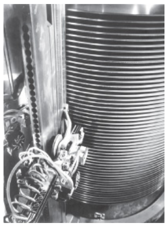
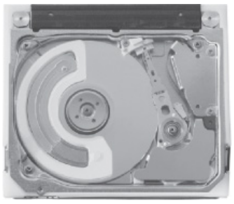
Disk
Overview & Physical Layer



1955: IBM RAMAC 305



Today: Hitachi MicroDrive

Reference: "Memory Systems: Cache, DRAM, Disk

Bruce Jacob, Spencer Ng, & David Wang

Today's material & any uncredited diagram came from chapters 16 & 17

Importance & Speed

- **Slowest form of on-line storage**
  - **but the most important**
    - » today: repository for the world's knowledge
    - » what do you care about more?
      - losing your computer or your files
- **2 roles for disks**
  - **bottom rung of the virtual memory ladder**
    - » slower and cheaper/bit than DRAM
    - » page fault ::= miss to disk
      - if it happens often – go to lunch
  - **file system**
    - » reliability & security become priorities
      - financial data centers
        - – duplicate everything
          - – data in a particular location – the usual RAIDx approach
          - – replicate locations such that
            - – natural or human disaster doesn't get them all

Offline Storage

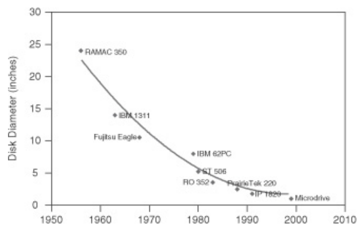
- Ignore it in what follows
- Removable disks
  - were an integral part of the computer center until the mid 70's
    - » mostly since disks didn't hold enough data
    - » and the sealed (a.k.a. Winchester) drives didn't show up until 1973.
  - now they are reserved for PC backup and transport
    - » e.g. USB or FireWire backup disks, thumb drives etc.
- Enterprise
  - several layers of backup
    - » 1st layer is disk based (access: seconds)
      - most recent snap-shots
    - » 2nd layer is tape (access: minutes – hours)
      - usually in the form of automated stackers
    - » vault (access: days)
      - holds the tapes

Comments

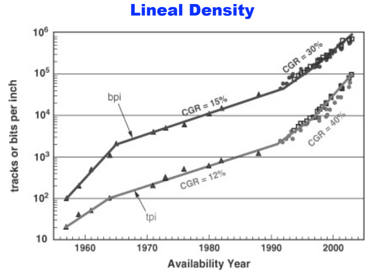
- Focus today is on hard-drive disks (HDD)
  - for on-line storage in computer systems
- Note some disks aren't really disks
  - Solid State Disk (SSD)
    - » a disk interface to a pile of chips
      - today this is FLASH based
      - PCRAM, FeRAM, NRAM, … possible future candidates
    - » significantly faster than HDD's but
      - more expensive
      - longevity issues
- Disks are pervasive in other digital gizmo's
  - iPod, DVRs, video cameras
    - » 1" & 1.8" form factors

CGR Better than Moore's Law

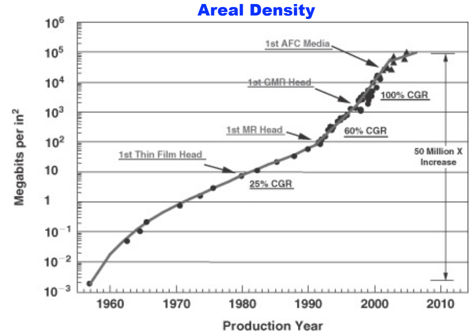
Form Factor



Lineal Density



Areal Density



Interfaces & Improvement

- **Interfaces**
  - **Control moves onto the disk**
    - » replaces motherboard control
    - » now – microprocessor and SRAM inside the disk
  - **Parallel to high speed serial interfaces**
    - » parallel SCSI – 1983, IDE/ATA – 1986
      - limited by short fat cable issues
    - » serial Fiber Channel – 1997, SAS, SATA
      - serial enables storage area networks (NAS)
- **Key improvement contributors**
  - **thinner magnetic platter coating**
  - **improvements in head design**
  - **lower flying height**
  - **accuracy of head positioning servo**
    - » hard to do cheaply
      - hence BPI CGR leads TPI CGR

