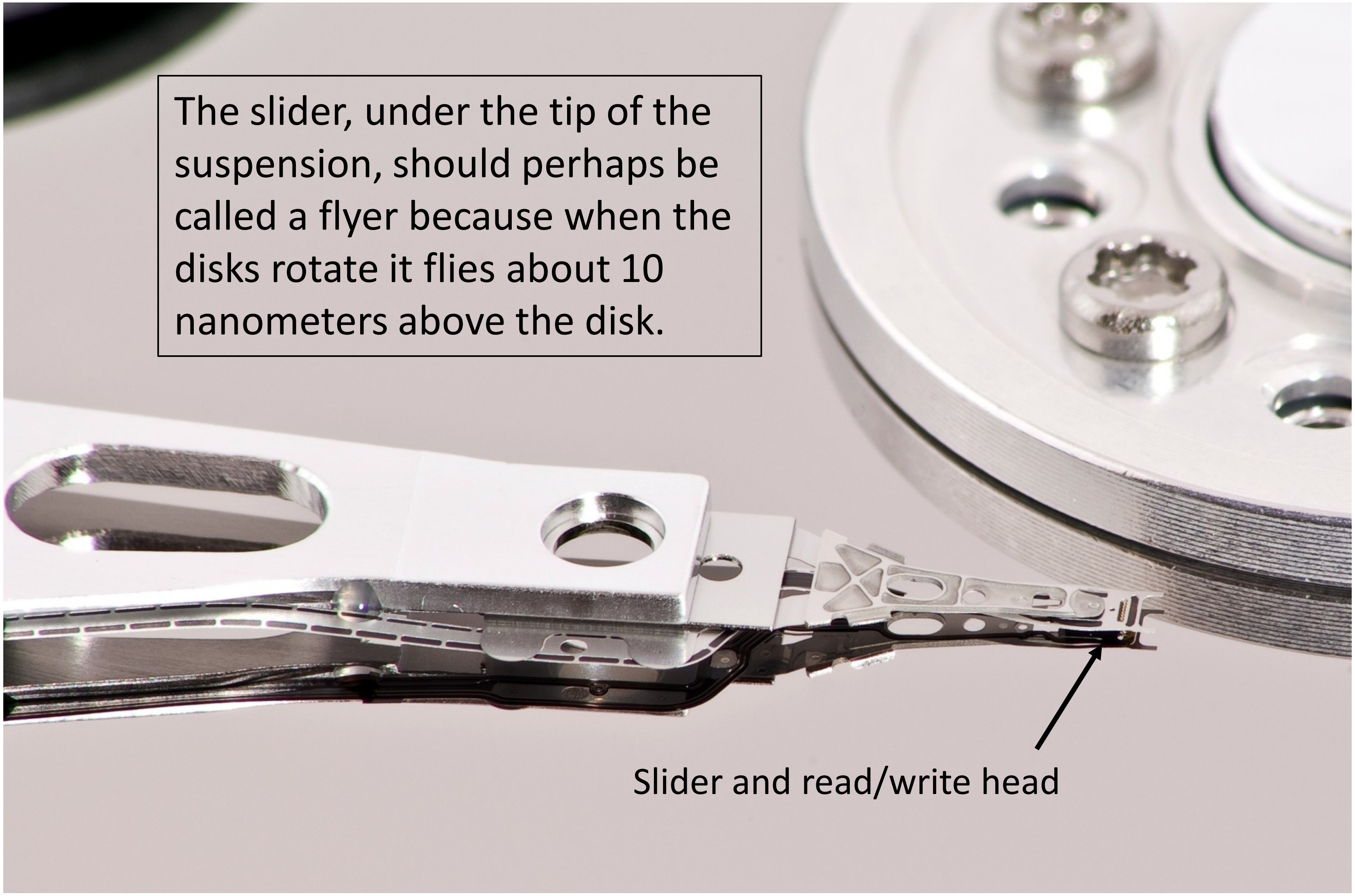
What’s Inside: Rotary Actuator and Other Parts



Not shown: Cover, Gasket, PCBA and Breather Filter

Arm Tip, Suspension, Slider

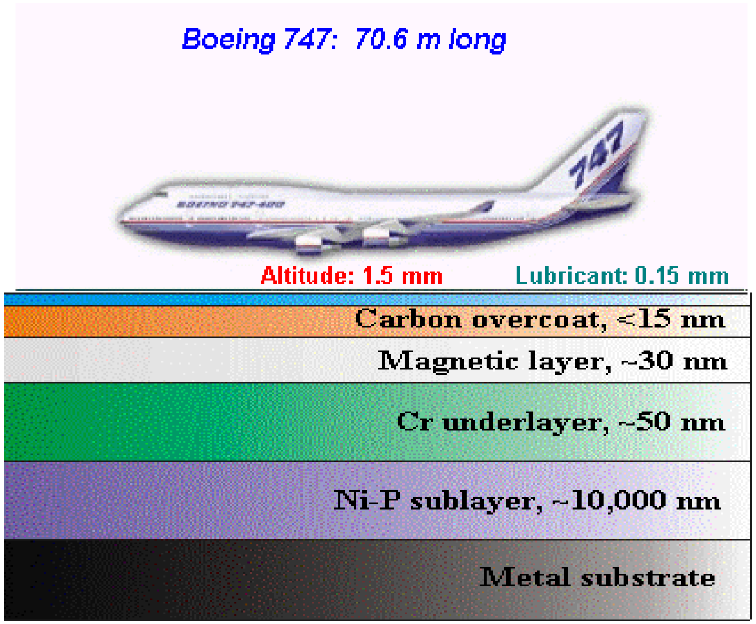
This one is stopped on the disk in the landing zone.



Comments on Fly Height

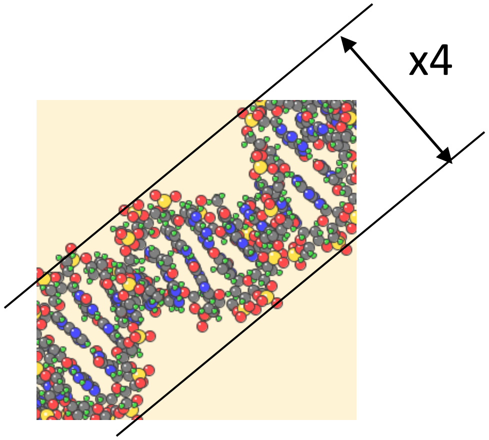
Now days (~2009) the head end of the slider "flys" about 10nm above the disk. That's about one-ten thousandths' of the diameter of a human hair or the width of four DNA molecules.

Why fly so low? Because the magnet flux density drops off like $1/R^3$. Ten times further from the disk the magnetic field strength is 1000 times weaker.



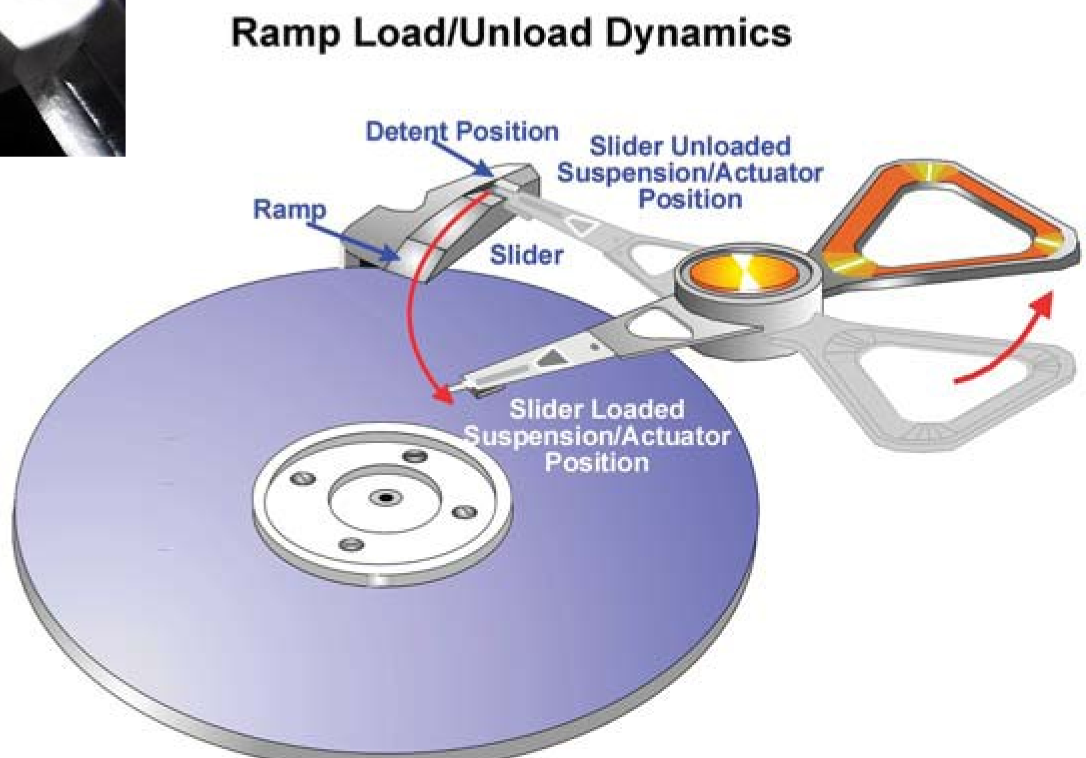
So if you want to pack in a lot of data (using small magnetic domains) you have fly close to the disk to get a strong enough signal.

What the head and processing electronics are actually trying to sense are the transitions from one magnetic polarity to another.



