Introduction to Hard Disk Drives

(What goes on inside your hard drive and more)

Larry Wittig

Lexington Computer and Technology Group Meeting

Original: 7 October 2009

Updated: 5 December 2012

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

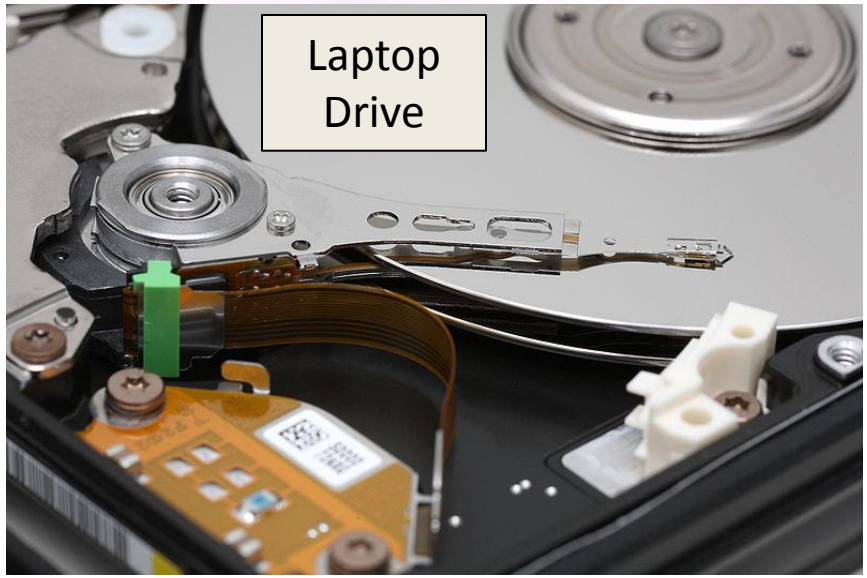
Modern HDDs

- Mechanical overview
- Read/Write Recording Heads
- Disks and Servo
- Actuator and Spindle motor
- Electronics and interfaces
- HHD advances in the works
- Solid State Drives (SSD)
- RAID
- How long can the improvements in HDD technology continue?

A Hard Disk Drive is a somewhat of a cross between a tape recorder and a record player

(Also it's digital as opposed to analog)

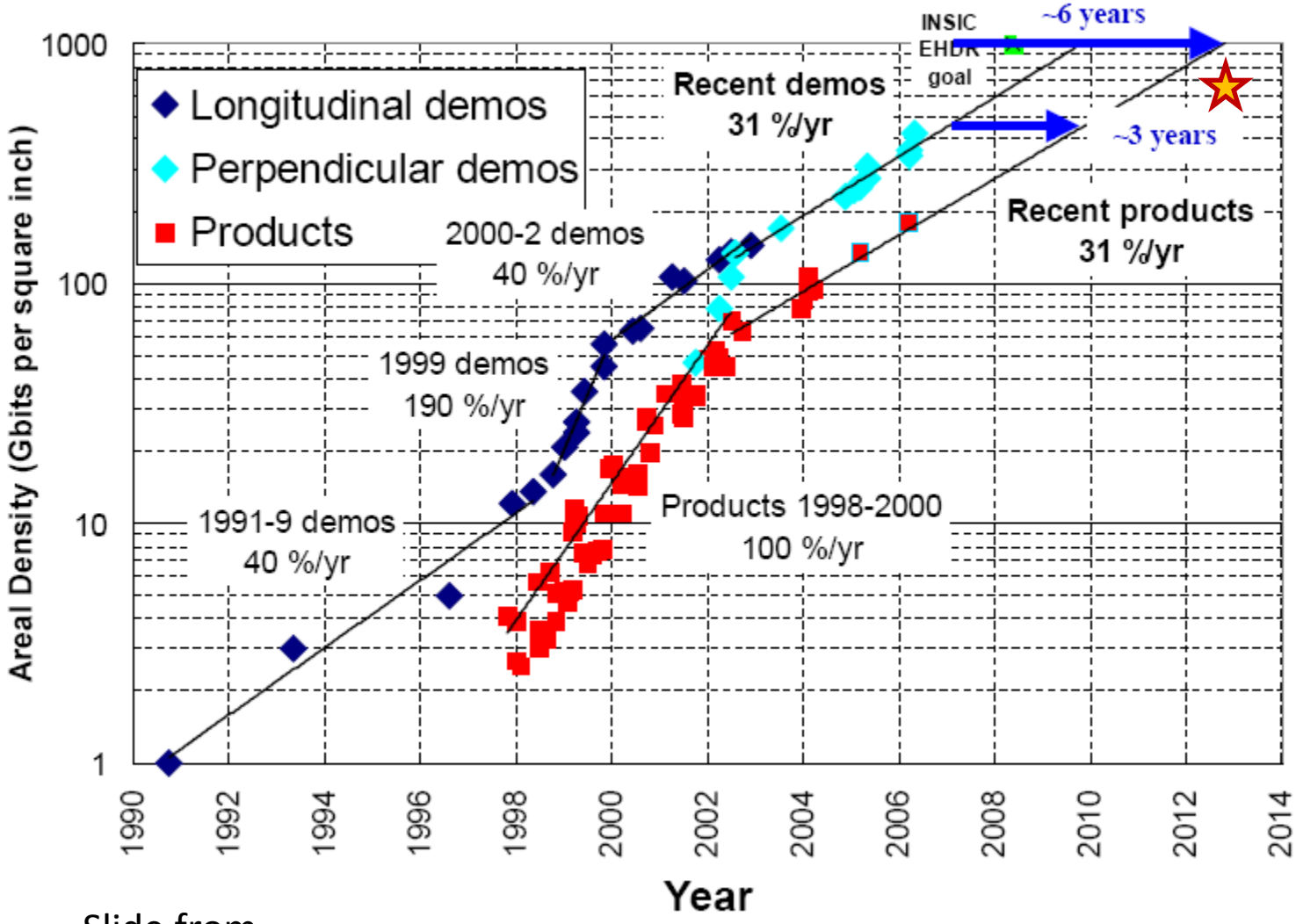
Desktop Drive



Laptop Drive

Mid 2012 ~ 600 Gbits/in^2. Approx. 400k Tracks/in x 1.7 Mbits/in. That's about 1000 tracks in the width of a human hair. That's ~1 TByte/3.5" disk ≈ 3 TB for a 3-disk HDD.

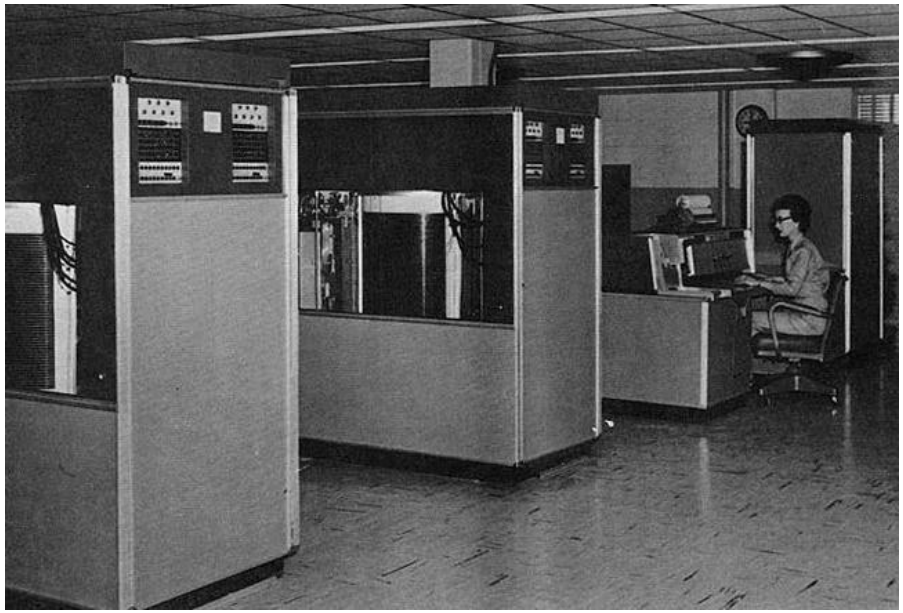
HDD Areal Density Trends – Demos & Products

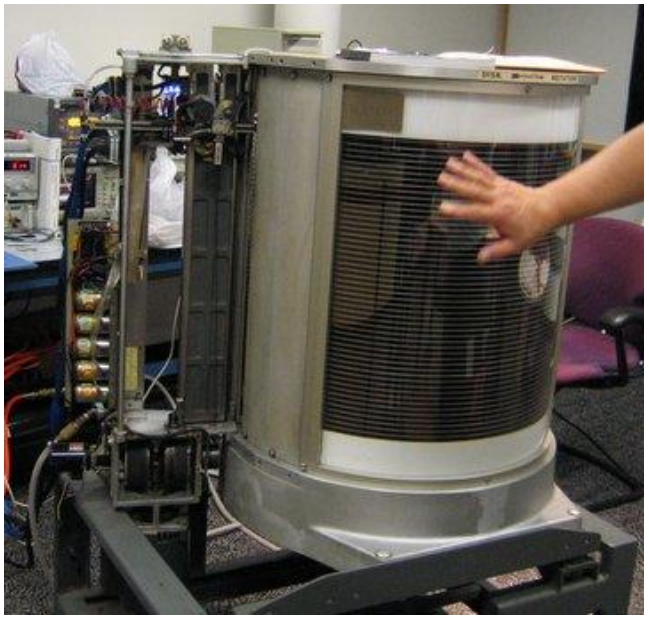


Slide from

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 3 TB desktop HDD stores 600,000 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 (or about $1.3M in 2012 dollars).





Size Progression of HDDs

From about 1980 to Present

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 (independent) 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.

  There are additional slides later on about RAID.

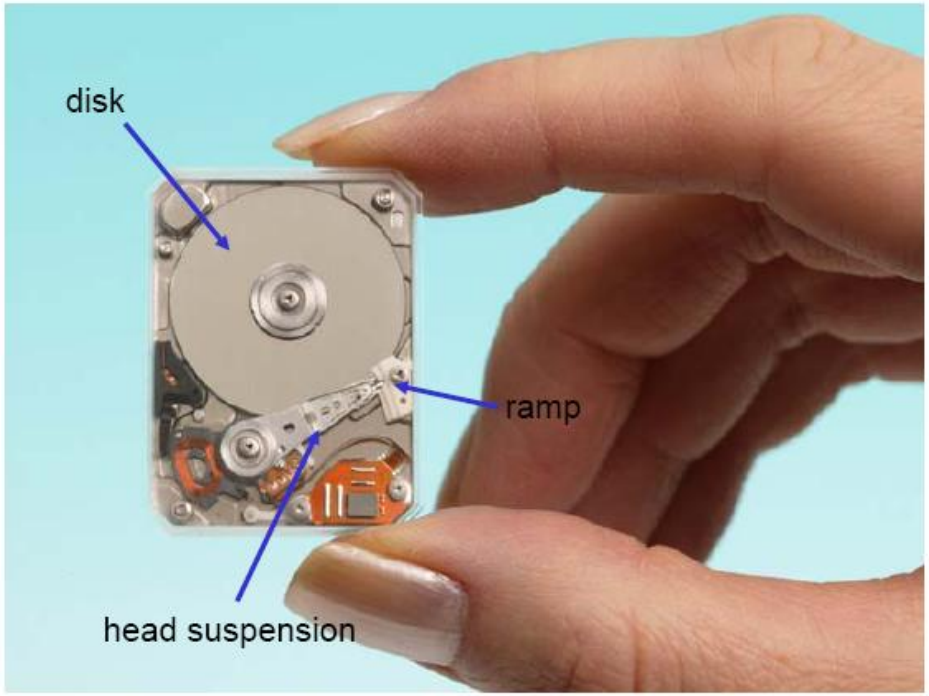
Micro Drives Have Not Succeeded Very Well Against Recent Advances in USB Flash Drives

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.

Which drive holds more data?

Which drive holds more data?

The small one holds about 4000 times more data then the large one.

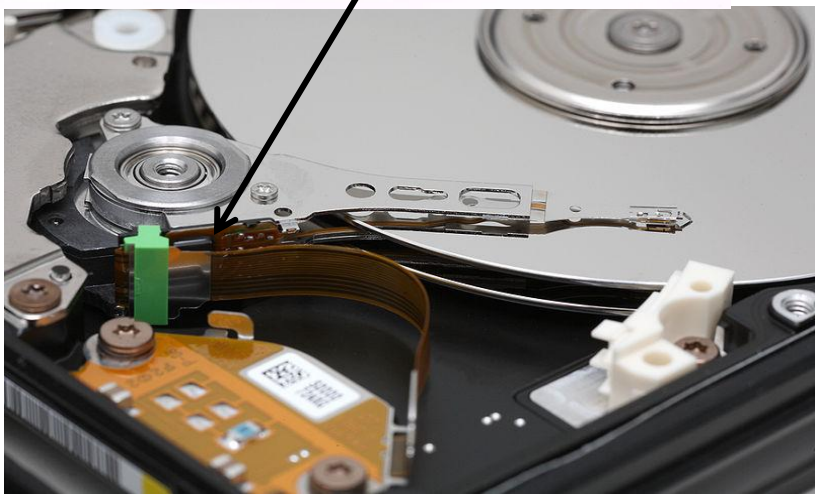
Back to what goes on inside your Hard Drive

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 <u>only one head in the drive is working at any one time</u>.

