Computer Science at the University of Virginia

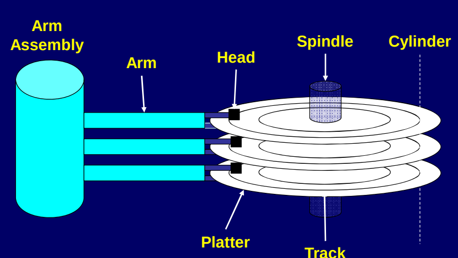
ABCs of Disk Drives

Sudhanva Gurumurthi

Hard Disk Drive (HDD) Components

- **Electromechanical**
  - Rotating disks
  - Arm assembly
- **Electronics**
  - Disk controller
  - Cache
  - Interface controller

HDD Organization



HDD Organization

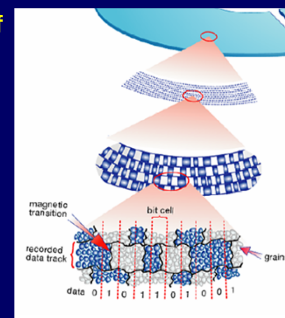
- Typical configurations seen in disks today
  - Platter diameters: 3.7", 3.3", 2.6"
  - RPMs: 5400, 7200, 10000, 15000
    - 0.5-1% variation in the RPM during operation
  - Number of platters: 1-5
  - Mobile disks can be as small as 0.75"
- Power proportional to: *(# Platters)*(RPM)$^{2.8}$(Diameter)$^{4.6}$*
  - Tradeoff in the drive-design
- Read/write head
  - Reading – Faraday's Law
  - Writing – Magnetic Induction
- Data-channel
  - Encoding/decoding of data to/from magnetic phase changes

Disk Medium Materials

- Aluminum with a deposit of magnetic material
- Some disks also use glass platters
  - Eg. Newer IBM/Hitachi products
  - Better surface uniformity and stiffness but harder to deposit magnetic material
- Anti-Ferromagnetically Coupled media
  - Uses two magnetic layers of opposite polarity to reinforce the orientation.
  - Can provide higher densities but at higher manufacturing complexity

A Magnetic 'Bit'

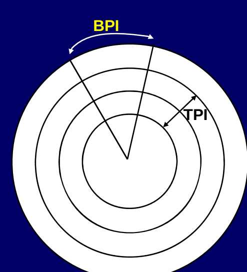
- Bit-cell composed of magnetic grains
  - 50-100 grains/bit
- '0'
  - Region of grains of uniform magnetic polarity
- '1'
  - Boundary between regions of opposite magnetization

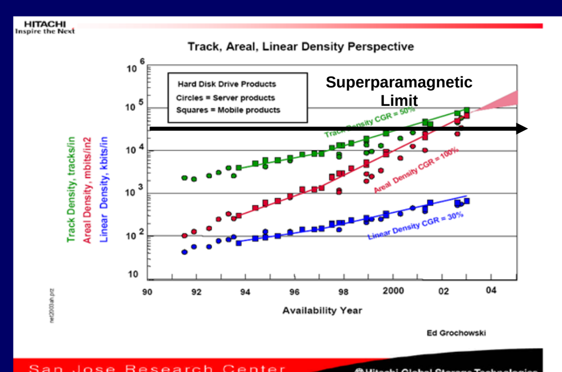


Source: http://www.hitachigst.com/hdd/research/storage/pm/index.html

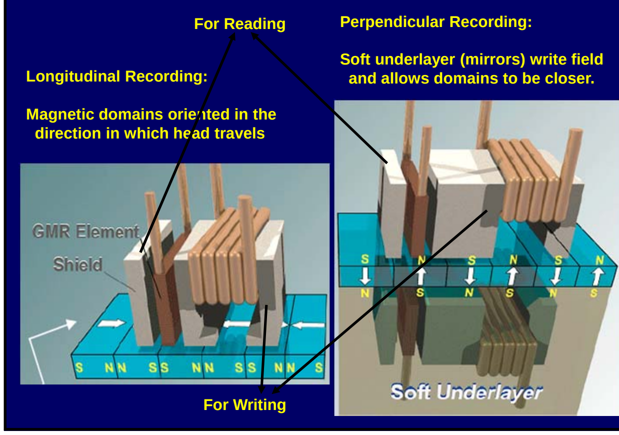
Storage Density

- Determines both capacity and performance
- Density Metrics
  - Linear density (Bits/inch or BPI)
  - Track density (Tracks/inch or TPI)
  - Areal Density = BPIxTPI