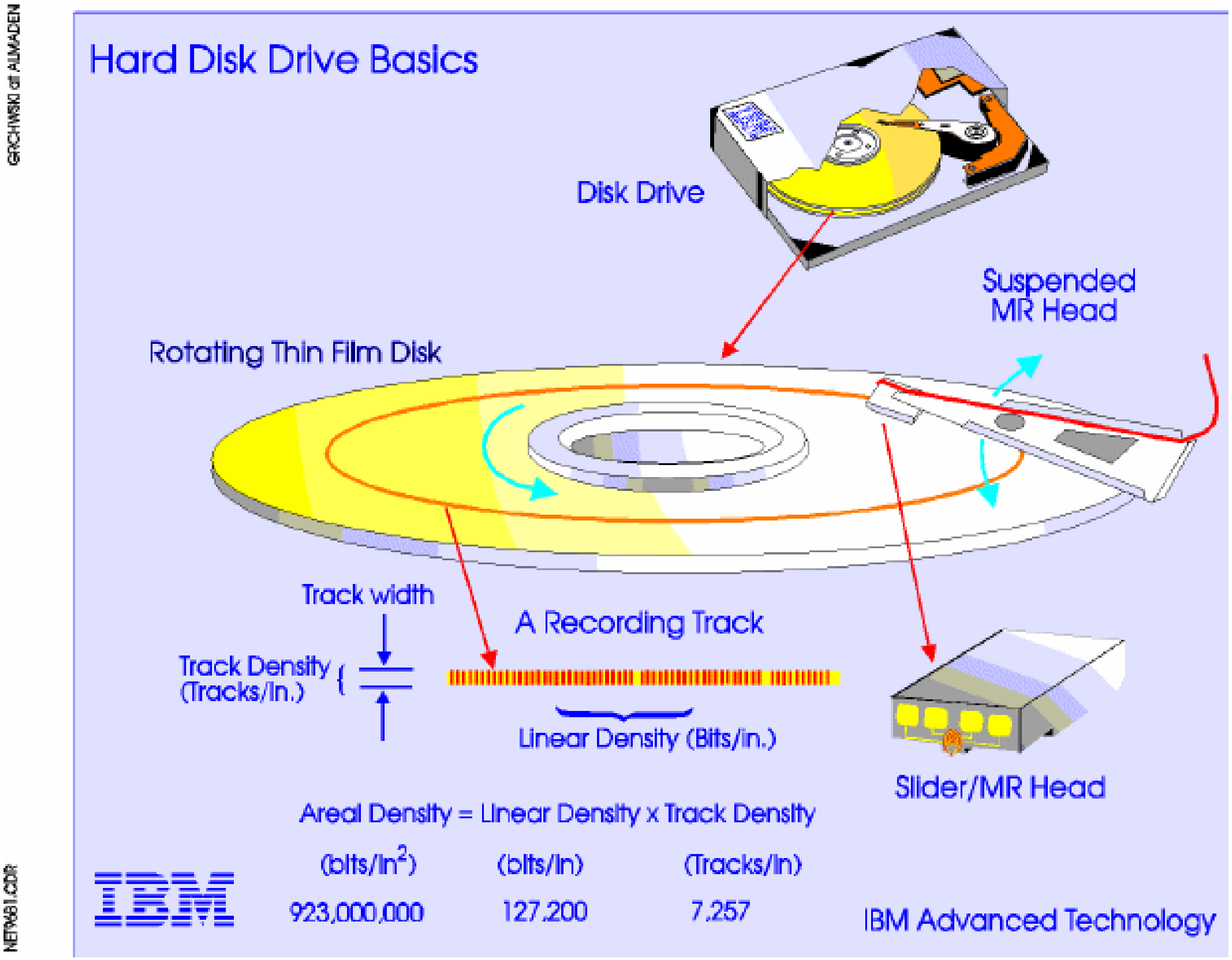
Ramp Load Allows for Smoother Disks (lower fly height) and Provides Shock Protection

Drives used in Laptop computers have ramp load for shock protection.



Laptop drives also have free-fall sensors and they "park the heads" on the ramp if an impact is about to happen.

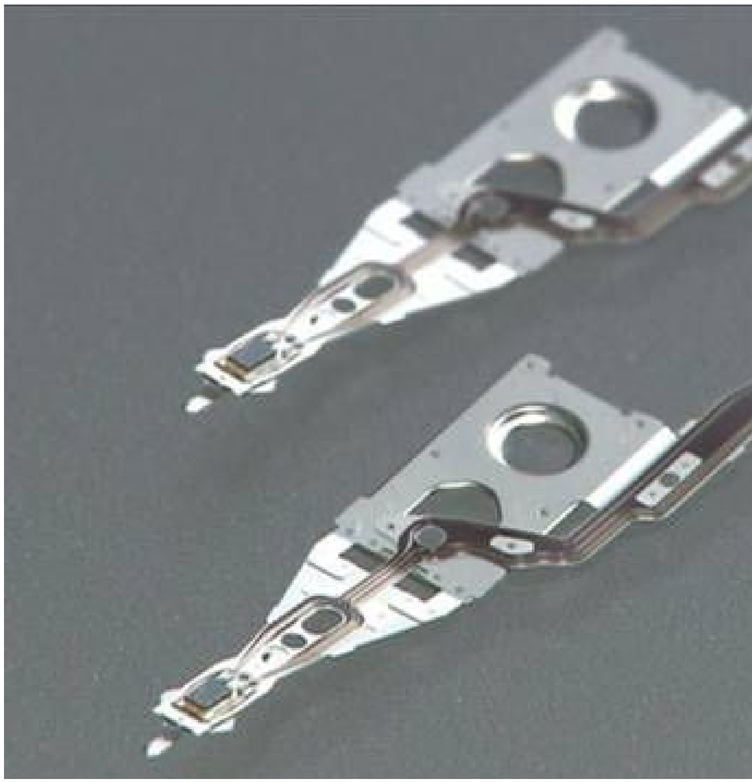
How Data is Stored and Recalled

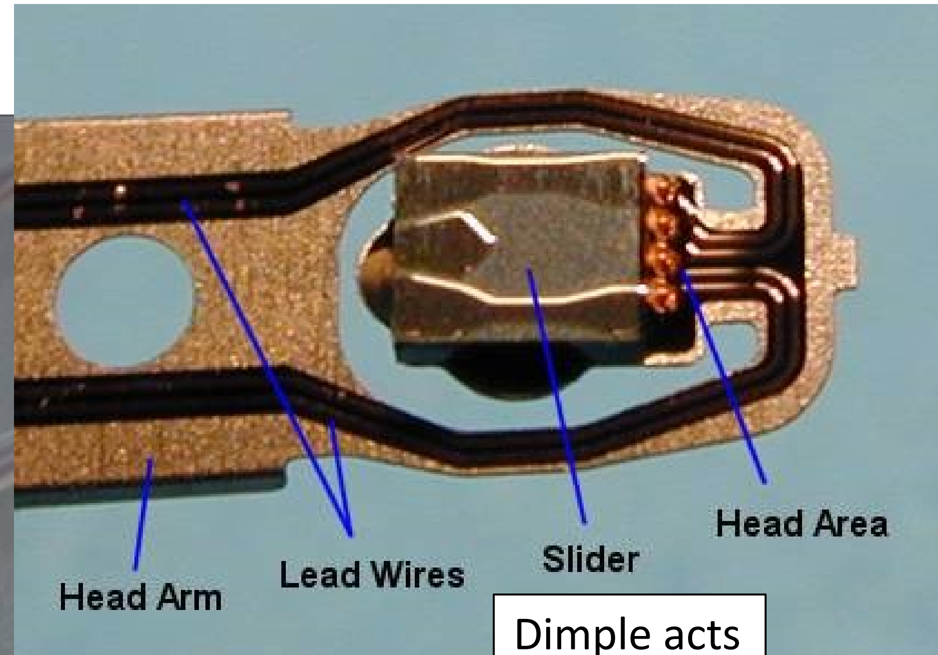


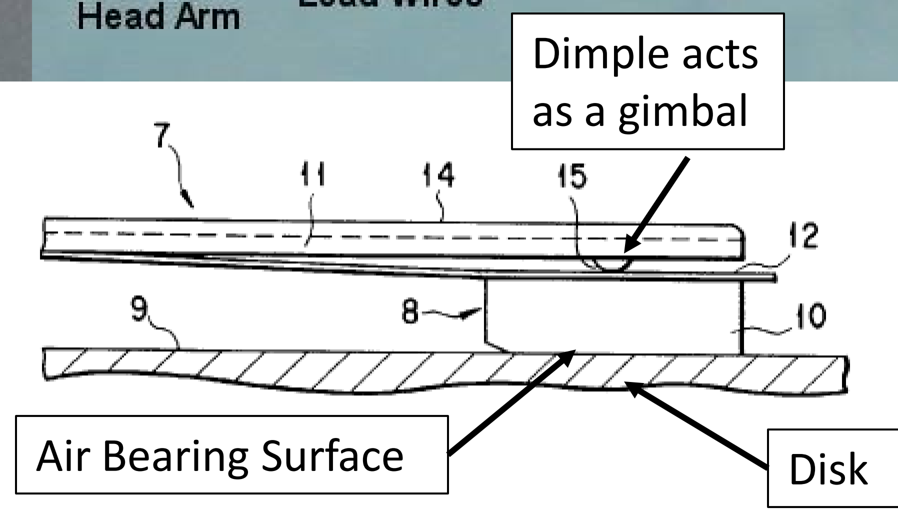
Now days (2009) the linear density is about 10x as high as shown here resulting in the ability to store 1 to 2 MBytes per rev. A 200 kB JPEG picture takes $1/10^{th}$ of a rev.

It takes 8 bits to make a Byte.