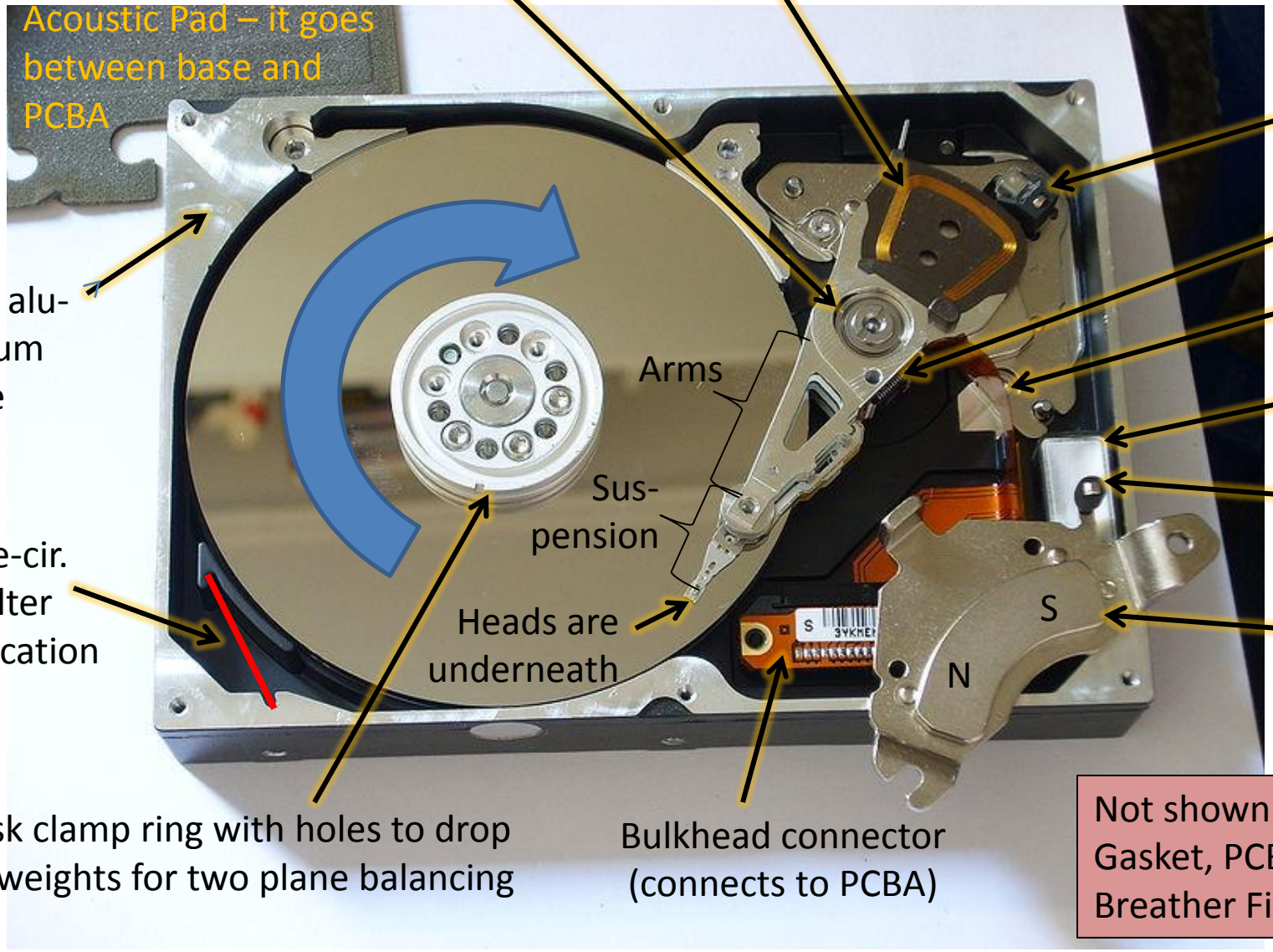
PCBA

"Stators" to reduce air turbulence.

Pre-Amp at actuator hub

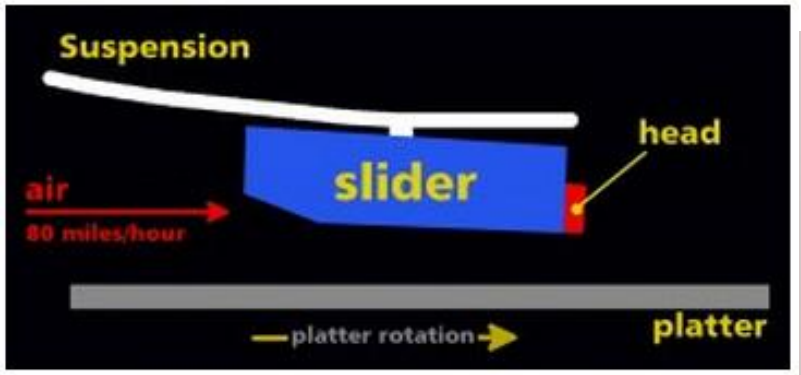


What's Inside: Rotary Actuator and Other Parts



Arm Tip, Suspension, Slider

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



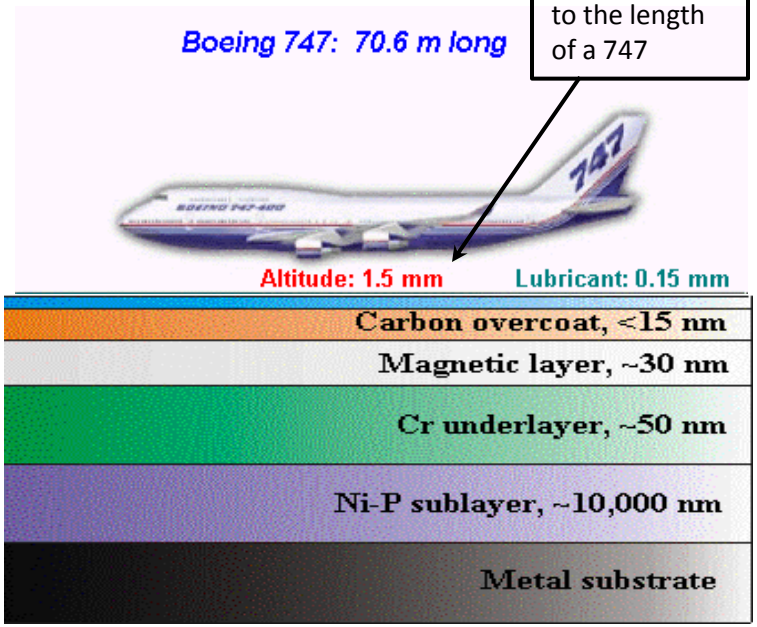
Close-up of Suspension

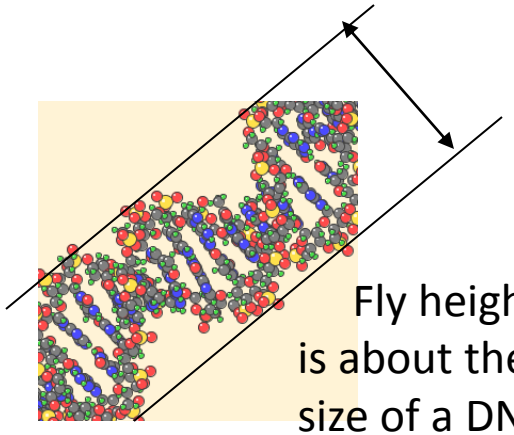
Comments on Fly Height

Now days (~2012) the head end of the slider "flys" at ~3 nm above the disk. That's about 4/100,000th of the diameter of a human hair or the width of a DNA molecule.

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.

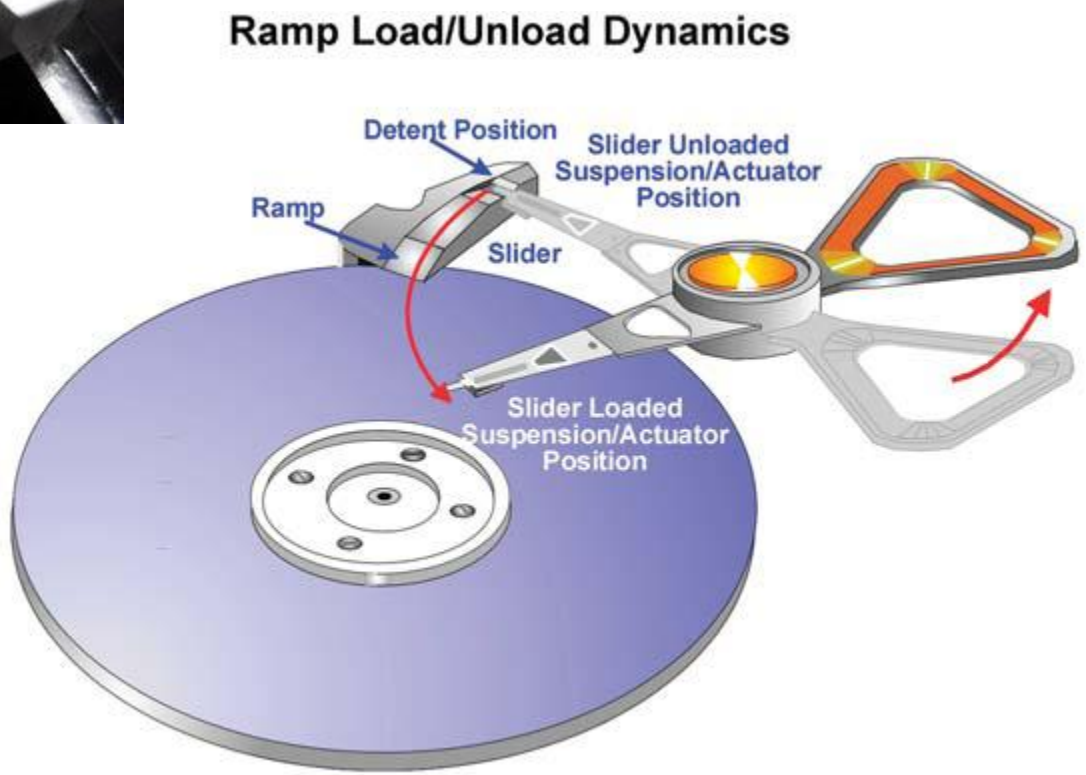




Fly height is about the size of a DNA molecule

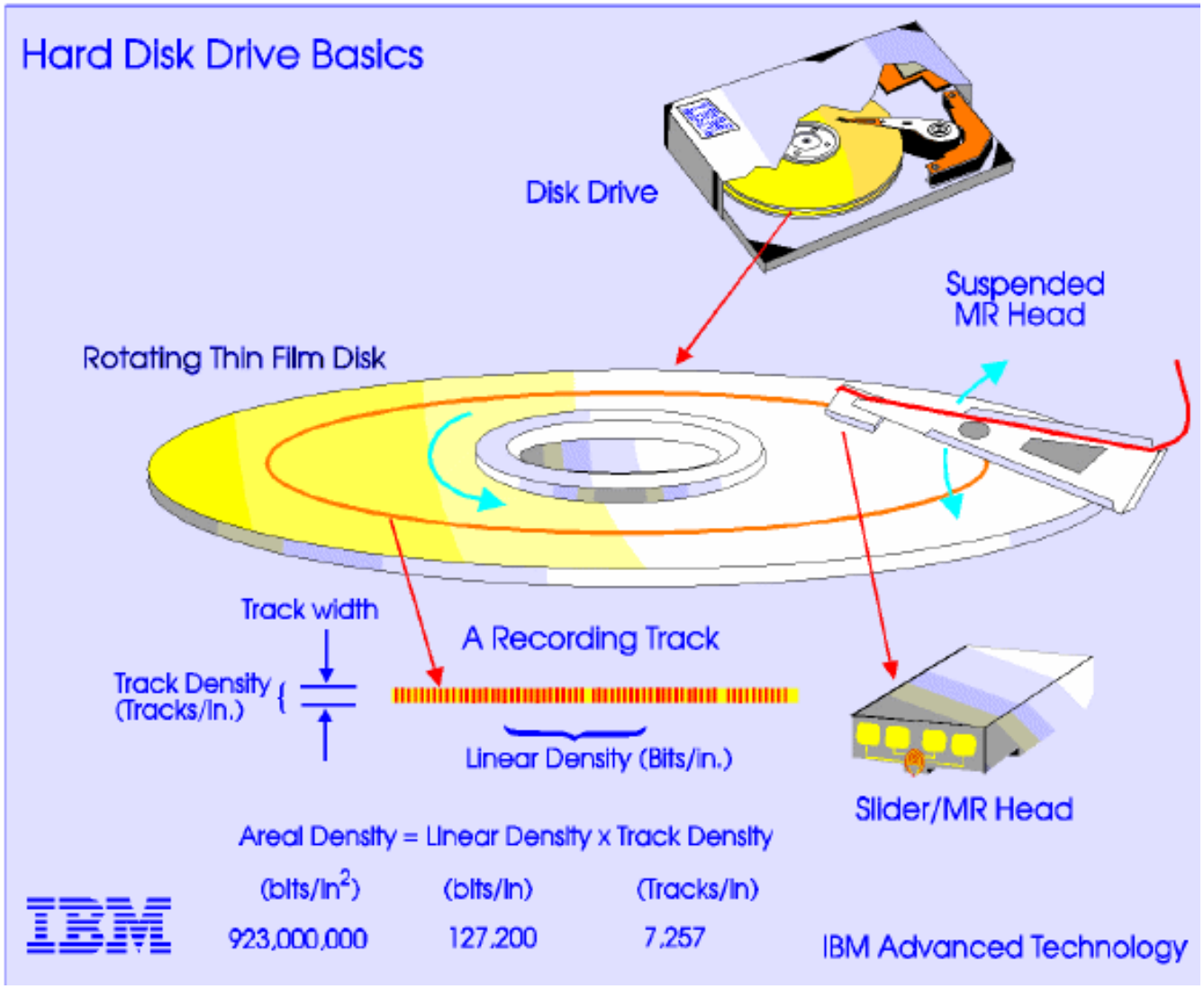
Ramp Load Allows for Smoother Disks (lower fly height) and Provides Shock Protection

Initially Ramp Load/Unload was used in Laptop computers for shock protection. But now all drives use it because it allows for smoother disks and lower fly height.