Access

- **A disk address**
  - **indirectly resolved to**
    - » **surface, radius, angle**
      - **polar coordinates resolve to cylinder & sector**
- **Performance**
  - **as always multiple metrics**
    - » **latency ::= response time**
      - **since seek and rotational latency varies significantly**
      - **response time usually averaged over large number of accesses**
    - » **bandwidth ::= transfer rate**
      - **transfer rate = IOPS*average block size**
        - – **dependent on disk RPM and lineal density (BPI)**
  - **multiple requests queued in disk controller**
    - » **hence response time looks exponential w/ increase in**
      - **throughput, request arrival rate, utilization**
      - **e.g. increased queueing delay**
    - » **optimization possible be reordering requests**

Workload Impact on Performance

- **Numerous factors**
  - **block size – larger block → longer transfer time**
  - **random vs. sequential access**
  - **footprint → # seeks and rotational scope**
  - **read vs. write → writes can be deferred**
  - **Q depth: deeper → better optimization opportunity**
  - **command arrival rate**
    - » **huge burst will increase Q occupancy time**
    - » **and longer service time**

Disk Futures

- **Disk demise oft predicted**
  - "greatly exaggerated" as Mark Twain said
- **Horizontal to vertical transition underway**
  - increased areal density should continue
- **MAID might threaten tape for offline storage**
  - massive array of idle disks
- **Reduced form factor**
  - may enable RAID
  - and server storage bricks may become available in PC's
    - » brick is a bunch of disks, controller, and battery
    - » idea: even if power goes down disk writes complete
- **Common saying**
  - Silicon Valley misnomer
    - » more money made due to FeO2 than Si

Disk Storage Layers

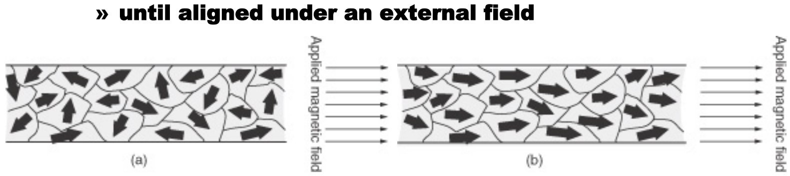
- **Physical Layer**
  - physics and engineering to just make disks work
- **Data Layer**
  - arrangement of data in blocks, sectors, stripes, ...
- **Internal Control Layer**
  - what the processor in the disk deals with
- **Interface Layer**
  - specifics of the drive interfaces
- **Cache or External Control Layer**
  - use of caches to improve performance
  - issues in management of multiple drives
    - » RAS issues such as RAID
    - » power issues such as MAID
    - » huge issue for the datacenter
- **2 lectures won't allow a deep dive into all of them**

Physical Layer

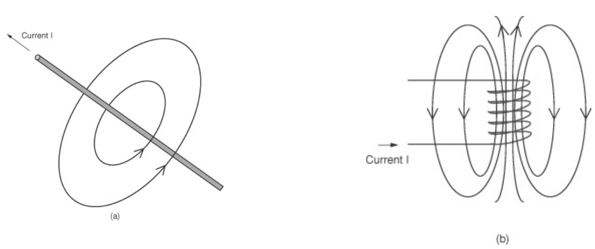
- **3 major components**
  - **magnetic recording physics**
    - » **ferromagnetic materials**
      - **magnetized by external field**
      - **stable after external field is removed**
      - **common elements: iron, nickel, cobalt**
      - **rare earth: gadolinium, dysprosium**
      - **rapidly quenched metal alloys form amorphous FM materials**
    - » **electron spin creates a magnetic field**
      - **non-FM materials consist of electron pairs w/ opposite spins**
      - **FM materials**
        - **non-paired valence shells**
        - **long range atomic ordering (aligned in parallel) to form a *domain***
    - » **beware the Curie temperature**
      - **above which the FM material loses to thermal entropy**
  - **electromechanical and magnetic components**
  - **integrated electronics in the drive**