The Suspensions is like a leaf spring that pushes the Slider against the disk







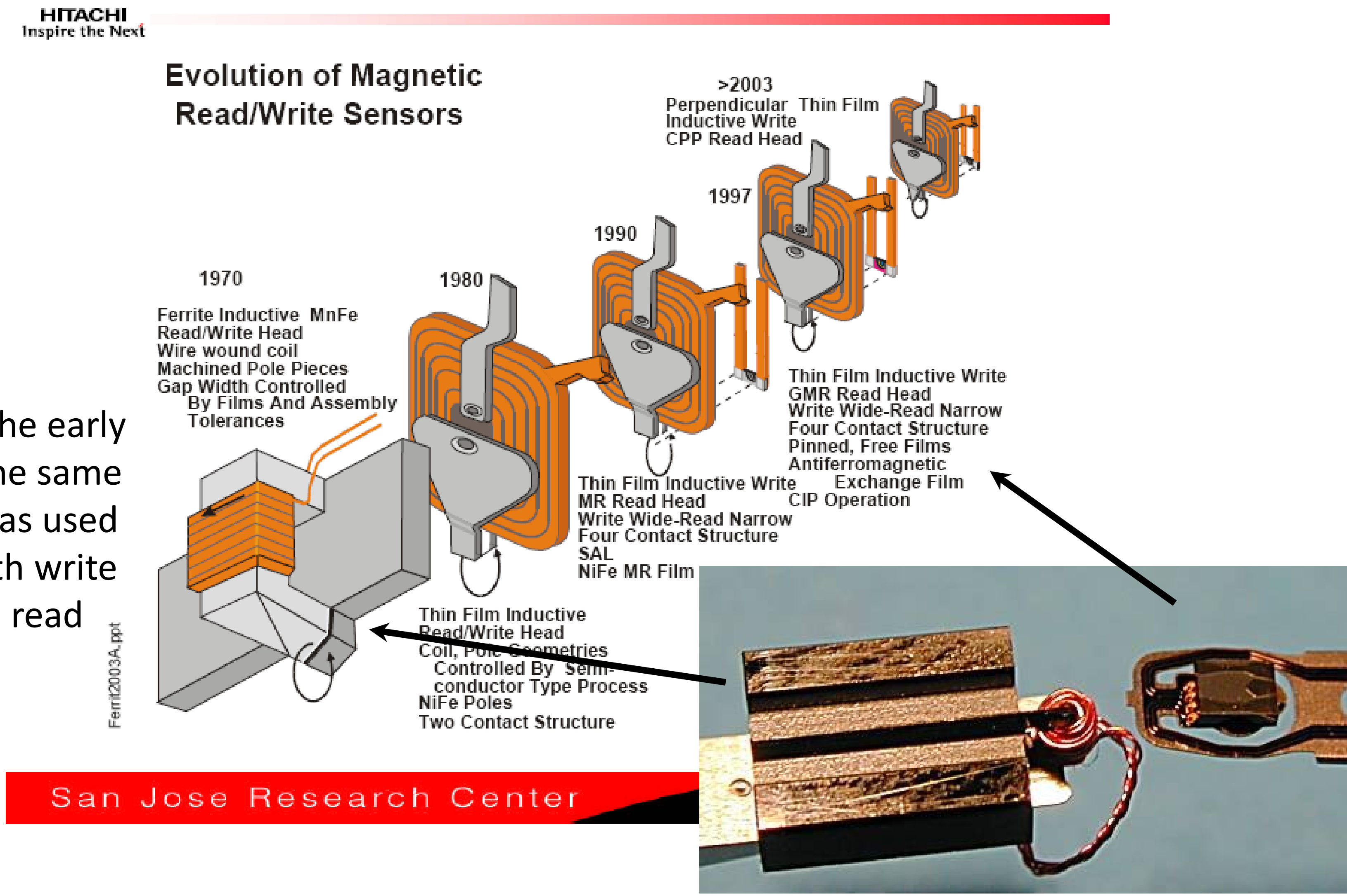
Sliders and Heads

The write gap width is about the width of the line

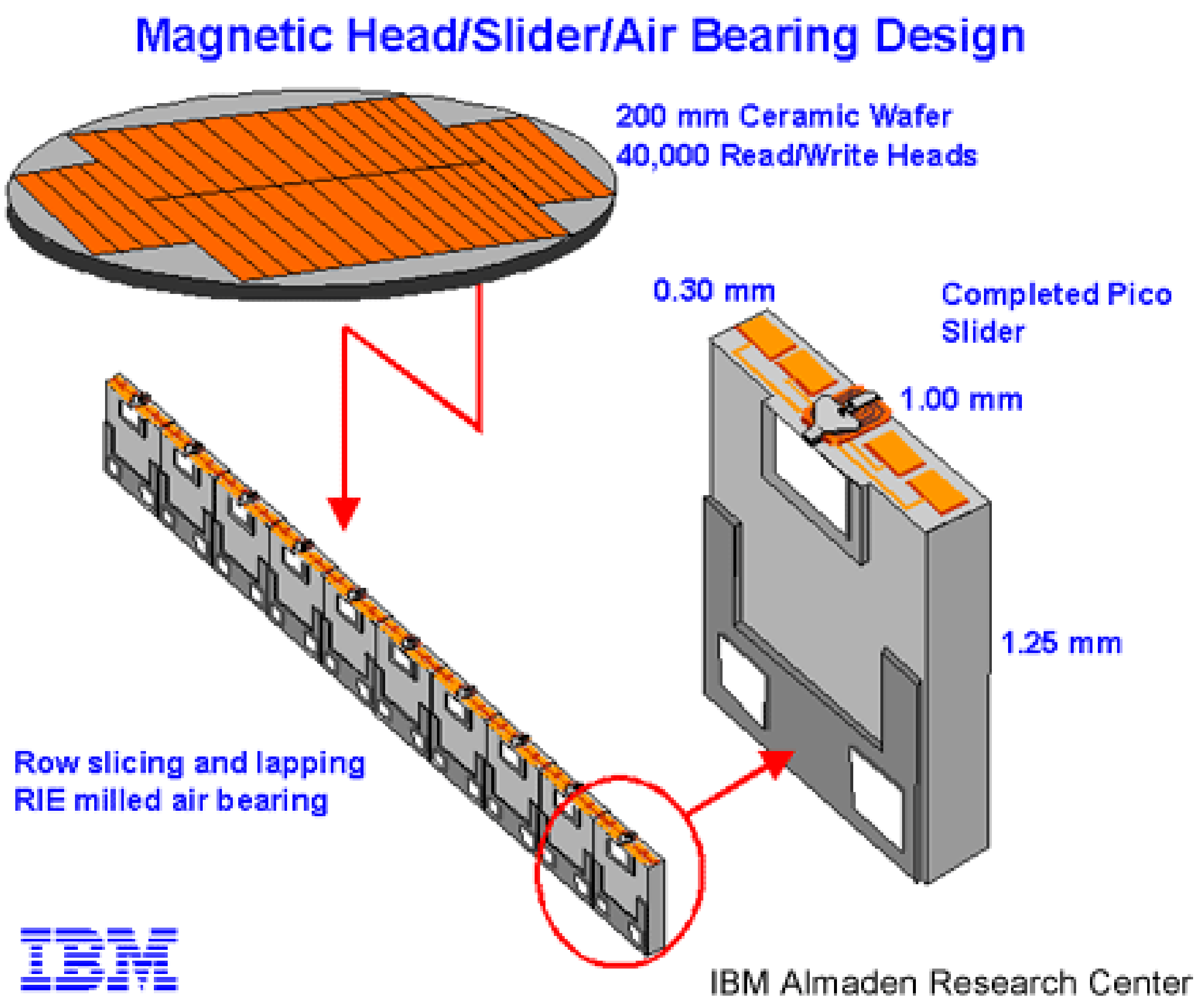
Major Innovations were (1) Thin Film Heads, (2) MR Readers and (3) Perpendicular Recording

Other non-head/disk related: Error Correction Codes, PRML and Reduced BAR

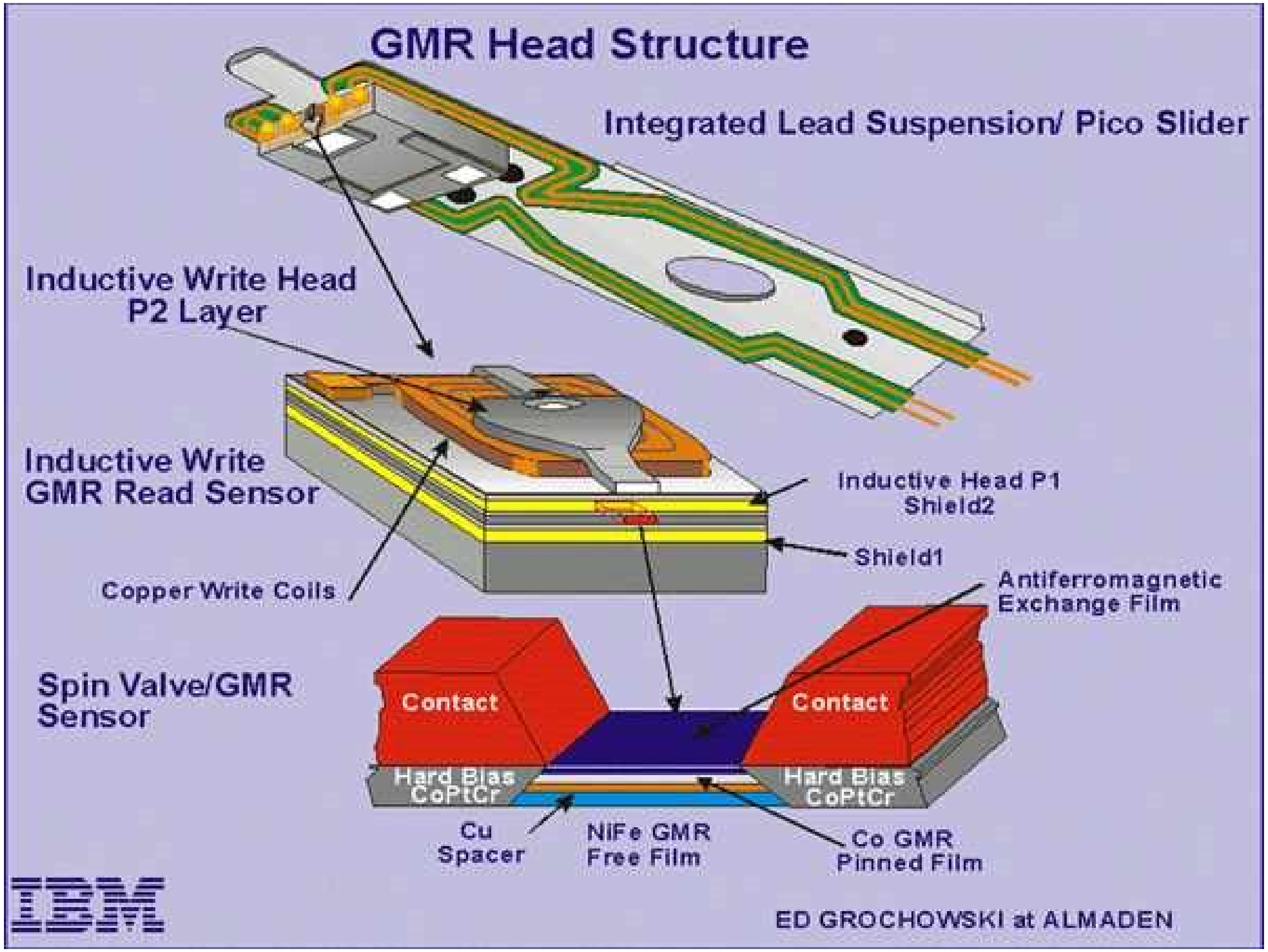
Until the early 90's the same coil was used to both write and read



Thin Film Heads are built on wafers similar to how computer chip are manufactured



GMR Heads – The IBM Physicists that developed MR & GMR materials won a Nobel Prize