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

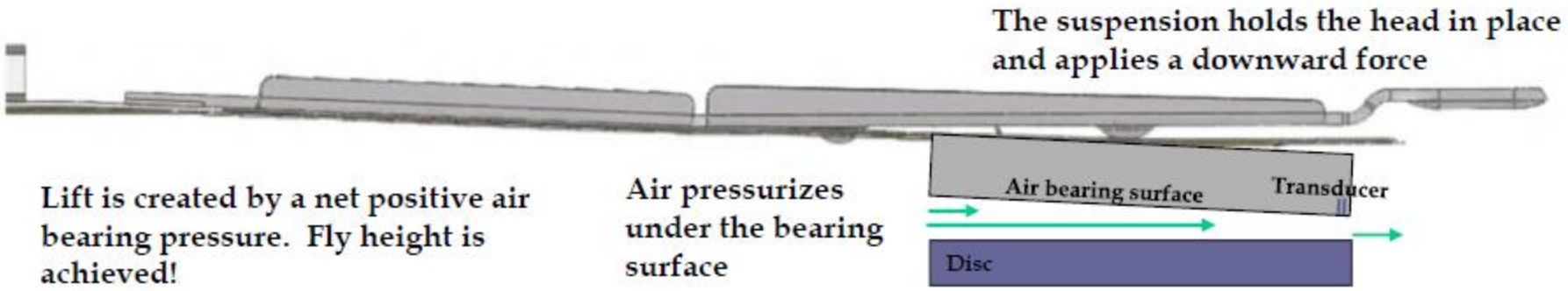
How Data is Stored and Recalled

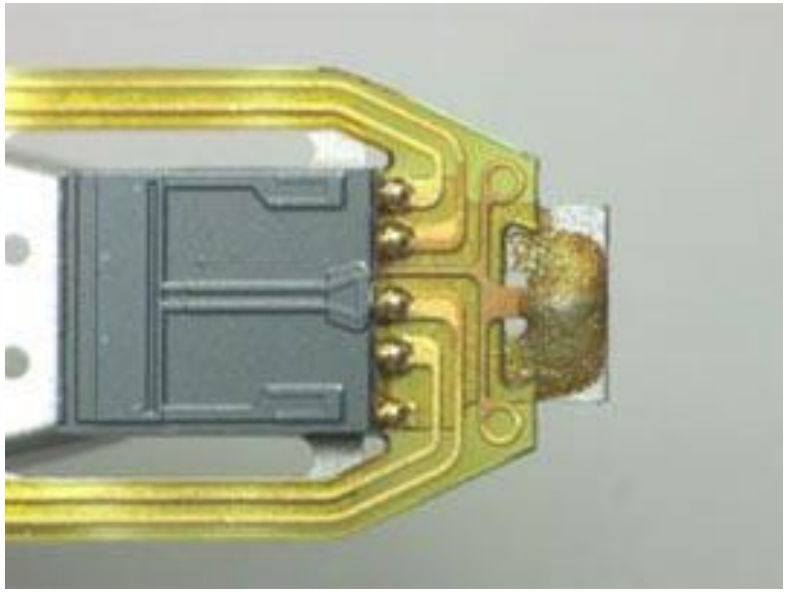


Now days (2012) the linear data density is about 15x as high as shown here resulting in the ability to store about 2 MBytes per revolution.

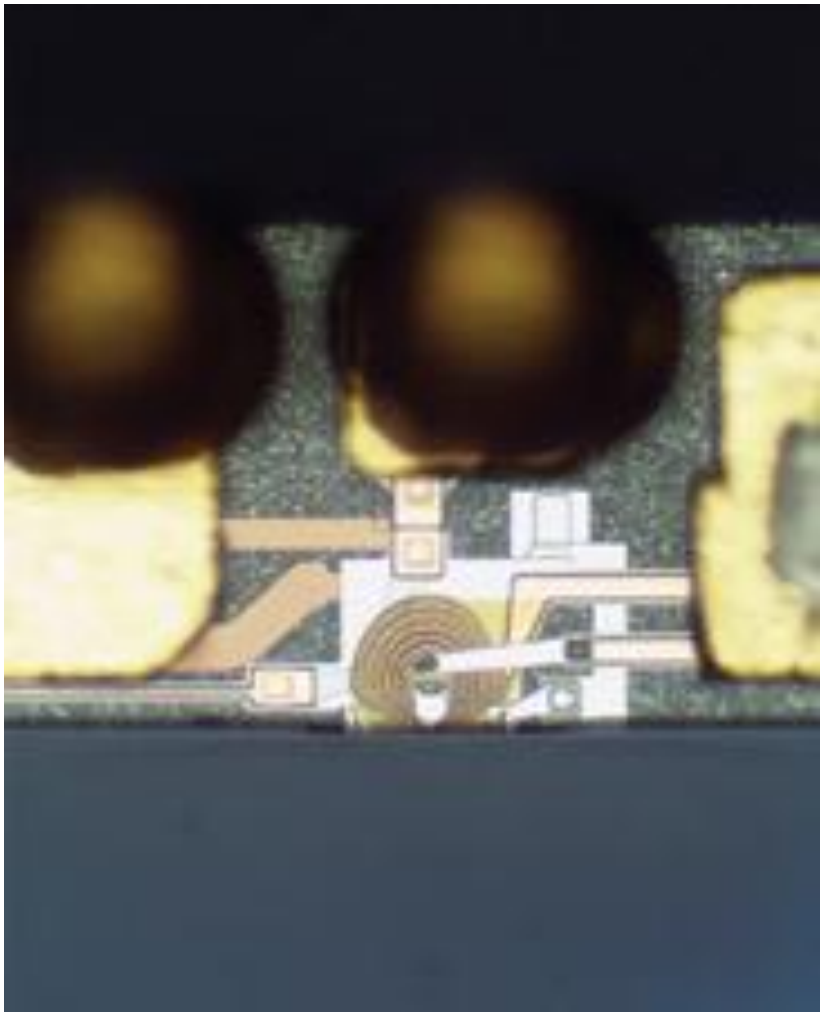
A 200 kB JPEG picture takes ~1/10th of a rev. and a type-written page takes about ~1/100th of a rev.

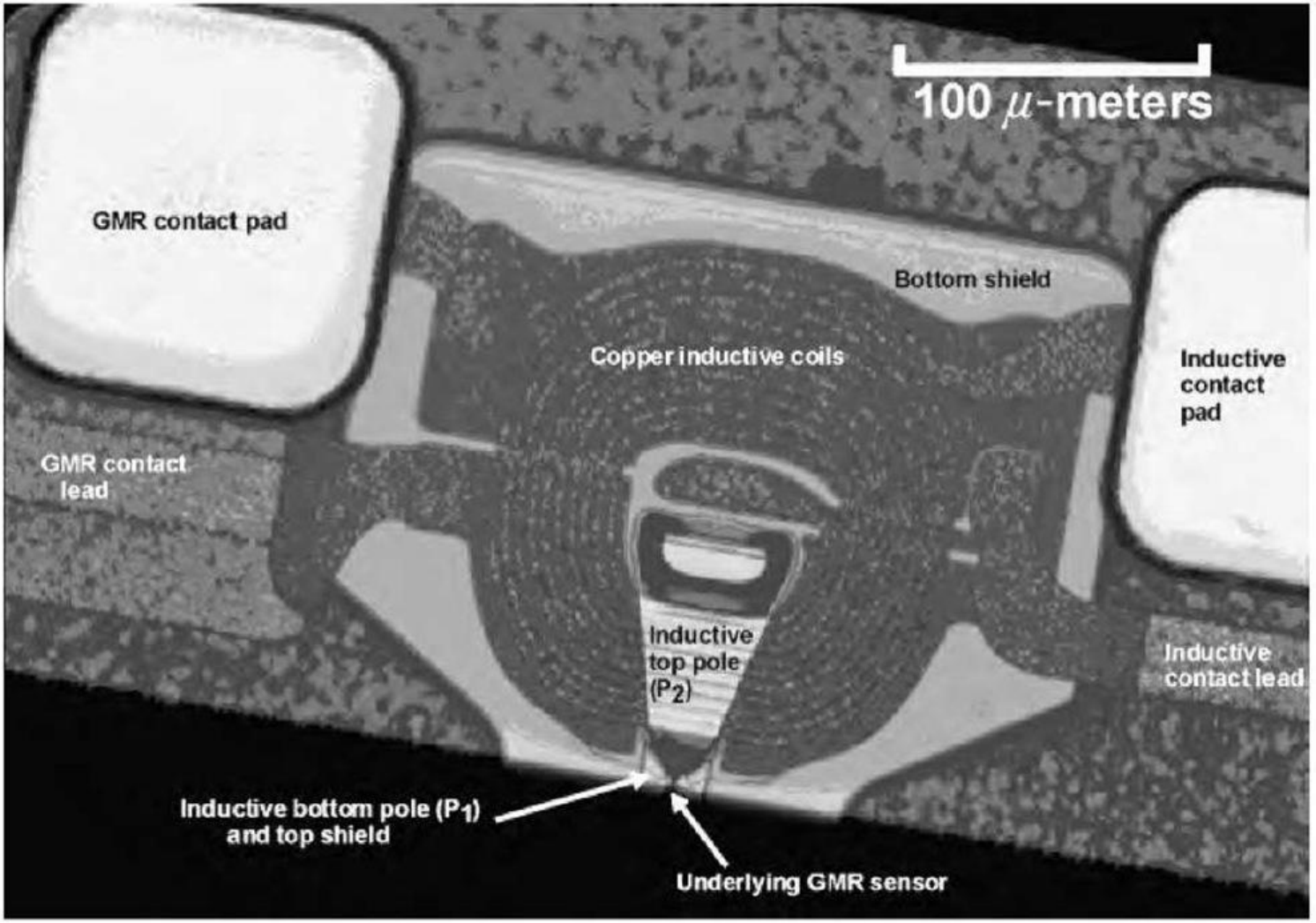
The Suspensions is like a leaf spring that pushes the Slider against the disk





Sliders and Heads




100 μ-meters
GMR contact pad
Bottom shield
Copper inductive coils
Inductive contact pad
GMR contact lead
Inductive top pole (P_2)
Inductive contact lead
Inductive bottom pole (P_1) and top shield
Underlying GMR sensor

Major Head 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

HITACHI
Inspire the Next

Evolution of Magnetic Read/Write Sensors

>2003
Perpendicular Thin Film
Inductive Write
CPP Read Head

1997

1990

1970

1980

Ferrite Inductive MnFe
Read/Write Head
Wire wound coil
Machined Pole Pieces
Gap Width Controlled
By Films And Assembly
Tolerances

Thin Film Inductive Write
GMR Read Head
Write Wide-Read Narrow
Four Contact Structure
Pinned, Free Films
Antiferromagnetic
Exchange Film
CIP Operation

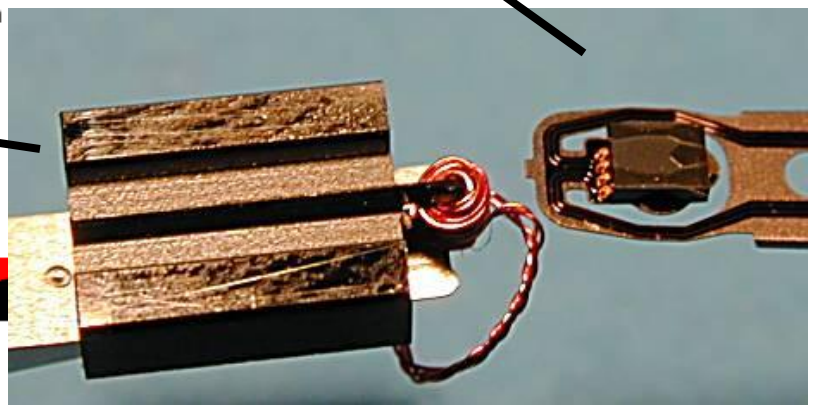
Thin Film Inductive Write
MR Read Head
Write Wide-Read Narrow
Four Contact Structure
SAL
NiFe MR Film