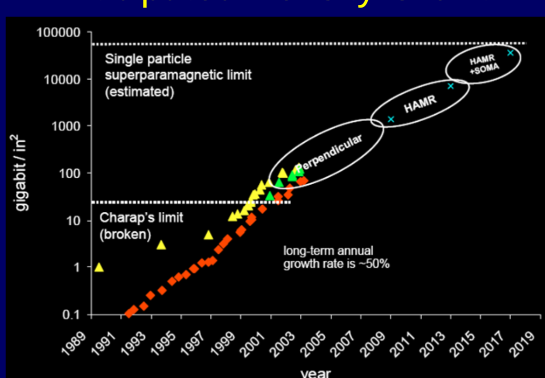
Source: Hitachi GST Technology Overview Charts, http://www.hitachigst.com/hdd/technolo/overview/storagetechchart.html



New Recording Technologies

- Longitudinal Recording now expected to extend above 100 Gb/sq-in.
- Perpendicular Recording expected to extend to 1 Tb/sq-in
- Beyond that:
  - Heat-assisted recording (HAMR)

Anticipated Density Growth

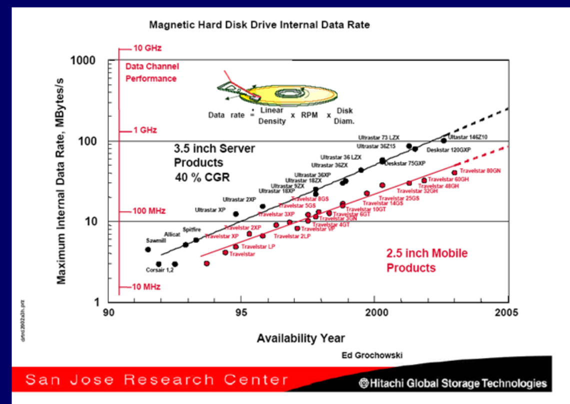


Tracks and Sectors

- Bits are grouped into sectors
- Typical sector-size = 512 B of data
- Sector also has overhead information
  - Error Correcting Codes (ECC)
  - Servo fields to properly position the head

Internal Data Rate (IDR)

- **Rate at which data can be read from or written to the physical media**
  - **Expressed in MB/s**
- **IDR is determined by**
  - **BPI**
  - **Platter-diameter**
  - **RPM**

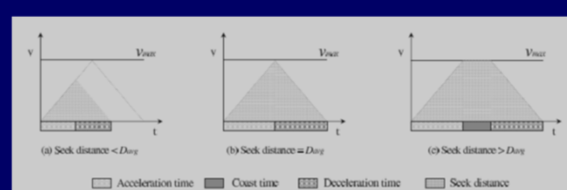


Source: Hitachi GST Technology Overview Charts,
http://www.hitachigst.com/hdd/technolo/overview/storagetechchart.html

Seeking

- **Seek time depends on:**
  - **Inertial power of the arm actuator motor**
  - **Distance between outer-disk recording radius and inner-disk recording radius (data-band)**
    - **Depends on platter-size**
- **Components of a seek:**
  - **Speedup**
    - **Arm accelerates**
  - **Coast**
    - **Arm moving at maximum velocity (long seeks)**
  - **Slowdown**
    - **Arm brought to rest near desired track**
  - **Settle**
    - **Head is adjusted to reach the access the desired location**

Physical Seek Operations



Seeking

[Speedup, Coast, Slowdown, Settle]