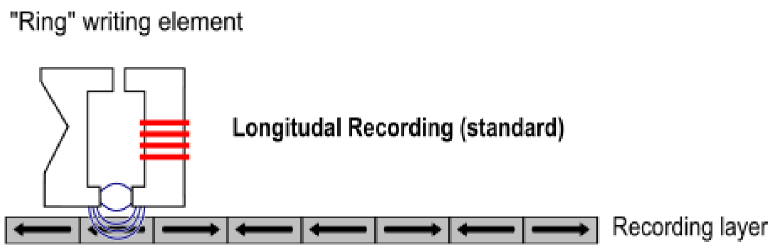


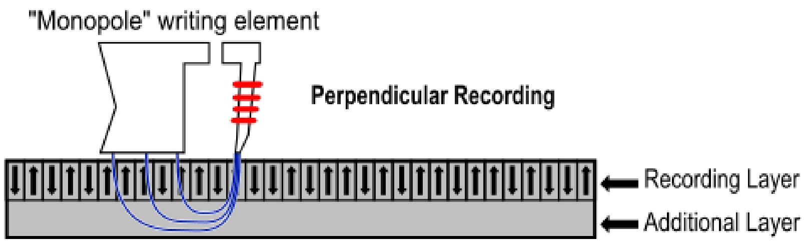
MR = Magneto Restrictive – the resistance of the material changes in the presence of a magnetic field

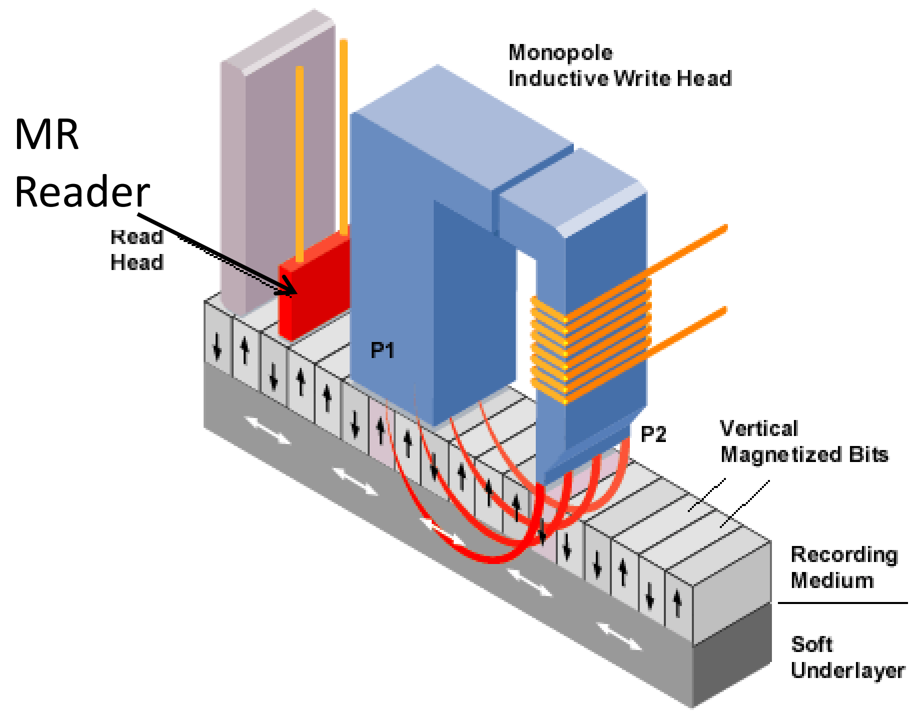
GMR = Giant MR effect

TMR = Tunneling MR

Perpendicular Magnetic Recording is now allowing the next increase in areal density

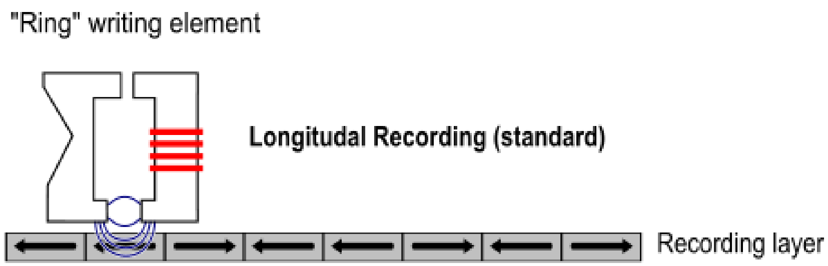




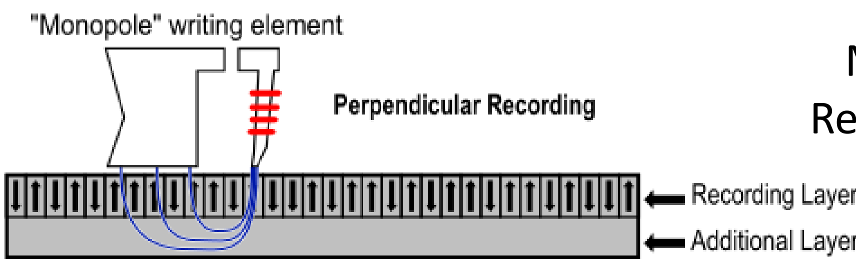


By standing the magnetic domains on end the bit density can be increased.

Perpendicular Recording is now allowing the next increase in areal density



If the magnetized volume is too small it's not stable



By standing the magnetic domains on end the bit density can be increased.

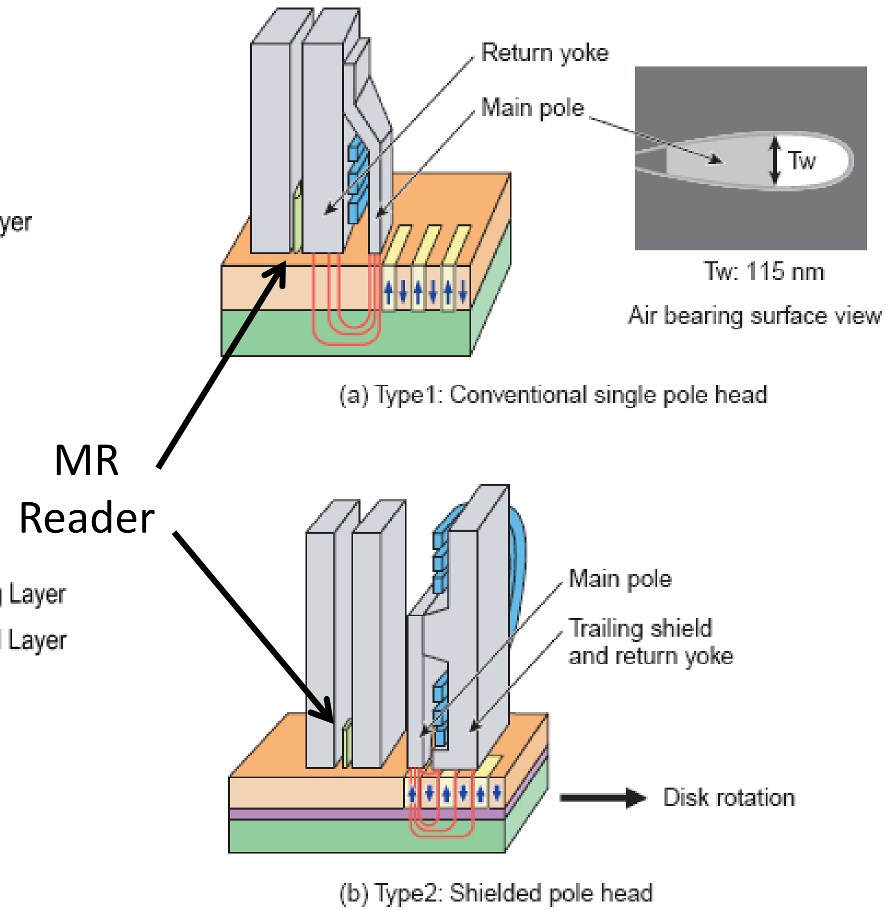
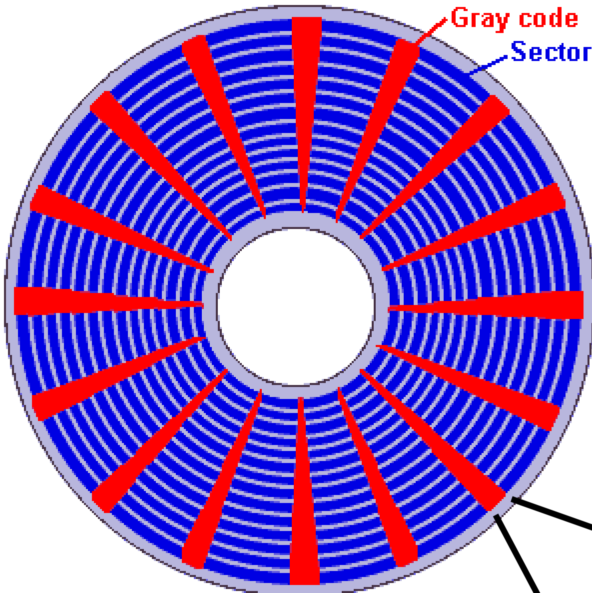


Figure 10
PMR head structures.

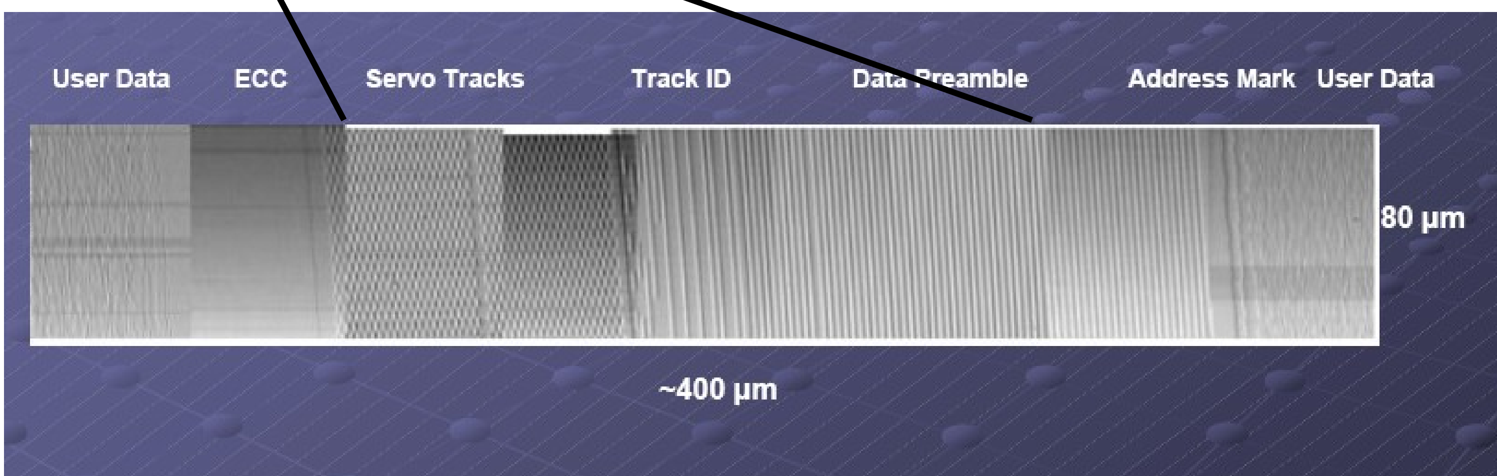
Servo Tracks are used to locate position on the disk