Thin Film Inductive
Read/Write Head
Coil, Pole Geometries
Controlled By Semi-
conductor Type Process
NiFe Poles
Two Contact Structure

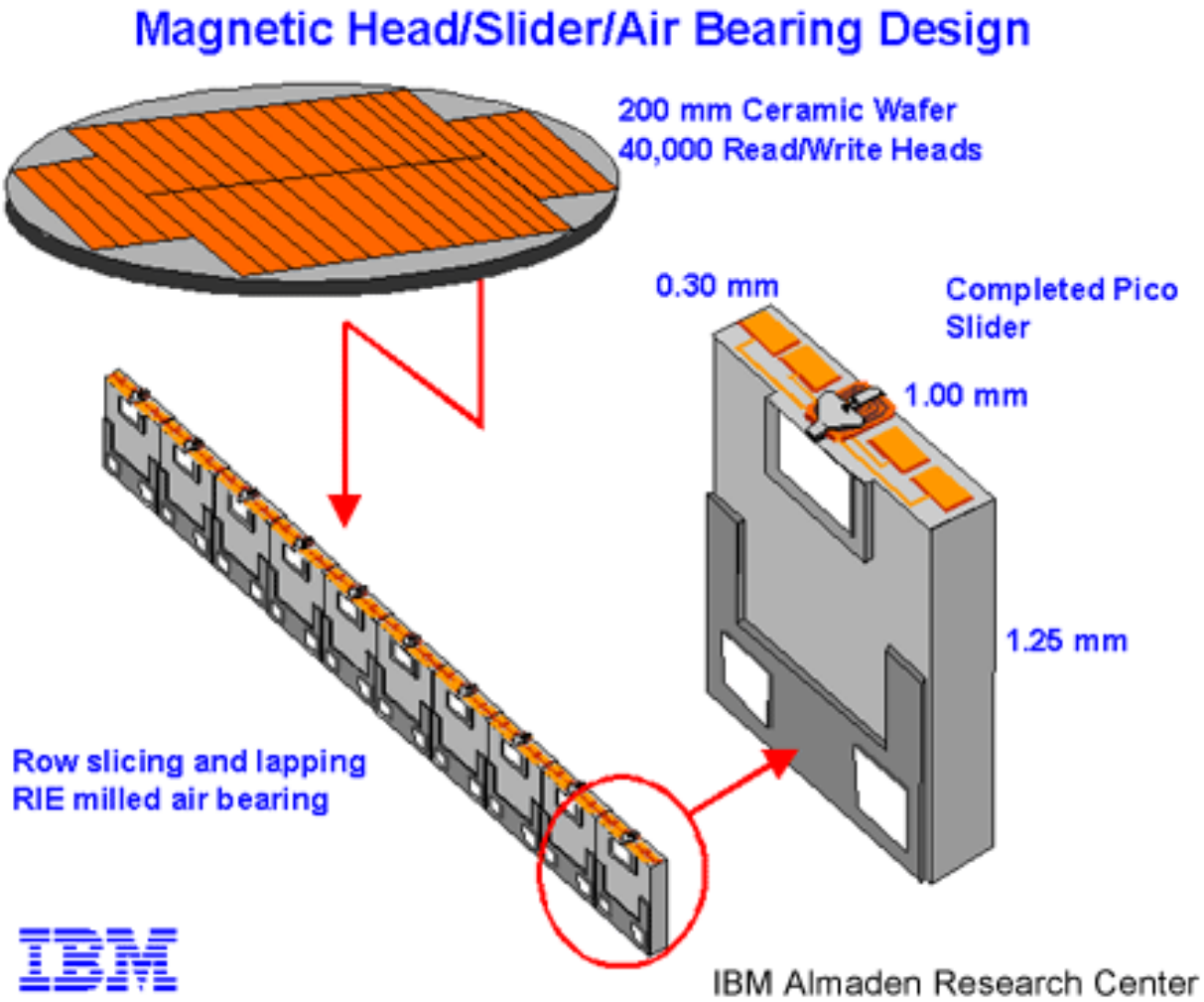
Ferrit2003A.ppt

San Jose Research Center

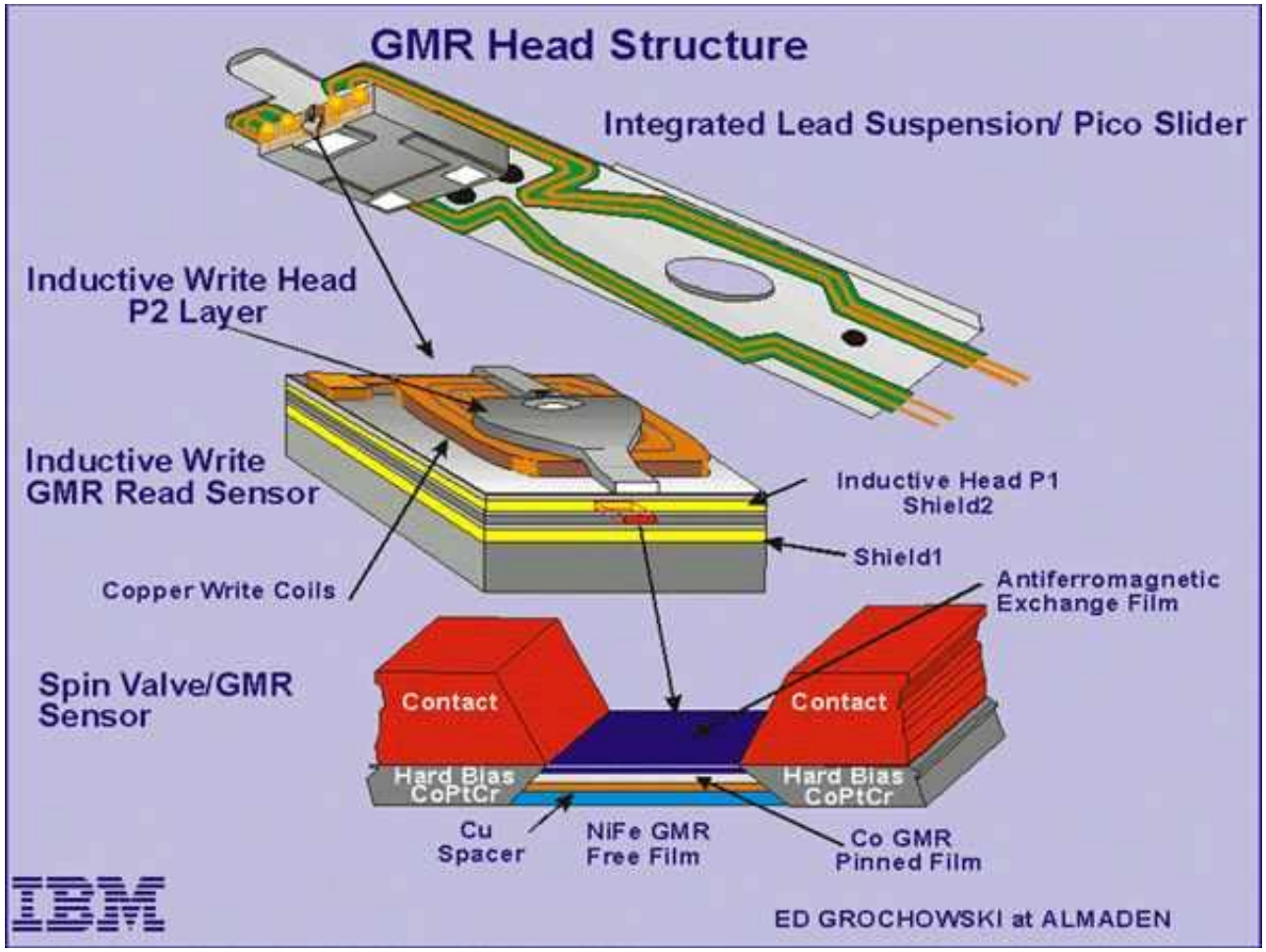
Until about 1990 the same coil was used to both write and read



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



GMR Heads – The IBM Physicists that developed GMR materials won a Nobel Prize -- TMR then replaced GMR

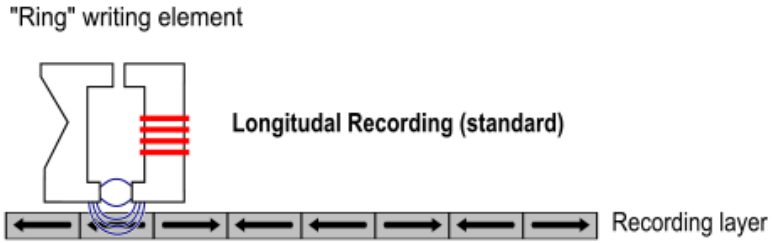


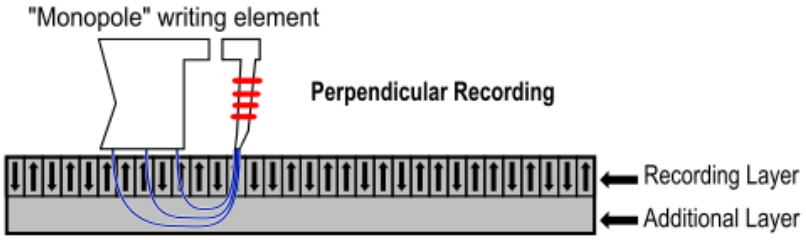
MR = Magneto Restrictive – the resistance of the material changes in the presence of a magnetic field (discovered by Joule and Kelvin ~ 1850).

GMR = Giant MR effect – won 2007 Nobel Prize for work done in 1988. Also used in MRAM.

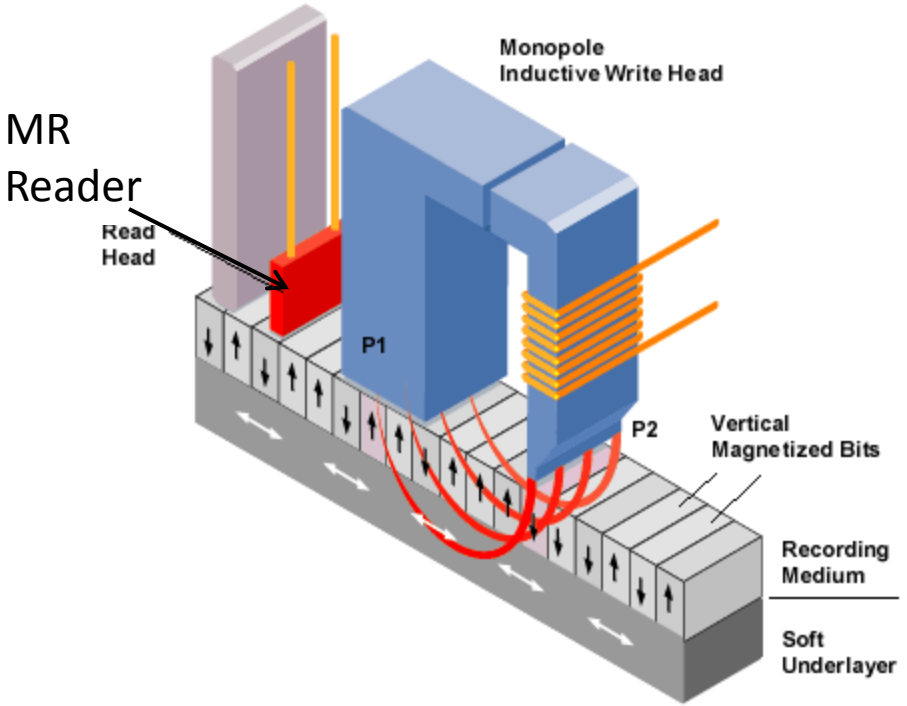
TMR = Tunneling MR

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



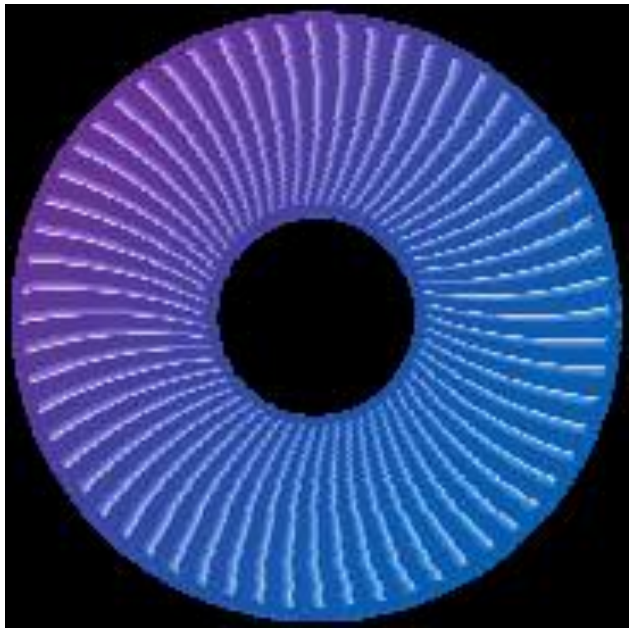


By standing the magnetic domains on end the bit density can be increased.