Domains

- **Bulk material**
  - **domains randomly aligned**
    - » **until aligned under an external field**



- » **current induced fields – right hand rule**

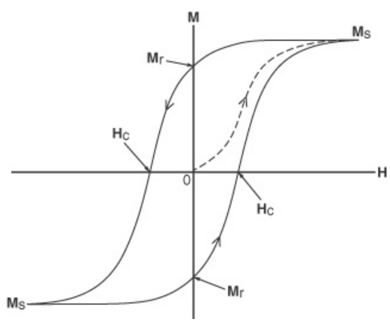


Magnetic Field properties

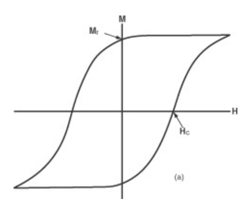
- **Measurements in MKS**
  - **things you might have forgotten from ugrad physics**
- **Field strength**
  - **H in amps/meter**
- **Dipole moment**
  - **field strength density: M – also in amps/meter**
  - **M is essentially the level of magnetization**
- **Flux density (a.k.a. magnetic induction)**
  - **B in webers/m²**
    - » **$B = \mu_0 \times H$**
    - » **where μ_0 is free space permeability = $4\pi \times 10\text{-}7$**

H-M Hysteresis

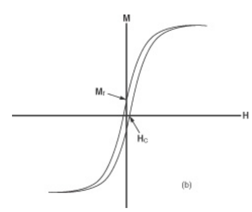
- **Key to magnetic recording**
  - **M is material state dependent**



Ms – M saturation
Mr – M remanent – non-volatile value
Hc – H coeorcivity – demagnetize



Hard material – high Mr x Hc



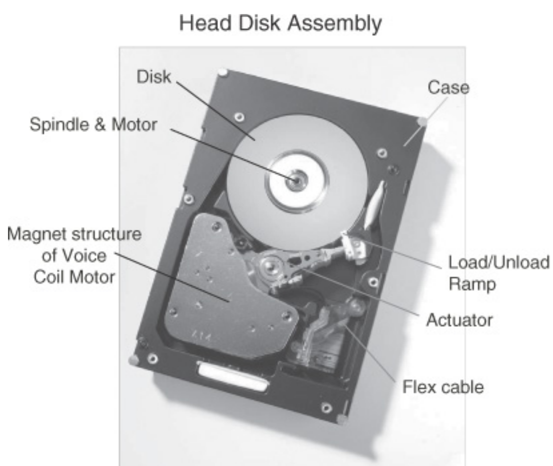
Soft material – low Mr x Hc

Axial Anisotropic: preferred axis horizontal (early) perpendicular (future)

Reading and Writing

- **Write**
  - **current in write head provides field**
    - » driven by write channel electronics
    - » ideally drive to Ms
      - highest signal to noise result since Mr separation is greatest
    - » in practice it's a suboptimal choice
      - high M compartment requires higher inter-bit separation
        - classic magnetic neighborhood problem
      - high H values on head requires more current (power)
        - and possibly more time
- **Read**
  - **option 1: read the weak magnetic fields**
    - » data value based on polarity
    - » problem – too hard to work in practice
  - **option 2: sense field reversal (easier)**
    - » 1 = reversal, 0 = no reversal
- **Required: balance read head sensitivity and write head capability**

HDD Anatomy



Recording Medium

- **Desireable properties**
  - thin (takes up less space)
  - light (less power to spin)
  - flat, smooth, rigid (low distortion allows head to fly lower)
  - High Hc (stable Mr under high areal density)
  - High Mr (improved signal to noise ratio)
  - tall thin rectangular hysteresis loop (not found in practice)
    - » max +Mr/-Mr separation
    - » smaller H currents for write efficiency
- **Substrate**
  - traditionally aluminum
    - » now plated with electroless nickel-phosphorus
      - polished to a smoother finish
  - now small form factor allows glass to be used
    - » more expensive but finer polish possible