- Very short seeks (2-4 cylinders)
  – Settle-time dominates
- Short seeks (200-400 cylinders)
  – Speedup/Slowdown-time dominates
- Longer seeks
  – Coast-time dominates
- With smaller platter-sizes and higher TPI
  – Settle-time becoming more important

Performing the Seek

- Amount of power to apply to the actuator motor depends on seek distance
- Encoded in tabular form in disk controller with interpolation between ranges.
- Servo information used to guide the head to the correct track
  – Not user-accessible
  – Gray code for fast sampling
  – Dedicated servo surface vs. embedded servo
    - Disks might use combination of both

Head Switch

- Process of switching the data channel from one surface to the next in the same cylinder
- Vertical alignment of cylinders difficult at high TPI
  – Head might need to be repositioned during the switch
  – Can be one-third to a half of the settle-time

Track Switch

- When arm needs to be moved from last track of a cylinder to first track of the next cylinder
- Takes almost same amount as the settle-time
- At high TPI, head-switching and track-switching times are nearly the same

Optimizing for settle-time

- Attempt reading as soon as head is near the desired track
- ECC and sector ID data used to determine if the correct data was read
- Not done for settle that immediately precede a write

Data Layout

- Logical blocks mapped to physical sectors on the disk drive.
- Low-Level Layout Factors
  - Zoned-Bit Recording
  - Track Skewing
  - Sparing

Zoned-Bit Recording

- Outer tracks can hold more sectors due to larger perimeter
- Per-track storage-allocation requires complex channel electronics
- Tradeoff:
  - Group tracks in zones
  - Outer zones allocated more sectors than inner ones
  - Due to constant angular velocity, outer zones experience higher data rates.
- Modern disks have about 30 zones

Track Skewing

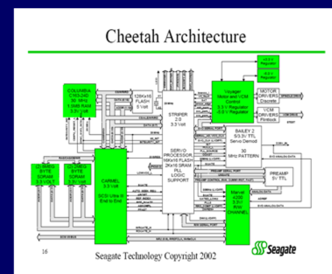
- To provide faster sequential access across track and cylinder boundaries
- Skew logical sector zero of each track by worst-case head/track switch-time
- Each zone has different skew factors

Sparing

- There can be defective sectors during the manufacture of disks
- References to them are remapped to other sectors
- Slip sparing
  - References to flawed sectors are slipped by a sector/track
- Stroke efficiency
  - Fraction of the overall disk capacity that is not used for sparing, recalibration tracks, head landing-zones etc.
  - Around 2/3 for modern disks

Drive Electronics

- Common blocks found:
  - Host Interface
  - Buffer Controller
  - Disk Sequencer
  - ECC
  - Servo Control
  - CPU
  - Buffer Memory
  - CPU Memory
  - Data Channel



Drive Electronics

- Host Interface
  - Implements the protocol between host and disk-drive eg. SCSI, ATA.
- Buffer Controller
  - To control access to the buffer memory between host interface, disk sequencer, ECC, and CPU.
  - Also controls data movement to and from host

Drive Electronics

- Disk Sequencer
  - To manage transfer between disk interface and buffer memory.
  - Also ensures that servo sectors are not over-written by user data
  - Controls timing of operations to/from disk to ensure constant data-rate

Drive Electronics

- ECC
  - Appends ECC symbols, performs error-handling operations
  - Current disks employ Reed-Solomon codes
- Servo Control
  - For necessary signal-processing for disk-rotation and head positioning
  - Needed due to motor variation, platter waviness (circumferentially and radially), stacking tolerances, vibrations, etc.
  - Additional spindle/actuator motor drivers are present for motion control
- CPU
  - DSPs to control the overall system
  - Typically the highest gate-count
  - Seagate uses 200 MHz ARM-based cores

Drive Electronics

- Buffer Memory/Disk Cache
  - Cache for data transferred between host and disk
  - Typically around 8-16 MB for modern disks
    - Use a single DRAM chip
  - Might also be used by the disk CPU as a data/code store
- CPU Memory
  - Could be ROM, SRAM, Flash, or DRAM
  - For storing CPU instruction op-codes
  - Could use a combination of volatile and non-volatile memory
- Data Channel
  - To transfer bits between controller and physical media