Go to Kryder Presentation

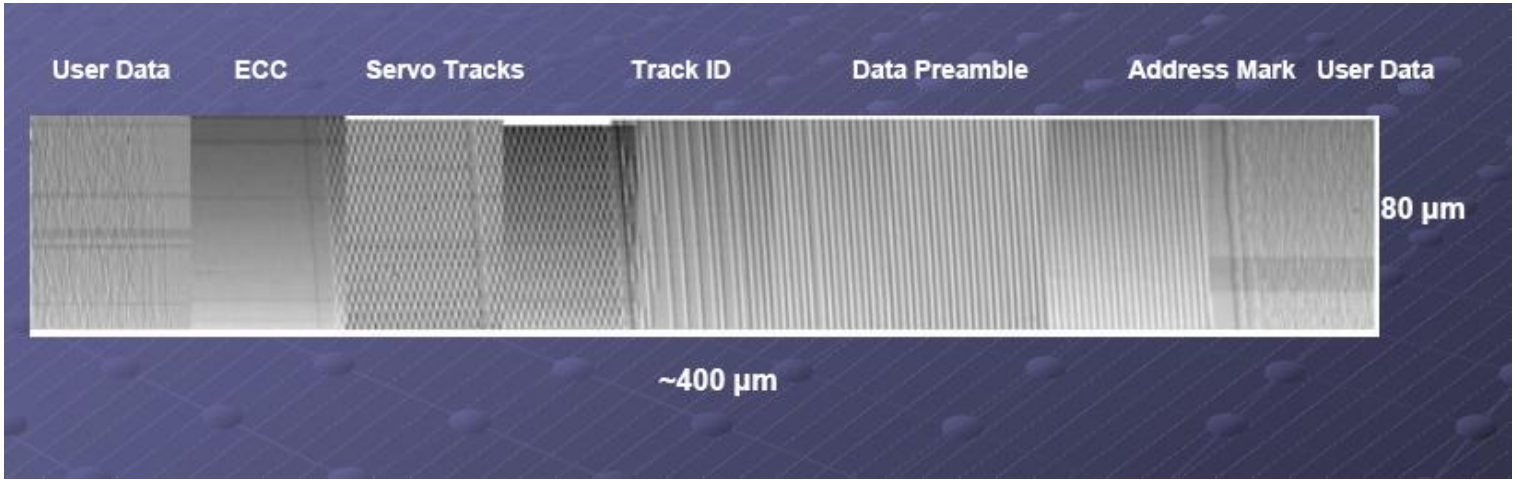
Servo Tracks are used to locate position on the disk



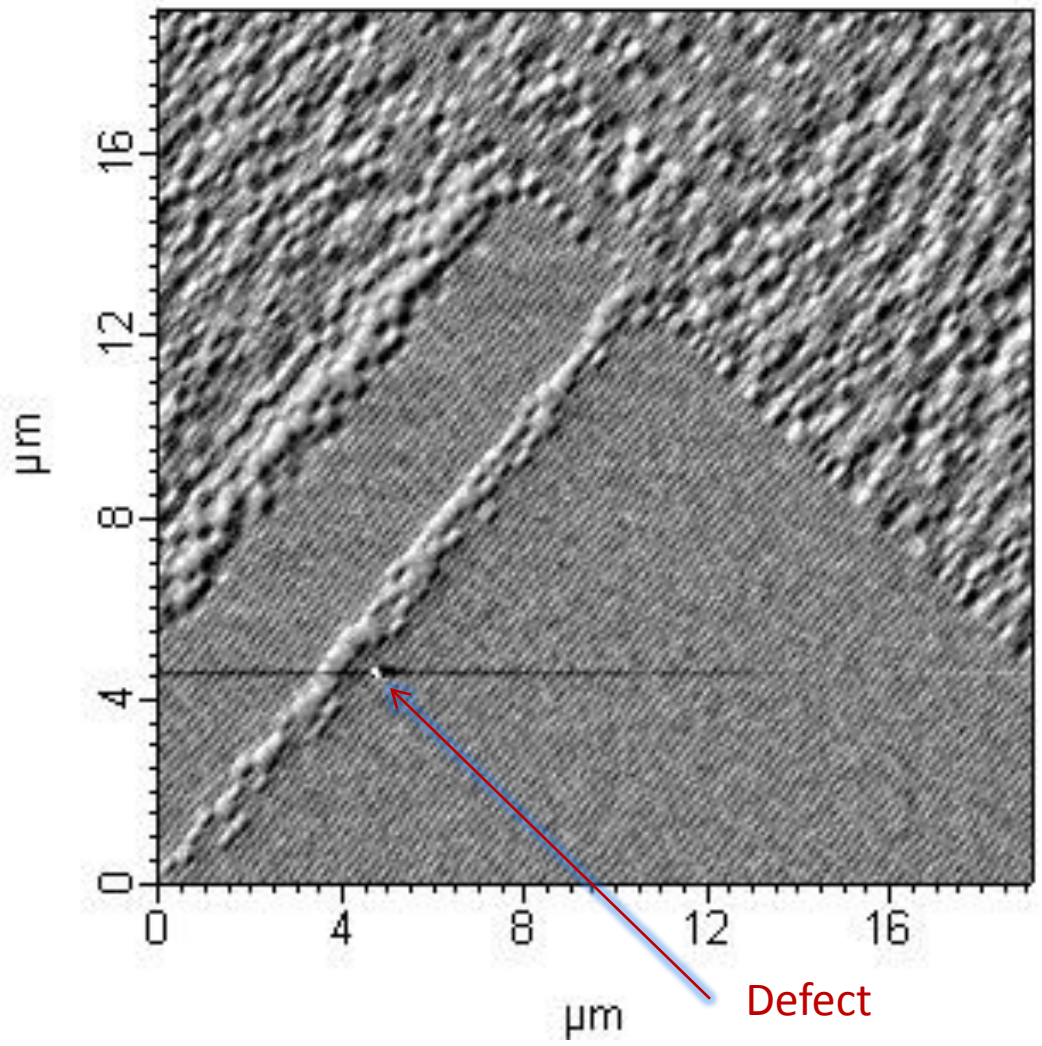
Servo wedges are “written” after the drive is assembled using the same heads that are used to write data. Wedges are used to (1) mark track and sector locations ,(2) provide timing info, and (3) provide “servo bursts” that are used to fine tune staying on track.

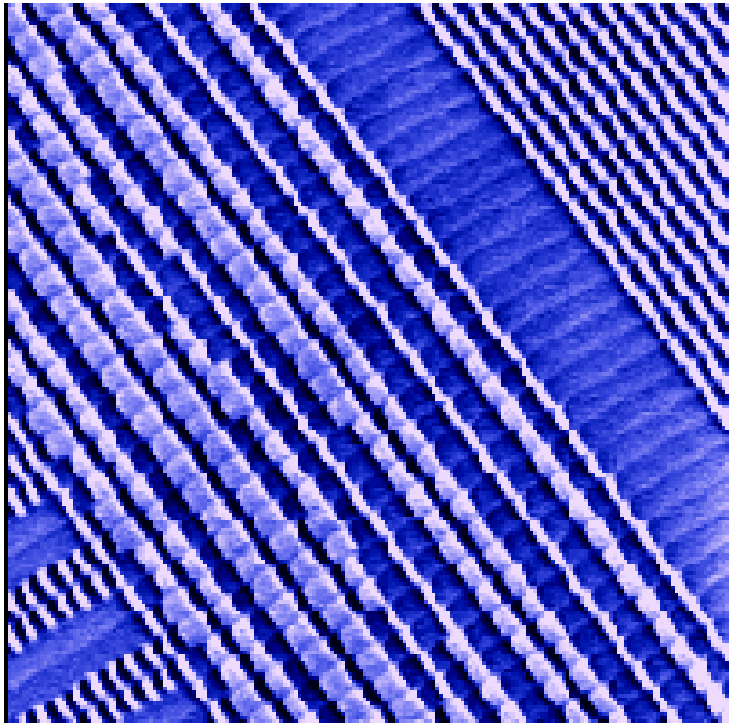
Approximately 75% of the surface is used for storing data. Present disks have ~ 200 wedges. After servo writing drives undergo about 30 hours of testing.