Magnetic Layer

- **1st 25 years**
  - particulate media
    - » magnetic particles in organic binder solution
    - » painted on spinning platter
      - high rpm creates relatively uniform coating
    - » bake in oven to bind and then polish
  - magnetic material
    - » gamma ferric oxide
    - » later: cobalt modified FeO, CrO, BaO_2
      - typically used for flexible media since they are less brittle
    - » HDD now – use thin film
      - sputtered magnetic material
        - – Ar plasma bonds material directly into substrate
      - magnetic material not diluted by binder → higher areal density
      - extremely uniform coating

Platter Cross Section

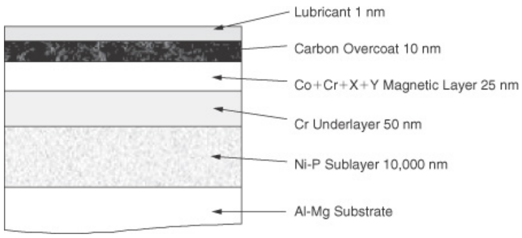
NiP – harder surface than Al-Mg

Cr – aids magnetic layer properties and bonding

Magnetic layer – Cr increases coercivity and squareness, grain size influenced by process – e.g. temp and rate of deposition

C overcoat – very thin hermetic seal to prevent rust

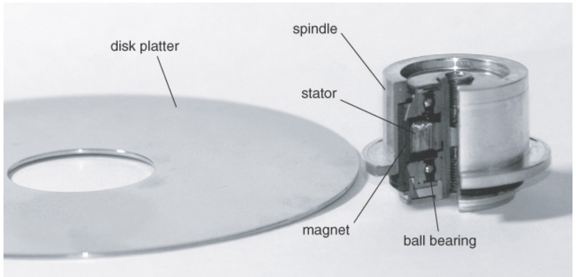
Lubricant – super thin, reduce wear between head and disk

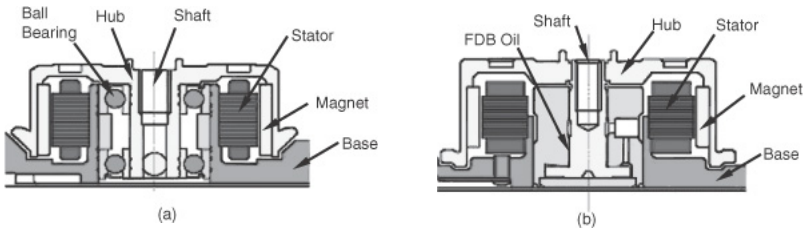


Spindle Motor

- **Today w/ high areal density**
  - » DC 3-phase 8-pole motors are common
  - » spindle integrated into motor
  - » platter attached to spindle
- **Ideal motor properties**
  - » reliable over years and thousands of start/stop cycles
  - » low vibration – so head doesn't impact surface
  - » minimal wobble – improves track registration
  - » low noise – customer appeal
  - » high shock tolerance – particularly for mobile
    - issue for non-motor components as well
- **Bearings are a big deal – see all of the above**
  - » ball bearings now replaced with FDB's
  - » fluid dynamic bearings)
    - high viscosity oil trapped in special sleeve
      - 10x improvement in wobble, 4db improvement in noise
      - better damping & reliability: larger contact surface