Read-Ahead Caching

- Actively reading disk data and placing in cache
- Variations:
  - Partial-hits
  - Large requests might bypass the cache
  - Discarding data after its had been read from cache
  - Read-ahead in 0
    - Disk continues to read where last request left-off
    - Good for sequential reads
    - Read-ahead could cross track/cylinder boundaries
    - Can degrade performance for intervening random accesses
- Could support multiple sequential read-streams by segmenting the disk cache

Write Caching

- Immediate Reporting
  - File-system can flag writes as being "done" as soon as they are written into the cache.
  - Immediate reporting disabled for metadata describing disk layout
- Use NVRAM
  - Provides write-coalescing for better utilization of disk bandwidth
  - The presence of many write requests allows for good disk scheduling opportunities

Other Issues in Disk Drive Design

- **Rotational Vibration**
- **Reliability**
  - **Duty-Cycle**
  - **Temperature**
- **Power Consumption**

Rotational Vibration

- **Caused by moving components near the drive eg. Bunch of disks in a enclosure**
- **Can cause off-track errors that can delay I/O activities or even prevent any operation to be reliably performed**
- **More of a problem at high TPI due to smaller tolerances**
- **Server-disks designed for a higher amount of vibration tolerance**

Reliability

- **Key metric – Mean-Time Between Failures (MTBF)**
- **Typical MTBF for SCSI disk = 1,200,000 hours**
  - **This is typically the first-year reliability**
  - **Assumes "nominal" operating conditions**

Factors Affecting Reliability

- **Duty Cycle**
  - **The amount of mechanical work required eg. Seek activity**
  - **Lower duty-cycles reduce the failure-rate**
    - ***For a 4-platter disk, reducing duty-cycle from 100% to 40% halves the failure-rate***
  - **Disks with more platters also increase mechanical stresses**
    - ***For 10% duty-cycle, failure rates for 1-platter and 4-platter disks are about 50% and 80% respectively***

Factors Affecting Reliability

- Temperature
  - Reliability decreases with increase in temperature
  - Includes drive temperature + heat transferred to it from external components
  - *A 15 C rise from room-temperature can double the failure-rate of the drive*
  - Drives are required to operate within a thermal-envelope for a given temperature and humidity
    - Usually 50-55 C with an external wet-bulb temperature of about 28 C

Power Consumption

- Disk power =~ *(# Platters)*(RPM)*$^{2.8}$*(Diameter)*$^{4.6}$
- Designers trade-off between them to achieve performance/capacity/power targets.
- Server disks have a higher power budget
  - Constrained only by the thermal-envelope
  - Bigger platters, faster RPMs, higher platter-counts
- Laptop disks
  - Need to be conscious of battery-energy
    - Lower power budget
  - Also might employ aggressive power-management to further reduce power consumption

Metrics for Drives

- Traditional
  - RPM
  - Seek time
  - Capacity
- New Metrics
  - Acoustics (drives in living rooms)
  - Power (battery, cooling, …)
  - Idle/Standby modes (Watts saved)
  - Shock/Vibration (cabinets, other drives, jogging)
  - Reliability (end-to-end protection)

Reading Material

- Required:
  - C. Ruemmler and J. Wilkes, “An Introduction to Disk Drive Modeling”, IEEE Computer, 27(3):17-29, March, 1994.
  - D. Anderson, J. Dykes, and E. Riedel, “More Than An Interface – SCSI vs. ATA”, FAST 2003.
- Supplemental:
  - James Jeppesen et al., “Hard Disk Controller: The Disk Drive’s Brain and Body”, ICCD 2001.
  - E. Grochowski and R.D. Halem, “Technological Impact of Magnetic Hard Disk Drives on Storage Systems”, IBM Systems Journal, 42(2):338-346, 2003.
  - D.A. Thompson and J.S. Best, “The Future of Magnetic Data Storage Technology”, IBM Journal of R & D, 44(3):311-322, May 2000.