~ 1000 tracks in 2012 density

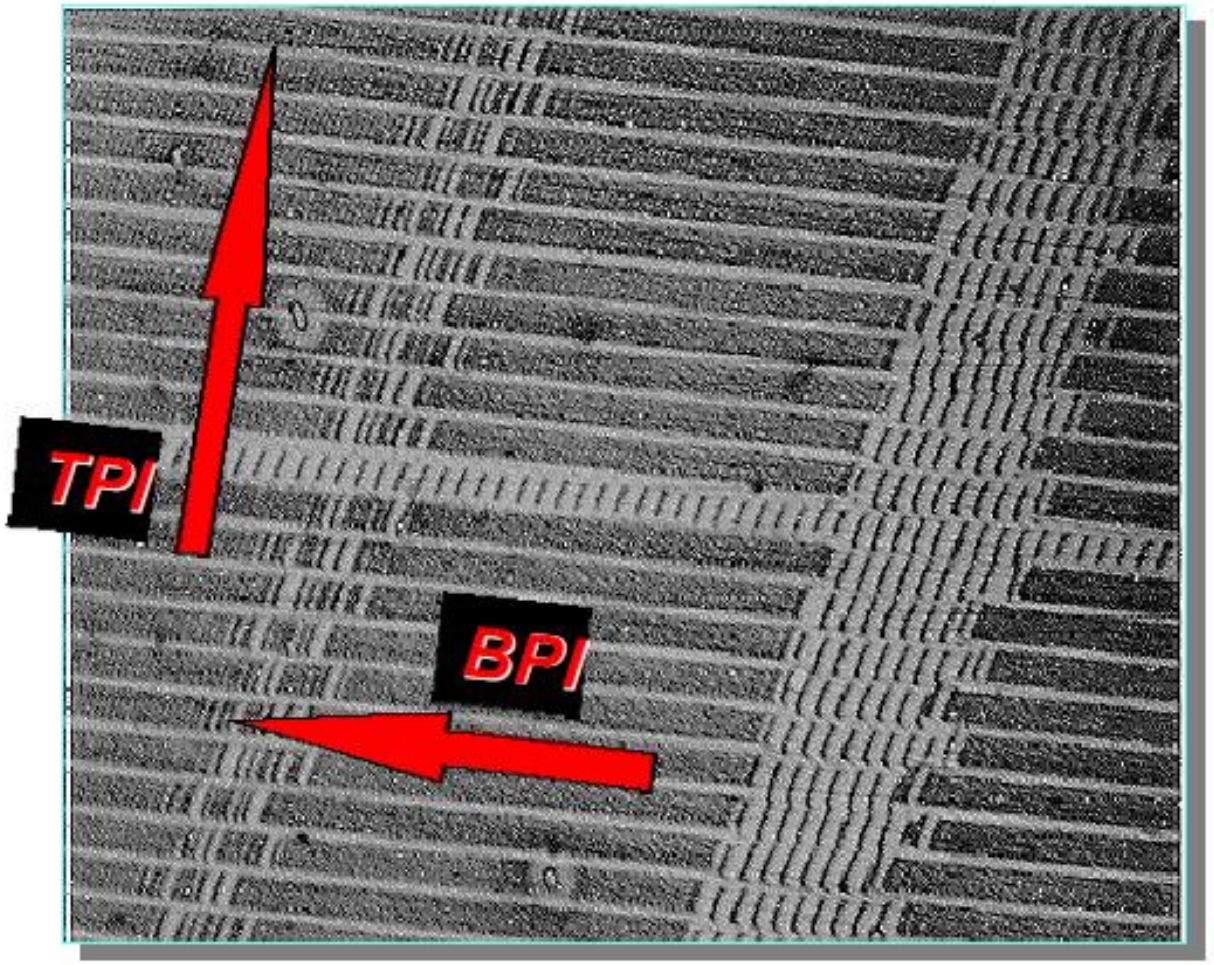


Magnetic Force Microscope Images of a Disk Surface Showing Servo Sector and Data

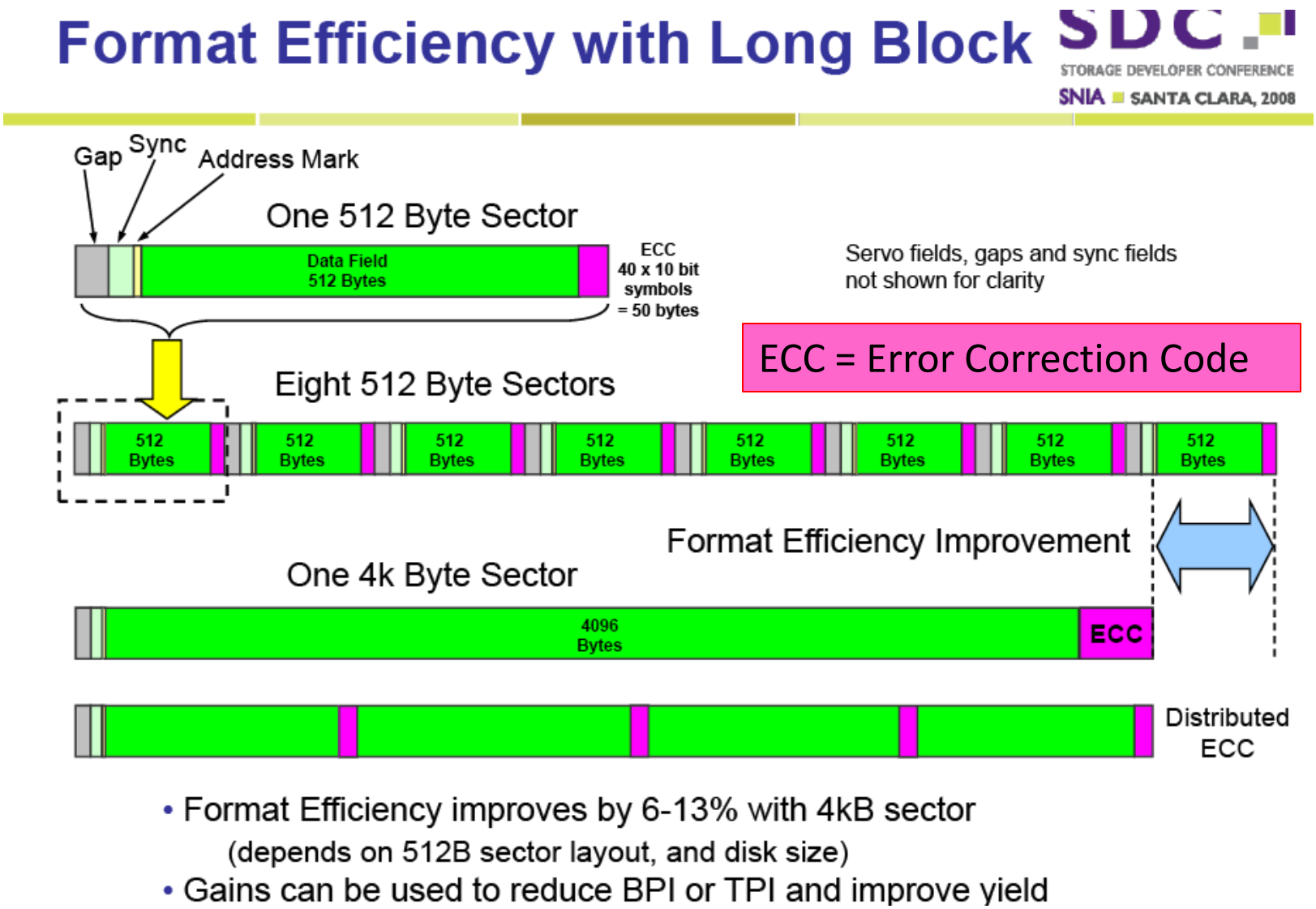




Magnetic Force Microscope Images of a Disk Surface Showing Servo Sector and Data



How the data is formatted between the servo wedges



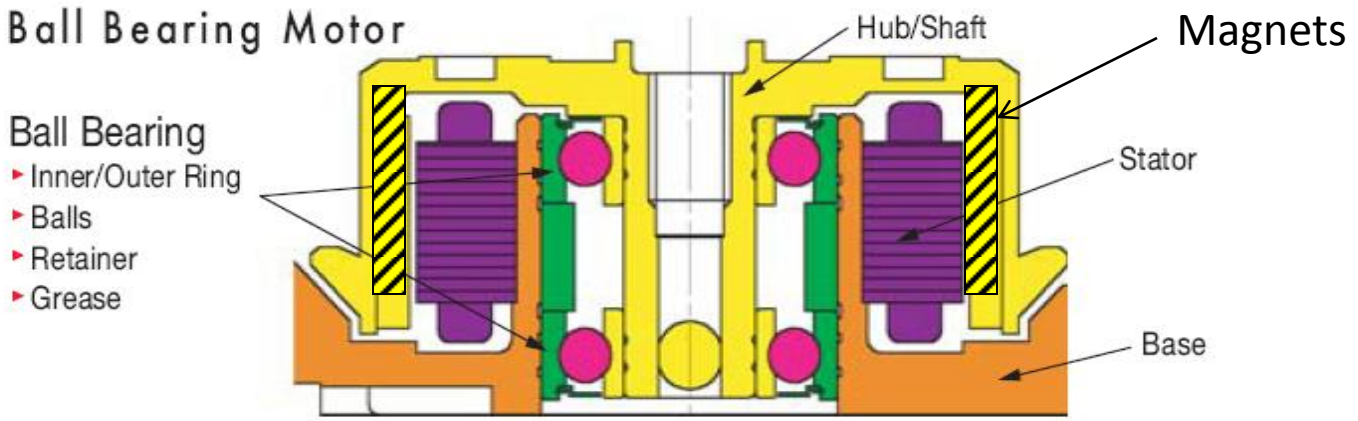
- Format Efficiency improves by 6-13% with 4kB sector
  (depends on 512B sector layout, and disk size)
- Gains can be used to reduce BPI or TPI and improve yield

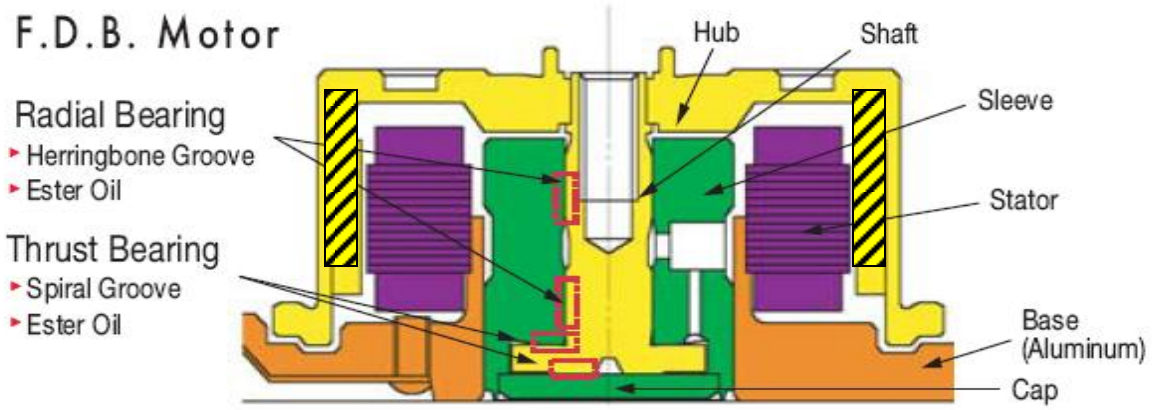
Error Correction Code

- Video on HDD in general, media material, and PRML
- Servo tracks, Gray code & ECC Web Page

Spindle Motor Cross Section

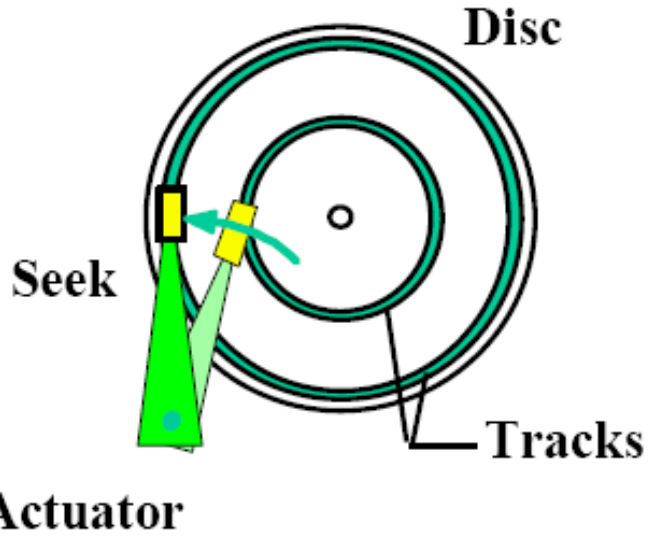
There was a change from Ball Bearing Motors to Fluid Dynamic Motors about 2002 because BB caused too much (off track) vibration. FDB are also quieter and less prone to failure.





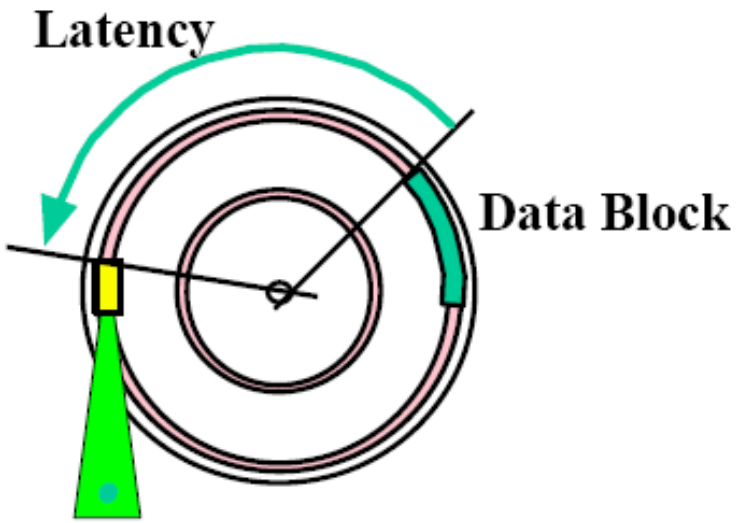
Mechanics: Getting to the Data

Seek Time



Keys: Swing length
Arm length, mass
Magnetic circuit

Latency Time



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

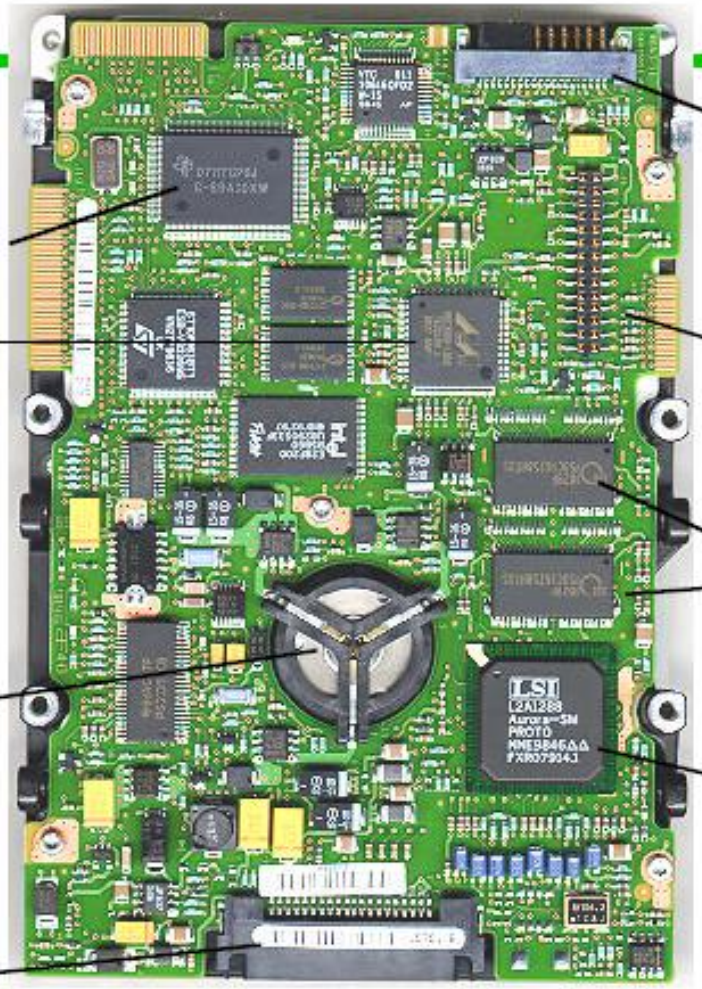
Seagate

Comments on Seek Time and Rotational Latency

- Seek Time is the time to move the actuator from one track to another. Average seek time ranges from about 15 ms on laptop drives down to about 2.5 ms on Enterprise drives. On a new drive 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 a slower rpm and seek time 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

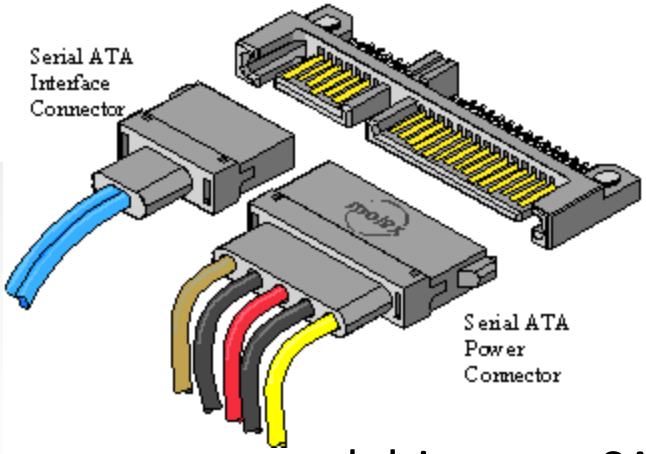
Newer Drives Use SATA

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.



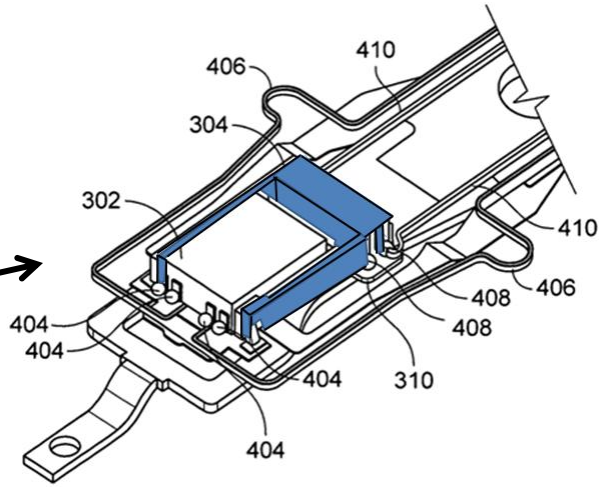
New computers and drives use SATA. A computer with an ATA interface will not accept a SATA drive without a convertor board.