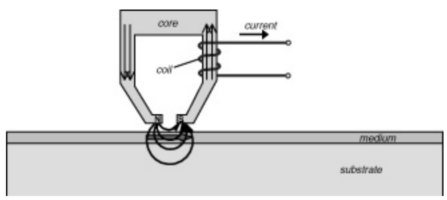
Motors Illustrated





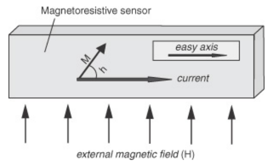
Write Heads

- **Inductive ring based head**
  - **electromagnet with a gap (no change over time)**
    - » **flux "leak" through gap passes through the recording medium**
  - **desireable characteristics (improved significantly)**
    - » **narrow (maximizes tpi)**
    - » **high flux density core (maximizes M)**
    - » **low inductance electronics (increases reversal speed – max bpi)**
    - » **strong – reduces contact damage**
    - » **light – easier to fly and move**



Read Heads

- **Significant changes have occurred**
  - **beginning – used same inductive head as for write**
    - » **field change induces a current in the coil**
  - **MR (magneto resistive) heads sense flux directly**
    - » **MR materials change resistance**
      - **function of angle between M and applied current flow**
        - $\Delta R = C_{MR} \times R \times \cos^2\theta$
      - **permalloy is one such material**
        - C_{MR} = .002 - .003
        - magnetically soft, 20% iron, 80% nickel
    - » **constant current applied to sensor**
      - **voltage change sensed: $\Delta V = I \times \Delta R$ (Ohm's Law)**



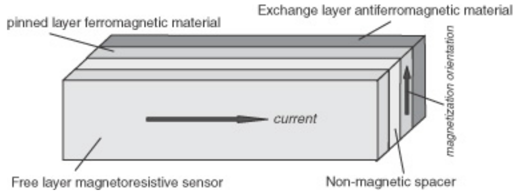
Read Head Issues

- **Clock recovery**
  - **since 1's occur with transitions**
    - » **there must be enough of them to recover the clock**
      - **hence encoding required**
- **Highest ΔR**
  - **occurs during the transition**
  - **hence bias Θ to be 45 degrees for $H_{external} = 0$**
  - **101 read waveform**

  - **MR heads drove big areal density increase starting in 1991**

Giant MR (GMR) Heads Next

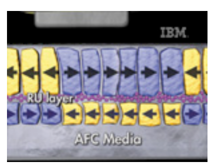
- **Composite design**
  - **made possible by molecular beam epitaxy**
  - **allows a free and pinned magnetic layer**
    - » **increases the resistance change**
      - **due to difference in field referenced to the pinned layer**
    - » **result is another increase in areal density**



    - » **video** http://www.research.ibm.com/research/demos/gmr/1.swf

AFC Media

- **IBM introduced in 2001**
  - **quadruples areal density w/ pixie dust sandwich**
    - » **3 atoms thing Ruthenium layer between 2 magnetic layers**
    - » **allows thicker material to appear thinner than it really is**
      - **circumvent the widely held "superparamagnetic" effect**
        - **beyond 20-40 Gb/in² domains are too small to hold their field polarity**
    - » **layers contain opposing polarities**



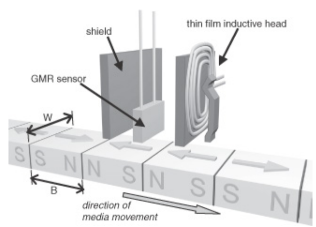
source: IBM

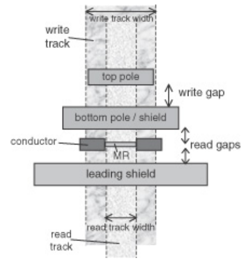
    - » **result 100 Gb/in² (and beyond claims IBM)**

http://domino.watson.ibm.com/comm/pr.nsf/pages/news.20010518_pixie_dust.html/$FILE/AFC4_mov.qt