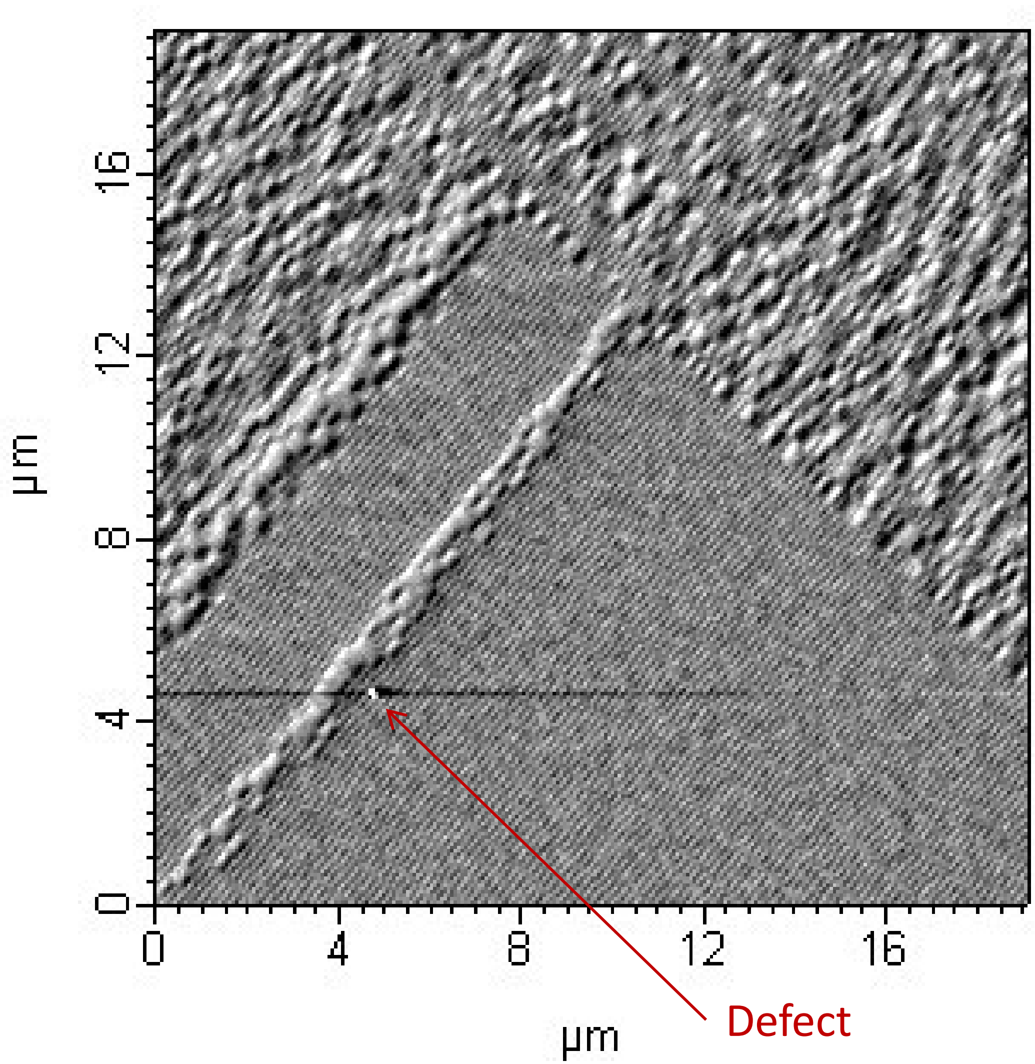
Servo wedges are "written" after the drive is assembled using the same heads that are used to write data. Wedges are used to (1) mark track and sector locations ,(2) provide timing info, and (3) provide "servo bursts" that are used to fine tune staying on track.

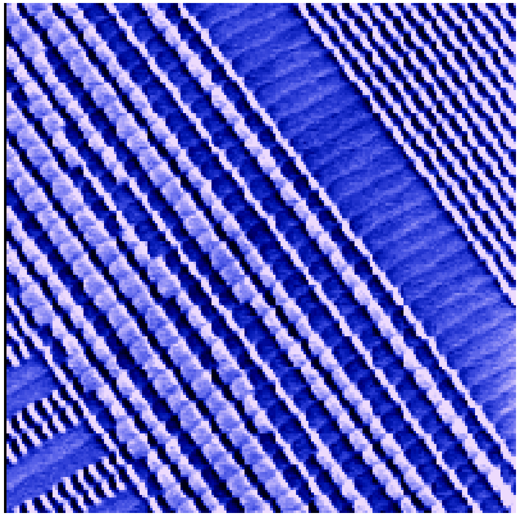
Approximately 75% of the surface is used for storing data. Present disks have ~ 200 wedges. After servo writing drives undergo about 30 hours of testing.

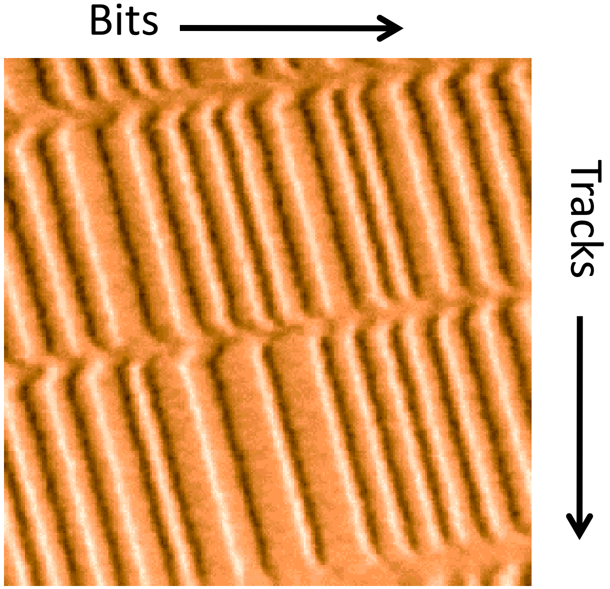
Approx-mately 100 tracks



Magnetic Force Microscope Images of a Disk Surface Showing Servo Sector and Data



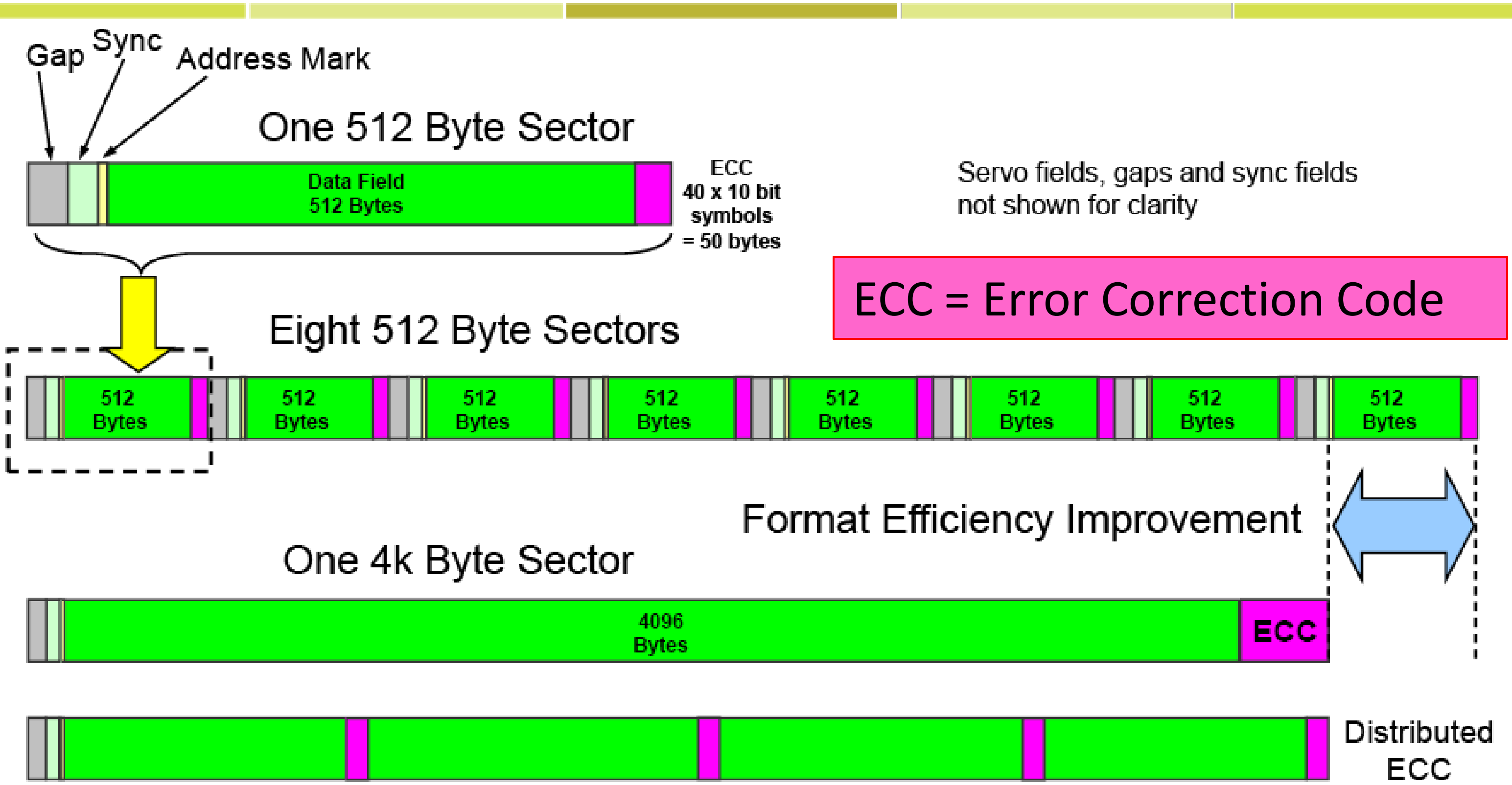




How the data is formatted between wedges

Format Efficiency with Long Block

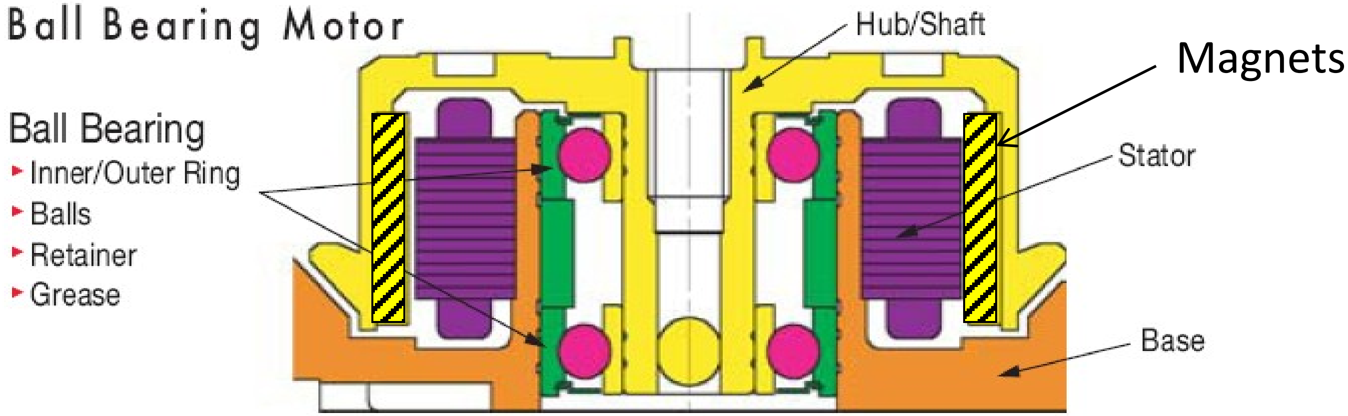
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2008