PATA: when SATA was introduce ATA was renamed Parallel ATA

More HDD Advances in the Works

Now in most HDDs

- Micro-actuators
  - Small second actuator located at the slider
- Fly-height adjust
  - Uses a very small heating coil to expand the material near the head

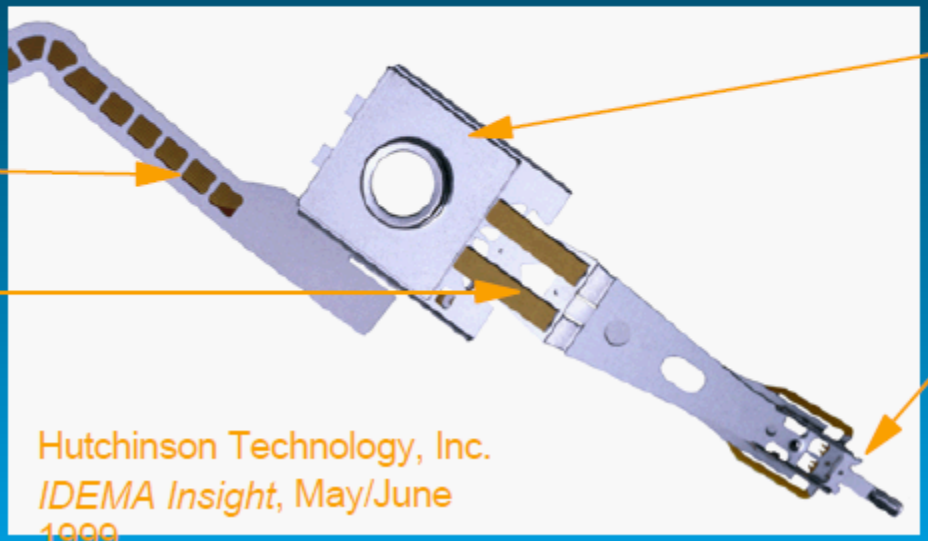


Just around the corner

- 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

Milli-actuator Suspensions

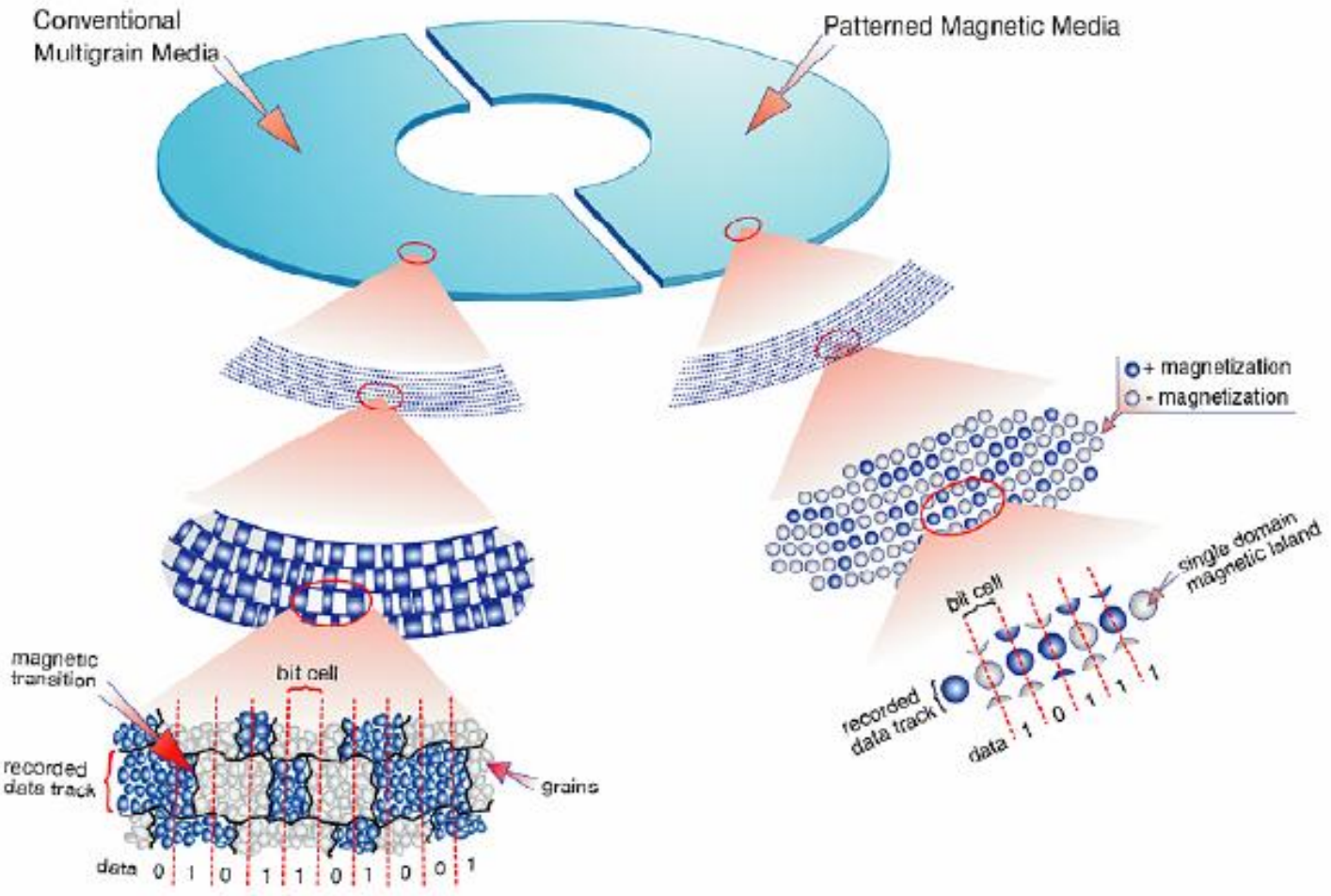
- Piezo "motor" on suspension
- Easier integration than microactuator
- Dynamics not quite as clean as microactuator



Hutchinson Technology, Inc.
IDEMA Insight, May/June 1999

Conventional Media vs. Patterned Media

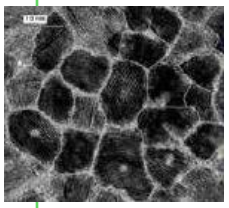
HITACHI
Inspire the Next



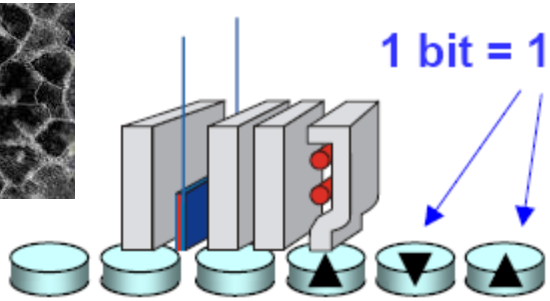
Beyond Conventional Perpendicular Recording

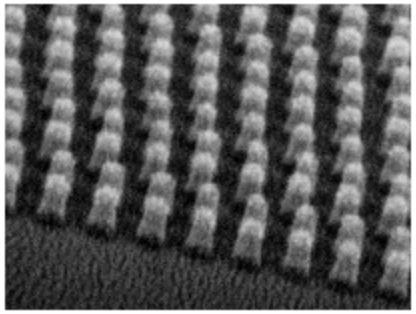
(two favorite technology options to extend thermal limit)

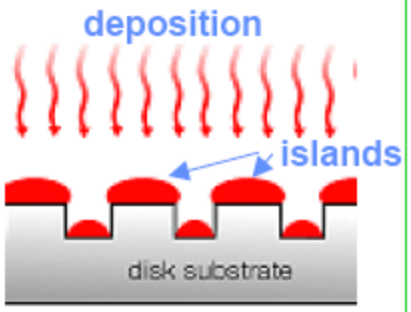
Patterned Media (increased V, utilizing 1 large "grain" per bit)



Now



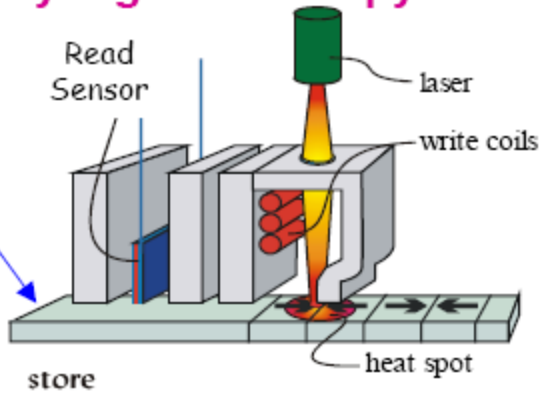


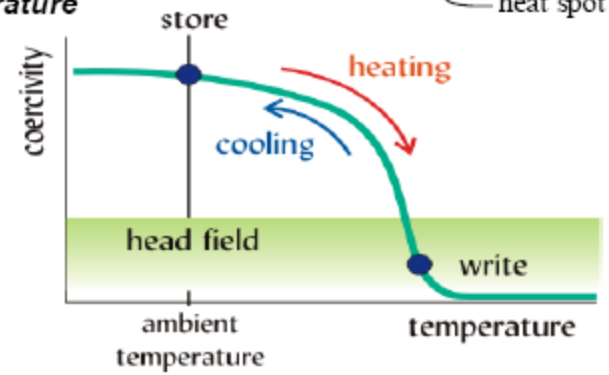


Thermal Assist (increased K_u, utilizing very high anisotropy media)

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 !

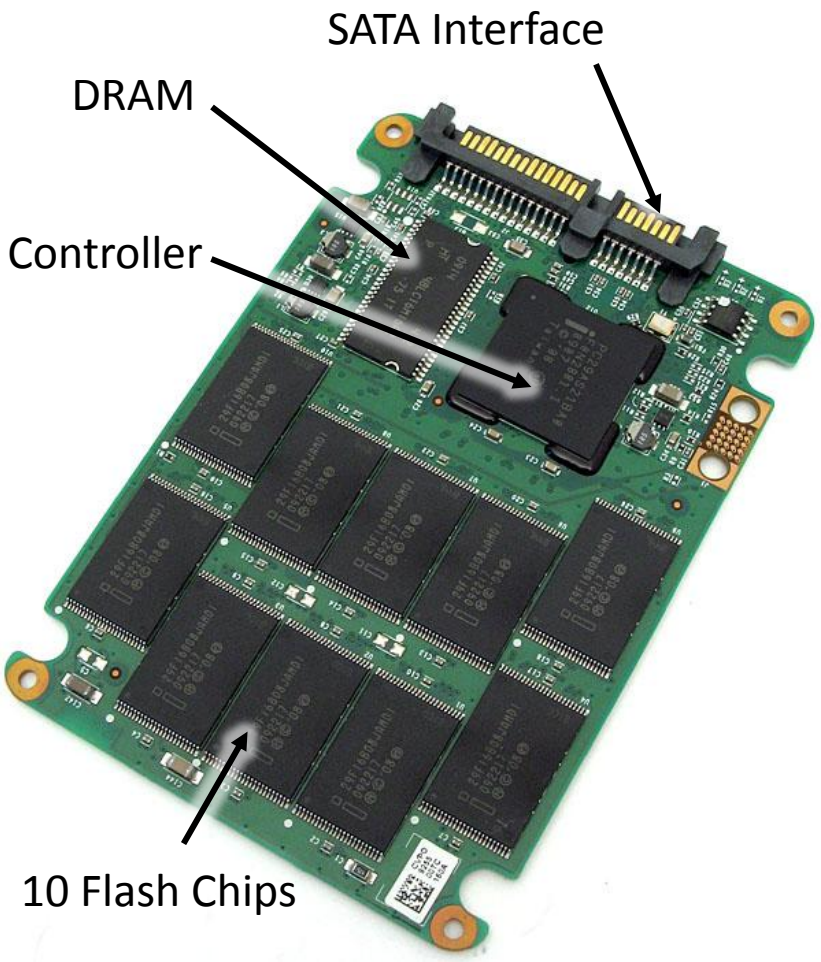
P. Frank & R. Wood, presented at APMRC2006

Helium Filled Drives

- Until just recently HDDs have been filled with air – the air cushion makes the heads “fly” – but now Hitachi (now WD) announced a drive filled with helium that will hit the market in 2013
- IBM had a patented using helium inside a drive in 1983
- Advantages of Helium Filled Drives
  - Lower Fly Height
  - Lower power (the power use of a HDD over it’s life is ~$100)
  - Less off-track disturbances due to reduced windage including less disk vibration
- Disadvantages of Helium
  - Helium is hard to contain – like throwing golf balls at a picket fence – cast aluminum baseplates need to be impregnated with epoxy
  - Extra manufacturing steps
  - It needs to last ~ 10 years
  - You need an additional inner structure because with changing external air pressure the outer structure will deform

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 – there is a wear out issue with extensive writes to the same area
- They are fast – if they use a fast interface like SAS or SATA (not USB2)
- The present max capacity is similar to that of a 2.5” HDD at ~250 GB
- BUT they are very expensive – about 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

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 now that drives are not as expensive.
- **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 understand it. It's probably something like a 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.