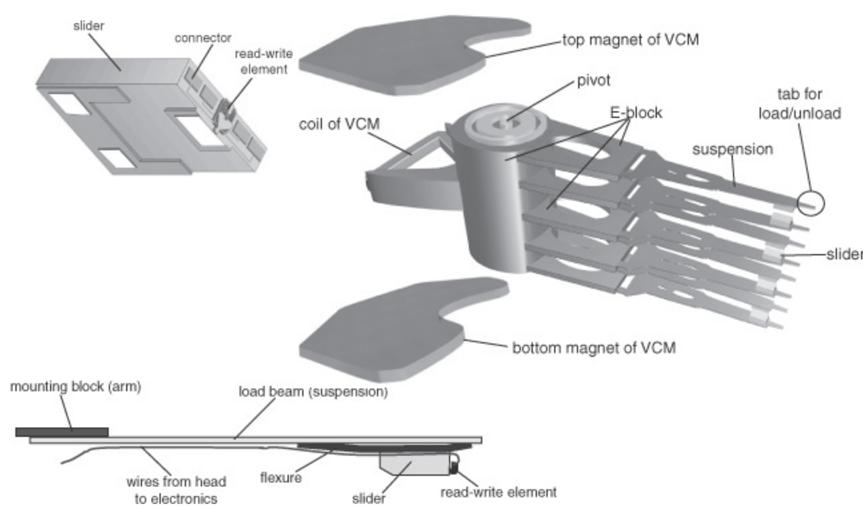
Other Issues

- **MR & GMR → separate read and write heads**
  - **each can be separately optimized**
    - » **placed in tandem**
  - **write wide read narrow is an option**
    - » **less sensitive to seek position**
  - **guard bands between tracks**
    - » **required to prevent fringe field writes affecting adjacent tracks**



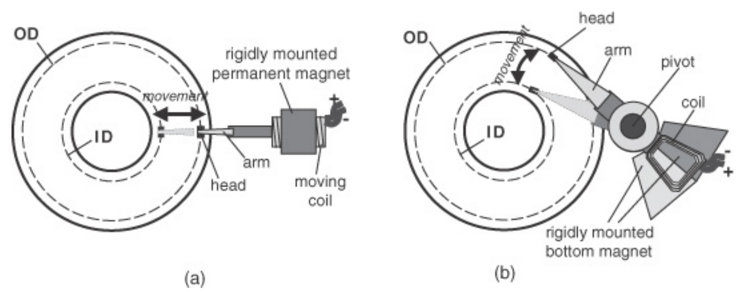


Flying Heads & Head Stack Assembly



Rotary vs. Linear Actuators

- **Rotary better**
  - if twist amount of pivot is accurate enough
  - for any track the head is tangential
    - » best signal/noise response of the read head

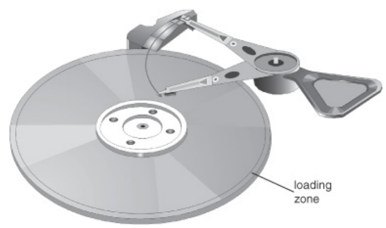


Single vs. Multiple Platters

- **Multiple platters improve capacity**
  - good idea when areal density was poor
  - problems:
    - » large % of power due to wind resistance
      - α RPM and therefore bandwidth
    - » weight of multiple arms → more powerful VCM
- **Similar issue for larger platter diameter**
  - wind resistance α area
  - increases seek stroke
- **Multiple platters better than bigger form factor**
  - due to power concerns
  - BUT single platter disks tend to be the winner

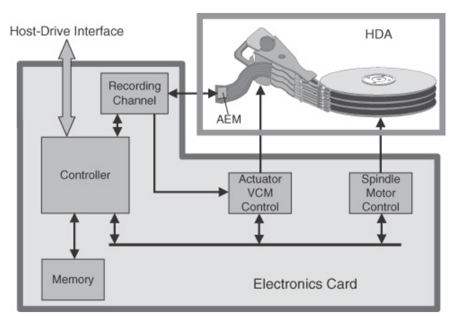
Start/Stop

- **2 approaches**
  - **contact start/stop (CSS)**
    - » let head contact platter surface as RPM's slow
      - air bearing for flying head disappears
    - » with today's high areal density
      - not a good idea
  - **load/unload**
    - » park head on a ramp before reducing RPM
    - » loading zone overlap matched to flying height