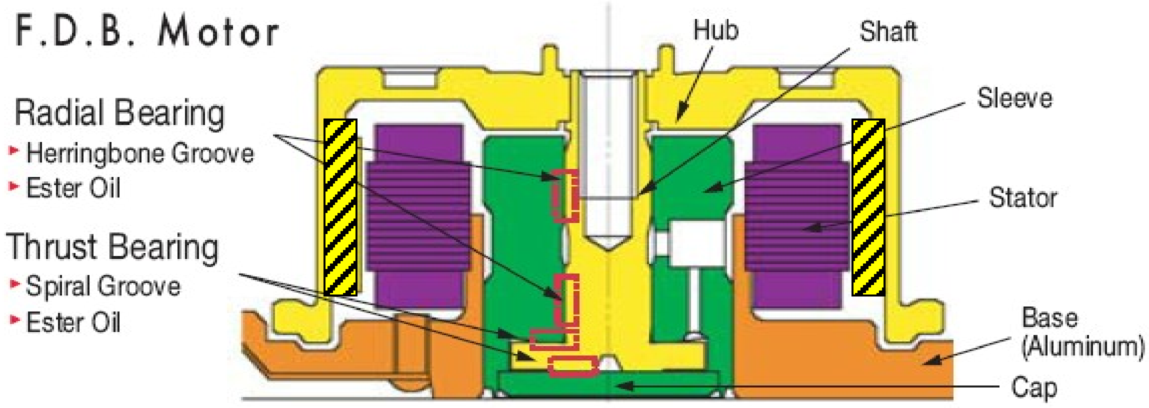


- Format Efficiency improves by 6-13% with 4kB sector
  (depends on 512B sector layout, and disk size)
- Gains can be used to reduce BPI or TPI and improve yield

Spindle Motor Cross Section

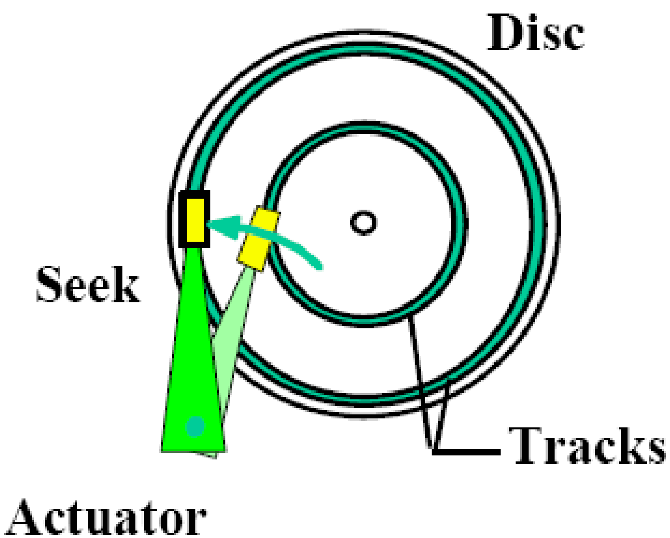
There was a change from Ball Bearing Motors to Fluid Dynamic Motors about 2002 because BB caused too much vibration. FDB are also quieter.





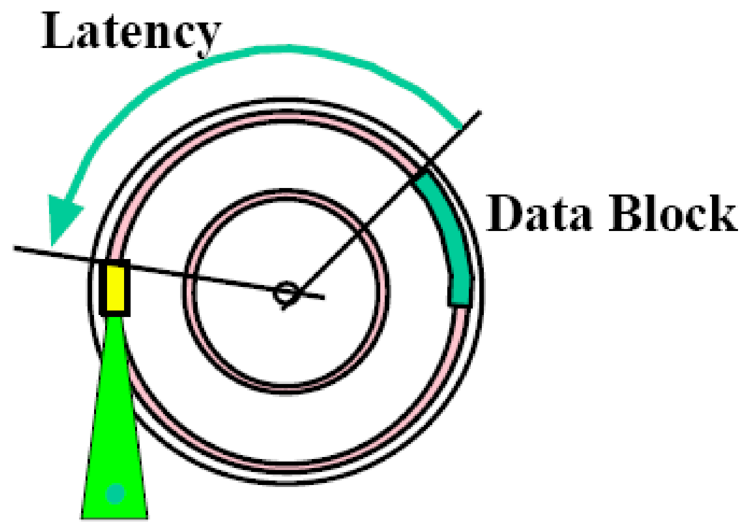
Mechanics: Getting to the Data

Seek Time



Keys: Swing length
Arm length, mass
Magnetic circuit

Latency Time



Power = ~RPM**3
= ~Diameter**5

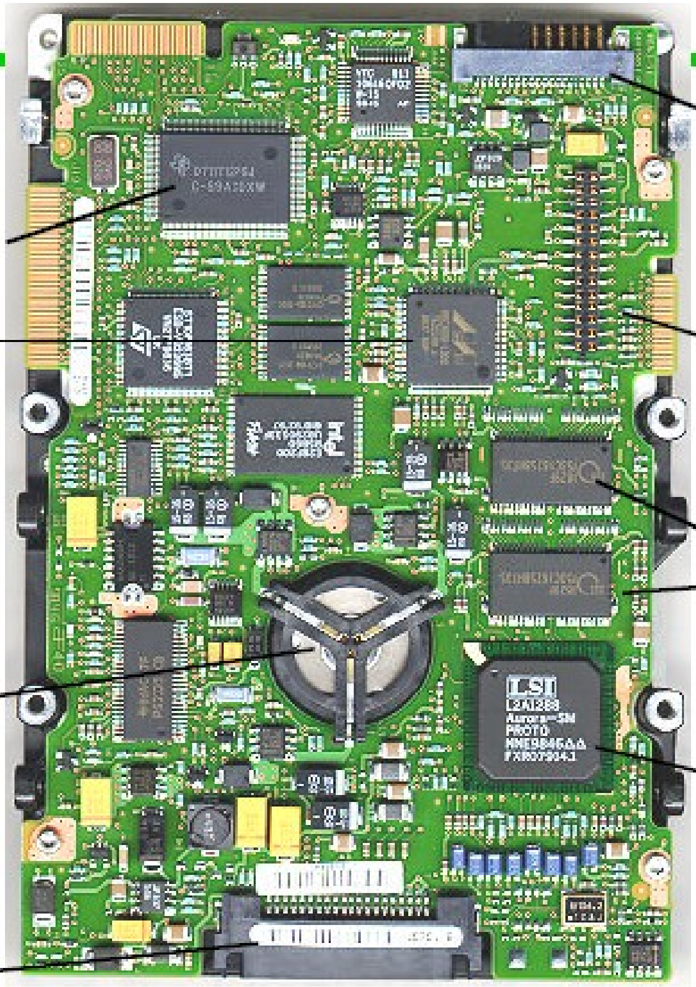
Seagate

Comments on Seek Time and Rotational Latency

- Seek Time is the time to move the actuator from on track to another. Average seek time ranges from about 15 ms on laptop drive down to about 2.5 ms on Enterprise drives. On a new drives with all data close together seeks are all short and the performance is better.
- Rotation latency is the time for one half revolution of the disks. It ranges from ~ 5.5 ms on a 5400 rpm drive down to ~2 ms on a 15,000 rpm Enterprise drive.
- For home use slow rpm and slow seek times are very acceptable. Most laptop drives even go into a low power mode between seeks which substantially reduces speed. But for a high end server a drive that is twice as fast can be worth two slower drives.
- Given commands to read or write several files Enterprise drives will optimize how it does this – this results in shorter seeks and consequently rotational latency is more important then seek time.

PCBA = Printed Circuit Board Assembly

PCBA - Cheetah



Position Microprocessor

Read/Write Channel

The R/W chip is propri-etary and its where error correction takes place.

Connector to Spindle Motor

Connector to Interface Adapter

Connector for Serial Port Test

Connector to Media

RAM

Controller

Seagate

Interfaces – the electrical interface to the computer

Desktop/Laptop Drives

- IDE, ATAx & SATA – see next page
- USB – Presently used for most external drives. It's slow, but fast enough for most home/small office applications .
- Firewire (IEEE 1394) – used for some external drives – faster than USB but did not catch on

Enterprise/Server Drives

- SCSI (Small computer system interface)
- SAS (Serial attached SCSI and has replaced SCSI)
- Fiber Channel – used in very high end arrays

SAS has about 10x the I/O rate of SATA for short files, can support many more drives per cable, and can talk to drives on a 10m cable vs. 1m for SATA.

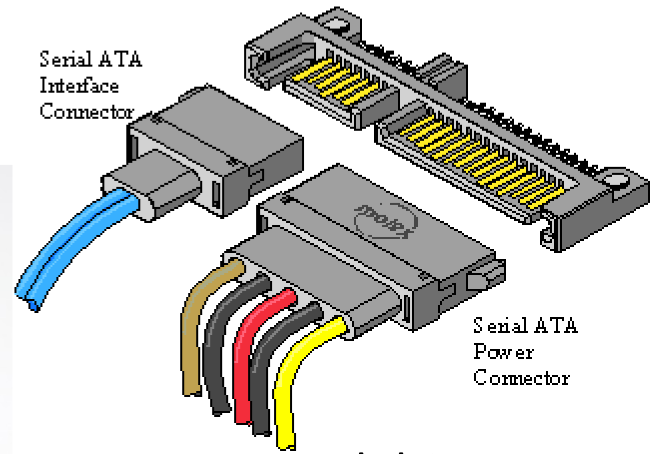
Interfaces for Desktop/Laptop Drives
Newer Drives Use SATA

IDE = Integrated Drive Electronics developed by Western Digital

ATA = Attachment Packet Interface; Standardized version of IDE; there were also EIDE and ATA2 thru ATA8 versions.