~500 HDDs x ~2TB/HDD = 1PByte

Google Drive Arrays

Google has several complexes around the country (and 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 The Dalles Oregon near a hydroelectric plant.

Link to NY Times Video and Article on Data Centers and Power Requirements

Video:

http://video.nytimes.com/video/2012/09/22/technology/100000001766676/what-keeps-a-data-center-going.html

Article:

http://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html?ref=technology

Google Power Consumption

After years of stony refusal to divulge its global power consumption, **Google has now revealed that the company consumed more than two billion kilowatt hours of energy worldwide** in 2010. That number includes energy consumed by servers in data centers all over the world, as well as energy required to keep the lights and refrigerators running at the Googleplex in Mountain View, Calif.

By Clint Boulton | Posted 2011-09-13

The average output of a nuclear power plant was 1 GW or about 8.8 billion kilowatt hours per year. So Google consumes about one-fourth of the power output of an average nuclear power 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.

Examples of the Need for Lots of Storage:

- The Large Hadron Collider collision data is being produced at approximately 25 petabytes (25, 000 TB = 25 million GB) per year, and the LHC Computing Grid had become the world's largest computing grid (as of 2012), comprising over 170 computing facilities in a worldwide network across 36 countries. That's about 10,000 2 TB drives per year.

- The proposed SKA (Square Kilometer Array radio telescope) when completed in ~2024, will generate about 25 exabytes per year or 1000 times that of the LHC.

- According to an IDC paper sponsored by EMC Corporation, 161 exabytes of data were created in 2006, "3 million times the amount of information contained in all the books ever written," with the number expected to hit 988 exabytes in 2010.

See Wikipedia article on exabyte.

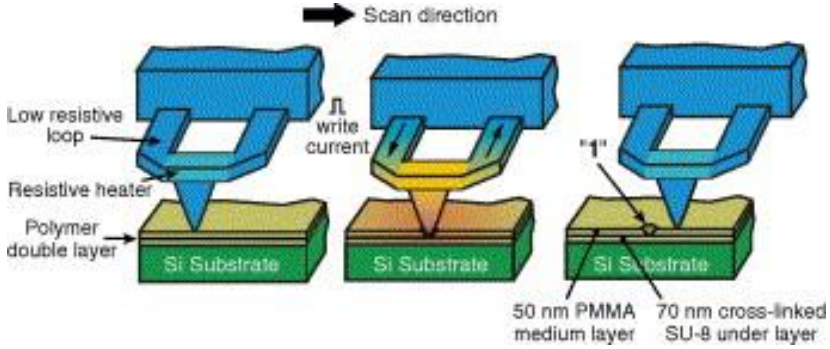
Memory/Storage Comparison

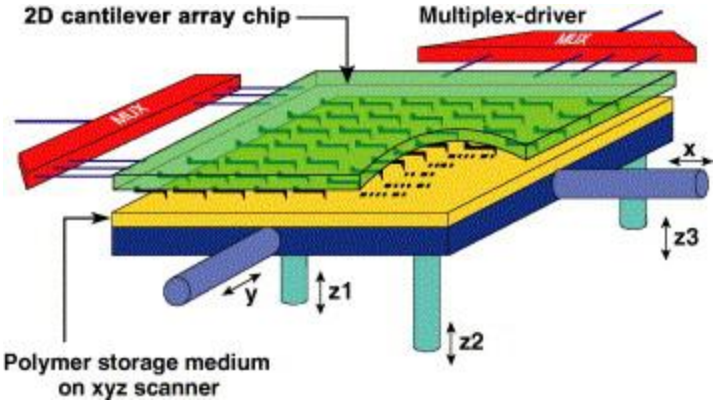
Device Type	HDD	DRAM	NAND Flash	FRAM	MRAM	STTRAM	PCRAM	NRAM
Maturity	Product	Product	Product	Product	Product	Prototype	Product	Prototype
Present Density	400Gb/in^2 [7]	8Gb/chip [9]	64Gb/chip [10]	128Mb/chip	32Mb/chip	2Mb/chip	512Mb/chip	NA
Cell Size (SLC)	(2/3)F^2	6F^2	4F^2	6F^2	20F^2	4F^2	5F^2	5F^2
MLC Capability	No	No	4bits/cell	No	2bits/cell	4bits/cell	4bits/cell	No
Program Energy/bit	NA	2pJ	10nJ	2pJ	120pJ	0.02pJ	100pJ	10pJ [11]
Access Time (W/R)	9.5/8.5ms [8]	10/10ns	200/25us	50/75ns	12/12ns	10/10ns	100/20ns	10/10ns [11]
Endurance/Retention	NA	10^{16}/64ms	10^5/10yr	10^{15}/10yr	10^{16}/10yr	10^{16}/10yr	10^5/10yr	10^{16}/10yr

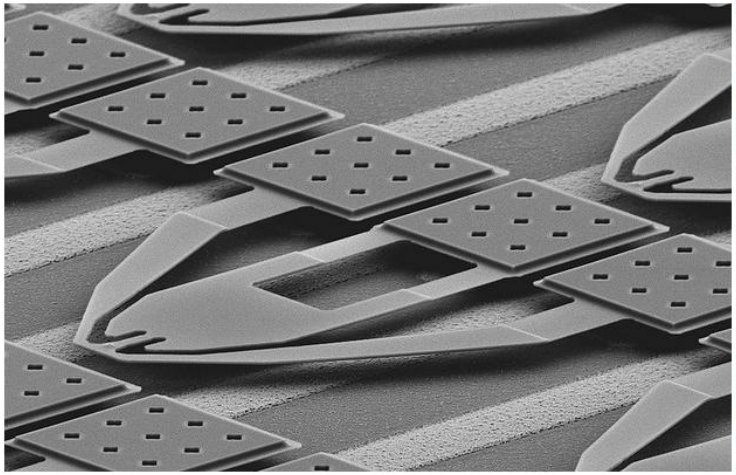
Device Type	RRAM	CBRAM	SEM	Polymer	Molecular	Racetrack	Holographic	Probe
Maturity	Research	Prototype	Prototype	Research	Research	Research	Product	Prototype
Present Density	64Kb/chip	2Mb/chip	128Mb/chip	128b/chip	160Kb/chip	NA	515Gb/in^2	1Tb/in^2
Cell Size	6F^2	6F^2	4F^2	6F^2	6F^2	N/A	N/A	N/A
MLC Capability	2bits/cell	2bits/cell	No	2bits/cell	No	12bits/cell	N/A	N/A
Program Energy/bit	2pJ	2pJ	13pJ	NA	NA	2pJ	N/A	100pJ [12]
Access Time (W/R)	10/20ns	50/50ns	100/20ns	30/30ns	20/20ns	10/10ns	3.1/5.4ms	10/10us
Endurance/Retention	10^6/10yr	10^6/Months	10^9/days	10^4/Months	10^5/Months	10^{16}/10yr	10^5/50yr	10^5/NA

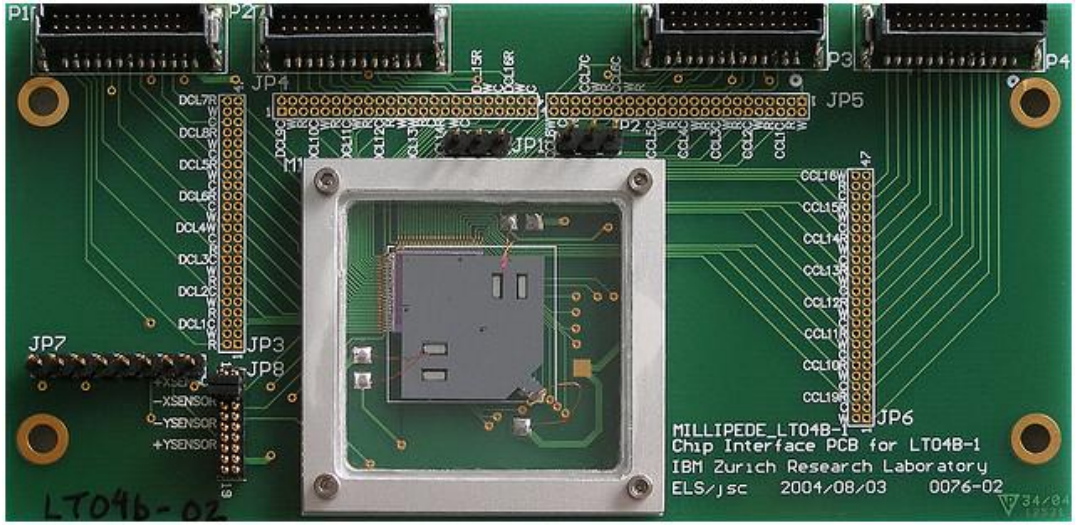
Source: After Hard Drives—What Comes Next? Mark H. Kryder and Chang Soo Kim,
IEEE TRANSACTIONS ON MAGNETICS, VOL. 45, NO. 10, OCTOBER 2009

IBM Millipede Concept









The storage density was about the same as a HDD or ~1 Tbit/in^2 in 2009.

Probe-based storage system: Small scale MEMS prototype compatible with SD form-factor: MEMS assembly (16.5 mm × 17.5 mm × 1.2 mm) encompassing the 2D cantilever read- and write-array, the micro-scanner, and the polymer medium.

How long can the rapid increase in HHD capacity continue?

[Note: bits/in^2 roughly equal Bytes for one side of a 3.5-inch HDD platter]

METHOD	Tb/in^2	COMMENTS
Present HDD (2012)	0.7	Presently available using perpendicular recording and TMR readers
HDD Limit	4 to 10	With HAMR **or** patterned media -- I upped this by 4x since 2009
	10 to 40	With HAMR **plus** patterned media
	~ 80	Single particle "super-paramagnetic limit"
Nano ferrite particles in carbon 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).

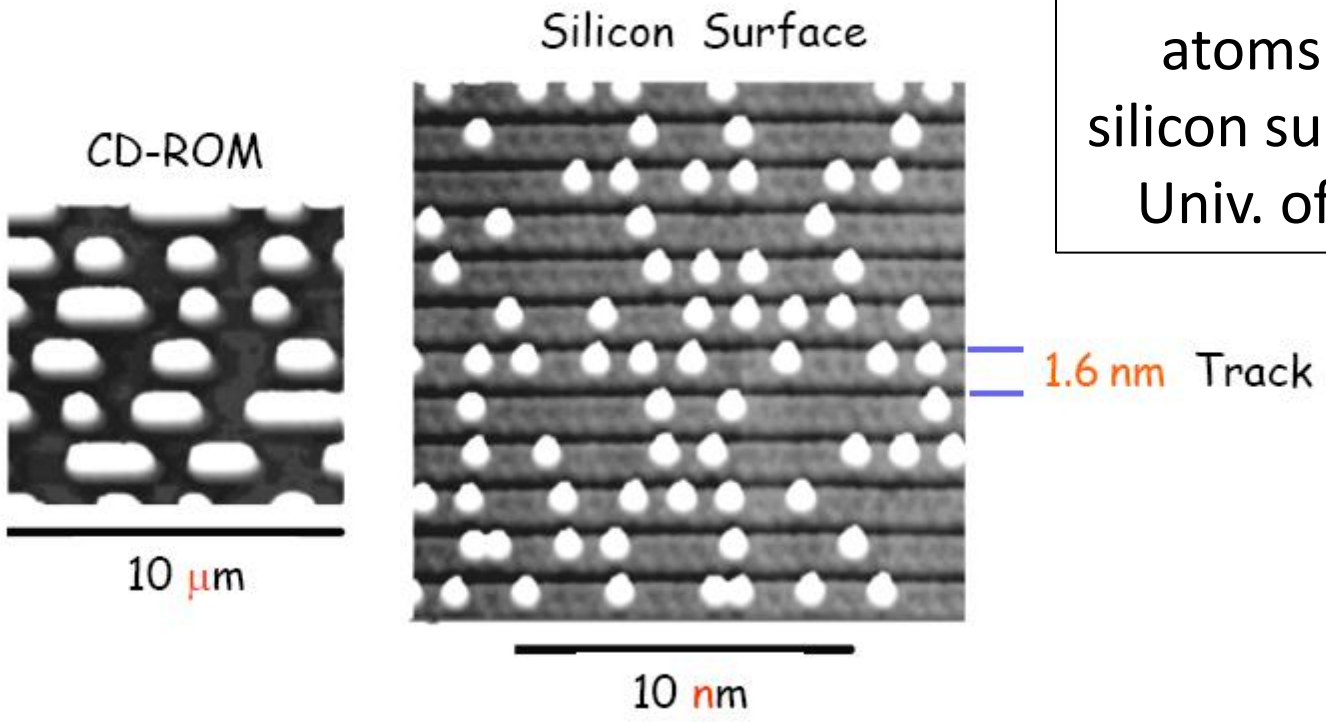
The end for HDD is in sight by ~ 2025

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.

This technique could store about 1 million times more data than an 2012 HDD can on the same area.

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

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