Electronics

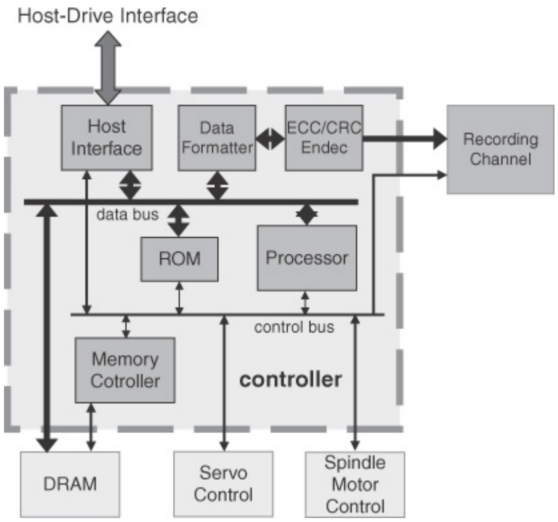
- **Small PCB inside**
  - **Controller**
    - » receive commands, schedule, and report back when command executes
    - » manage the disk cache
    - » interface with HDA – e.g. seek and sector targets
    - » error recovery and fault management
    - » power management
    - » start/stop control



Controller Components

- **ROM**
  - **holds code for the μP**
- **Memory controller**
  - **w/ larger caches SRAM moved to DRAM**
  - **simple DRAM controller & cache/write_buffer manager**
- **Host Interface**
  - **protocol specific: FC, SATA, etc.**
- **Data Formatter**
  - **move data from memory and partition into sector sized chunks**
- **ECC/CRC**
  - **usual BUT**
    - » **areal density improvement if bit compartments are allowed to be a little flakey**

Controller Illustrated

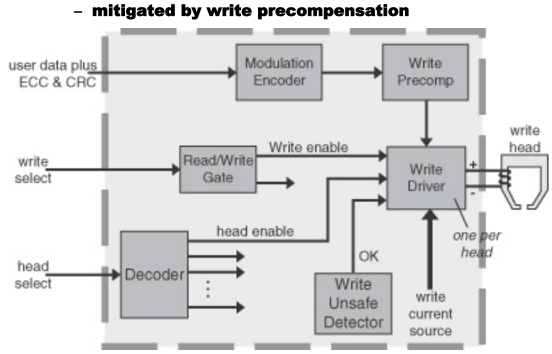


Memory

- **3 distinct roles**
  - **scratch-pad**
    - » on power up
      - load protected data from platter
        - defect maps
        - ID tables
        - adaptive operational parameters
    - » queue of commands
  - **speed matching**
    - » interface and disk bandwidths and timing differ
  - **cache**
    - » read pages
    - » write buffer

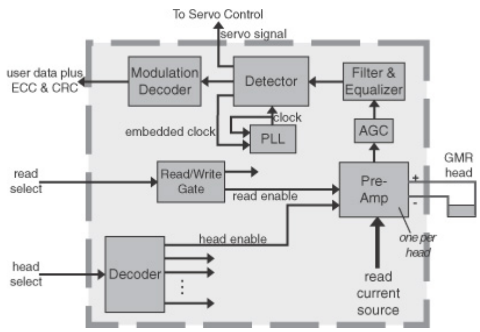
Write Channel

- **Several duties**
  - **limit run length of 0's**
    - » no transitions for too long ruins clock recovery
    - » several modulation codes possible
      - obvious 2 bits/logical_bit (50% efficient)
      - need to consider ISI (inter-symbol interference)
        - mitigated by write precompensation



Read Channel

- **GMR yields < 1mv ΔV**
  - **differential preamp located in the AEM**
  - **then AGC (auto gain control)**
  - **low pass filter to reduce high-freq noise**
- **Detection, clock recovery, & decode**



And Finally

- **Motor controls**
  - **simple ADC/DAC**
  - **but with adaptive correction**
    - » **for positioning drift & thermal issues**