There are 40 pins. 16 pins are used for data, and the rest are used for control or power.

SATA = Serial ATA, there are two wires for everything except power.



New computers and drives use SATA. A computer with an ATA interface will not accept a SATA drive without a convertor board.

PATA: when SATA was introduce ATA was renamed Parallel ATA

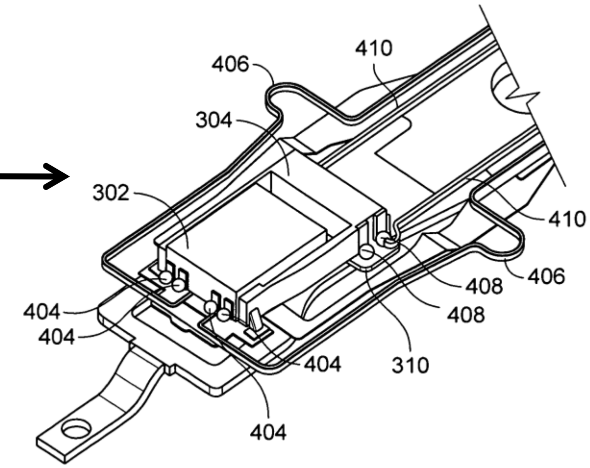
Be kind to your Hard Drives
and back them up frequently.

Appendix
Additional Material

- More HHD advances in the works
- Solid State Drives (SSD)
  - An expensive but fast and robust alternative to HDDs
- A Few Comments on RAID
- Some slides of a more philosophical nature on how much storage is enough

More HDD Advances in the Works

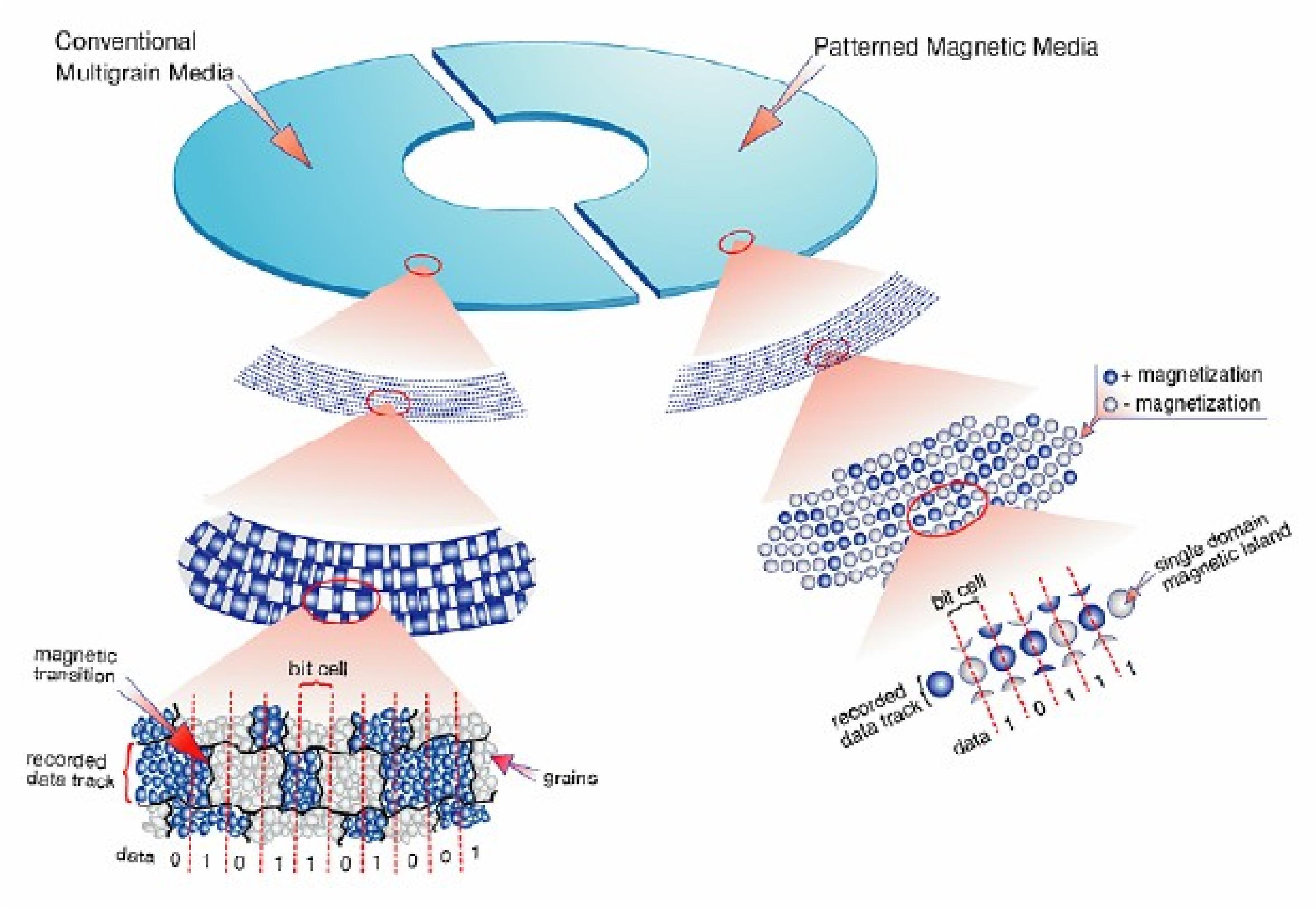
- Micro-actuators – now in some drives →
- Fly-height adjust -- now in some drives
  - Uses a very small heating coil to expand the material near the head



- Patterned media
  - An idea that has been around for a long time
  - Instead of ~50 small grains per bit, it would use one large grain per bit
- HAMR = Heat Assisted Magnetic Recording
  - Seagate invested lots of $$$ on this but then seemed to back off

Conventional Media vs. Patterned Media

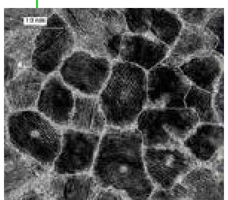
HITACHI
Inspire the Next



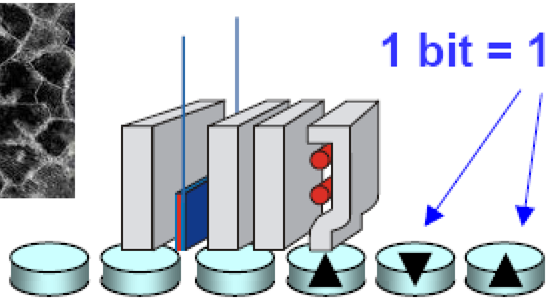
Beyond Conventional Perpendicular Recording

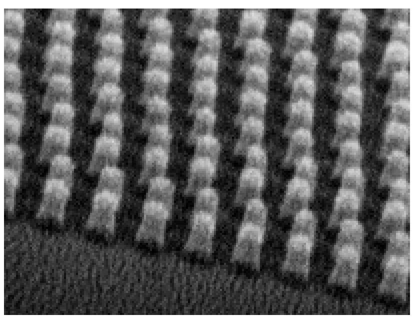
(two favorite technology options to extend thermal limit)

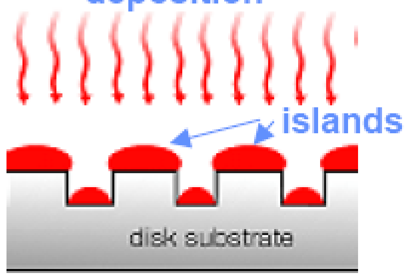
Patterned Media (increased V, utilizing 1 large "grain" per bit)



Now



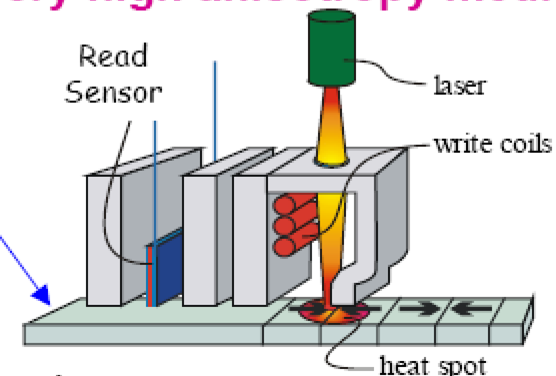


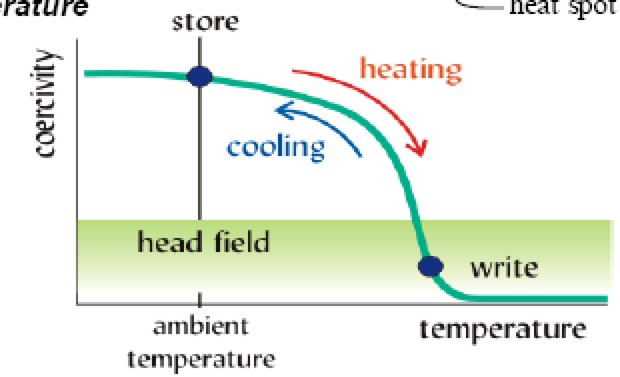


Thermal Assist (increased K_u, utilizing very high anisotropy media)

Heat-Assisted Magnetic Recording

high anisotropy medium sensitive to temperature





Challenges: *Disk Manufacture*
Lithography/Stamping

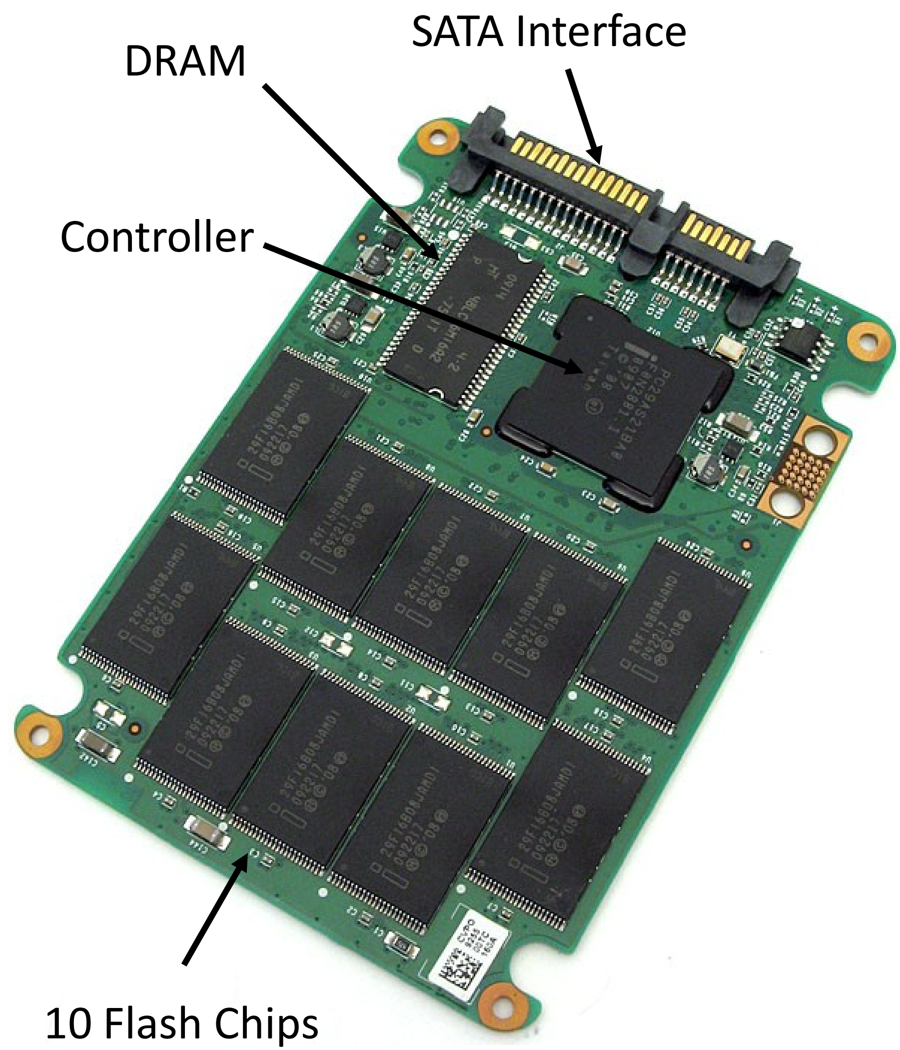
Challenges: *Head Integration*
New Media Development

… plus all the engineering challenges of scaling dimensions for > 1 Tbit/in² !

P. Frank & R. Wood, presented at APMRC2006

Solid State Drives (SSD) -- An Alternative to HDDs

- Solid State Drives are just a collection of FLASH memory chips (like those in a USB Thumb Drive) arranged in the format of a HDD.
- They are robust and quiet because there are no moving parts
- Reliability not clear – backups still advised
- They are fast – if they use a fast interface like SAS or SATA
- The present max capacity is similar to that of a 2.5” HDD at ~250 GB
- BUT they are very expensive – now at least 10x as much for the same amount of storage as a HDD, and even higher for the largest capacities



A Few Comments on RAID

Redundant **A**rrays of **I**nexpensive **D**rives

(**I**ndependent)

This ↗ Not this

There are several types of RAID giving different trade-offs of protection against data loss, capacity, and speed.

JBOD and RAID 0 discussed below are usually included in RAID discussions but are not really RAID in that they are not redundant.

- **JBOD** – Just a Bunch Of Drives – it's a disk array but without redundancy.
- **RAID 0** distributes data across several disks in a way that gives improved speed at any given instant. If one disk fails, however, all of the data on the array will be lost. Using RAID 0 to increase speed is no longer much of an issue because of the speed of newer drives, especially for home users.

RAID levels 1 and 5 are the most commonly found, and cover most requirements for homes and small offices.

- **RAID 1** mirrors the contents of the disks, essentially a real time backup. The contents of each disk in the array are identical to that of every other disk in the array. This differs from simple backups in that **the data is written to both drives at the same time**. This is a good simple approach for homes or small businesses.

- **RAID 5** (striped disks with parity) combines three or more disks in a way that protects data against loss of any one disk. The storage capacity of the array is reduced by one disk. This is a good approach for small businesses (or a city government). It is very economical in that N+1 drives can store N drives worth of data.

- **RAID 10** (or 1+0) uses both striping (Raid 0) and mirroring Raid 1). "01" or "0+1" is sometimes distinguished from "10" or "1+0": a striped set of mirrored subsets and a mirrored set of striped subsets are both valid, but distinct, configurations.

EMC Symmetrix RAID Array

84 Hard Drives shown here.

Some enclosures are filled top-to-bottom with drives. RAID implementations also have redundant servers and power supplies.

Exactly how firms like EMC implement redundancy may be a trade secret, and you most likely need a Ph.D. in reliability theory to under-stand it. It’s probably some-thing like Redundant Arrays of Redundant Arrays .

RAID can vary from Two Laptop Size Drives to 480 Desktop Size Drives

The two-drive array above is appropriate for home use. On some desktop computers RAID can be implemented internally with two or more drives.

This array would cost you as much as a very nice home in Lexington.

Google Drive Arrays

Google has several complexes around the country (and perhaps the world) of multiple buildings each with the area of a football field full of the servers and RAID arrays. They locate them near places where they can buy electricity cheaply. The one above is on the Colombia River in Dallas Oregon near a hydroelectric plant.

Typical File Storage Requirements

A typewritten page	2 kilobytes
A low-resolution photograph	100 kilobytes
The range for typical PDF files	100 to 800 KB
A short novel	1 megabyte
The contents of a 3.5 inch floppy disk	1.44 megabytes
A high-resolution photograph	2 megabytes
An MP3 (music) downloadable file	3 to 5 MB
The complete works of Shakespeare	5 megabytes
A video or audio downloadable file	500 KB to 10 MB
A minute of high-fidelity sound	10 megabytes
One meter (or close to a yard) of shelved books	100 megabytes
The contents of a CD-ROM	500 megabytes
A pickup truck filled with books	1 gigabyte
The contents of a DVD -- A short Std. Def. Movie	4.7 gigabytes
A collection of the works of Beethoven	20 gigabytes
A library floor of academic journals -- Highest reported BlueRay disc capacity 2009	100 gigabytes
50,000 trees made into paper and printed	1 terabyte
An academic research library -- Highest 2009 Hard Drive Capacity	2 terabytes
The print collections of the U.S. Library of Congress	10 terabytes
All U.S. academic research libraries	2 petabytes
All hard disk capacity developed in 1995	20 petabytes
All printed material in the world	200 petabytes
Total volume of information generated in 1999	2 exabytes
All words ever spoken by human beings	5 exabytes

Not only is a picture worth a 1000 words, it looks like it takes 1000x as much storage space.

Pictures, music & video/movies take substantially more storage space than text.

That's only a million 2-tera-byte drives; this is well below the production runs of almost all HDD models.

A note on doubling capacity

Why multiple short time periods between doubling the storage density of HDDs has resulted in a million-fold increase in storage capacity.

Chess board story – As an award the inventor of Chess asked the king for one grain of wheat on square one of the chessboard, two on square two, four on square three, etc. Up to 2^{63} grains on the last square. The total after the board is filled is ~1.84 followed by 19 zeros which is approximately 46 times the total wheat production of the world in 2007.

$$1 + 2 + 4 + 8 + \bullet\bullet\bullet + 2^{N-1} = 2^N - 1$$

$$2^0 + 2^1 + 2^2 + 2^3$$

Square No.	Grains on Square	Sum up to this square	
1	1	1	
2	2	3	
3	4	7	
4	8	15	
8	128	255	Row 1
16	32,768	65,535	Row 2
24	8.39E+06	1.68E+07	Row 3
32	2.15E+09	4.29E+09	Row 4
40	5.5E+11	1.1E+12	Row 5
48	1.41E+14	2.81E+14	Row 6
56	3.6E+16	7.21E+16	Row 7
64	9.22E+18	1.84E+19	Row 8

How long can the rapid increase in HHD capacity continue? Not too much longer, but there are other techniques in the wings.

METHOD	Tb/in^2	COMMENTS
Present HDD (2009)	0.5	Presently available using perpendicular recording and TMR readers
HDD Limit	2 to 5	With patterned media and perhaps HAMR
Nano ferite particles in nano tubes	10 to 20	U.S. DoE Berkeley National Laboratory, UC Berkeley and UMass Amherst
Small cluster of atoms on substrate	500	Demo'ed by U of Wisc. About 20 gold atoms per bit.
Quantum holography demo	3000	Highest density record per Wikipedia. Demo'ed by Stanford.

A Tbit (Terabit) is 1000 Gbits (Gigabits).

A Pbit (Petabit) is 1000 Tbits (Terabits).

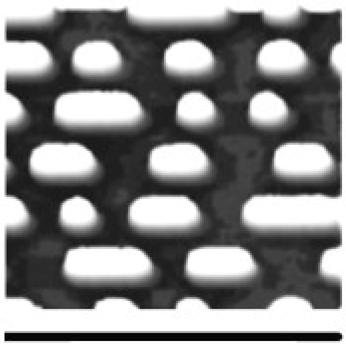
These very high density techniques are painstakingly slow.

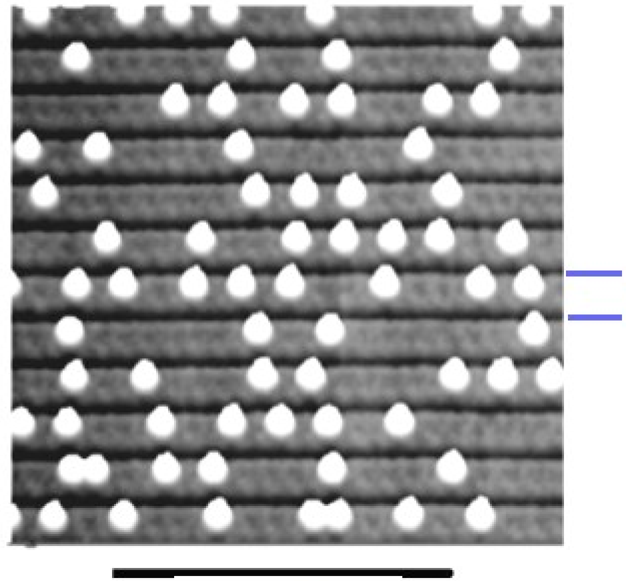
THERE'S PLEANLY OF ROOM AT THE BOTTOM – Richard Feynman

In Pursuit of the Ultimate Storage Medium :

1 Bit = 1 Atom

Clusters of gold atoms on a silicon substrate. Univ. of Wisc.





Density × 1 000 000

This technique can store 1 million times more data than a CD can on the same area.

From: http://uw.physics.wisc.edu/~himpsel/talktech.pdf