Lecture 8: Magnetic Disks

Topics

- Some history and trends
- Structure and terms
  - Platters -> Tracks -> Sectors
  - Data storage
  - Metric: Area density
- Operation and performance
  - Actuator-> Seek, Rotate, Transfer
  - Data access
  - Performance: Disk latency
- Disk arrays (RAIDs)

SONGS ABOUT COMPUTER SCIENCE

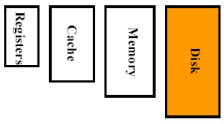
SAVE THE CODE

Written by Mikolaj Franaszczuk
To the tune of: Save Tonight
by Eagle Eye Cherry
http://www.cs.utexas.edu/users/walter/cs-songbook/save_the_code.html

... ...
save the code
and fight the break of dawn
come tomorrow
tomorrow the set is due

there's a bug in the program
and it's bad, can you fix it?
tomorrow comes with one desire
to fail me away... it's true
it ain't easy, to write in Scheme
student please, don't start to cry
'cause girl you know it's due right now
and lord I wish it wasn't so
... ...

Magnetic Disks



- Purpose
  - Long term, nonvolatile storage
  - Large, inexpensive, and slow
  - Lowest level in memory hierarchy
- Hard drive technology
  - still very much with us
  - shows no sign of going away

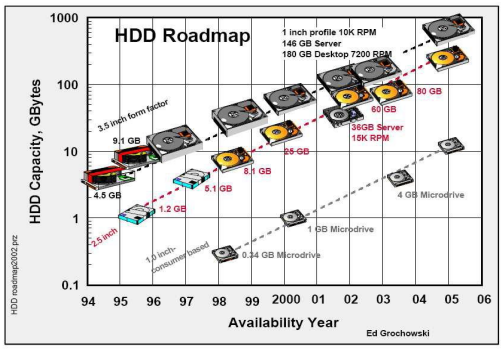
  .. despite the rise of newer technologies such as SSDs
- Hard disks
  - Data storage: rotating platters coated with magnetic surfaces
  - Data access: moveable actuator with read/write heads to access the disks

Disk History

Disk History

Year	Product	Highlight	Feature
1956	IBM 305 RAMAC	• Random Access Method of Accounting and Control • 1st HD introduced with both movable heads and multiple disk surfaces	• 50 24" disks; • Capacity of 5 MB; • areal density is a mere 2,000 bits per square inch; • data throughput 8,800 bits/s
1962	IBM ---	1st removable hard disks	
1970	IBM ---	Mainframes [IBM 370] → microcode	Invent of floppy disk drive
1970s	---	Mainframes	14 inch diameter disks
1980s	---	Minicomputers, Servers	8", 5.25" diameter disks
1990s	---	PCs Laptops, notebooks	3.5 inch diameter disks 2.5 inch

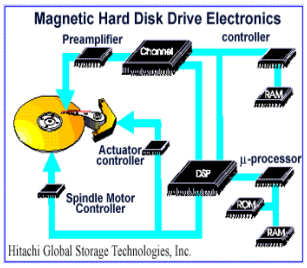
Disk Evolution – Capacity

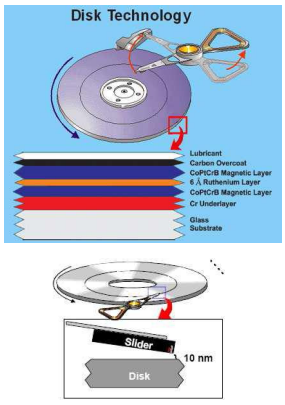


San Jose Research Center ©Hitachi Global Storage Technologies

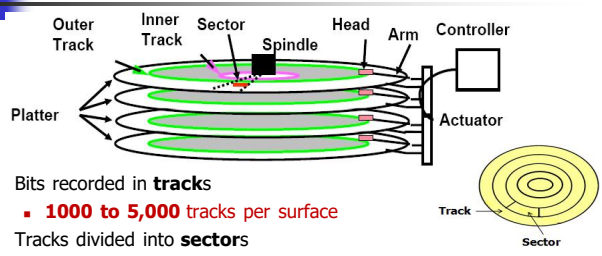
Hard Disk Drive Inside

Hard Disk Drive Inside – cont.





Platter Organisation and Data Placement

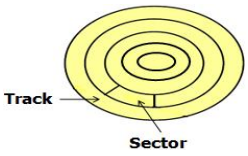


- Bits recorded in **track**s
  - **1000 to 5,000** tracks per surface
- Tracks divided into **sector**s
  - Up until the 1980's all tracks used to have same number of sectors
  - Now: constant bit size (**512 bytes/sector**), more sectors on outer tracks (**64 to 200 sectors per track**).
  - A sector is the smallest unit that can be read or written (sector #, gap, information of sector+CRC, gap, …)

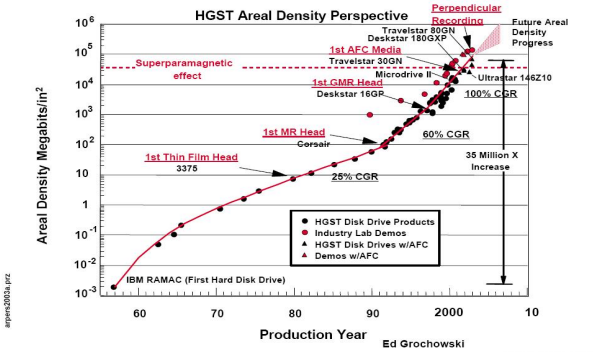
Disk Terminology and Typical Numbers

Areal Density

- Bits along track
  - Metric is Bits Per Inch (**BPI**)
- Number of tracks per surface
  - Metric is Tracks Per Inch (**TPI**)
- We are interested in bit density per units area
  - Areal Density: Metric is Bits Per Square Inch
  - **Areal Density = BPI x TPI**
- Max values achieved in "normal" products:
  - ~2005 up to **80 Gbit/in²** (not gigaBytes)
  - 5 Sep 2006 Seagate world record: **421 Gbit/in²** (laboratory testing).
  - 4 May 2011 Seagate world record: breaking the **1Tbit** areal density barrier.
  - 2016 -> **2Tb/in²** and 2020 -> **4Tb/in²**.



Disk Terminology and Typical Numbers

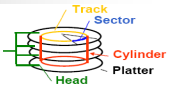


Actuator Operations and Data Access

- Actuator
  - **Seek**: moves head (end of arm, 1 per surface) over track and select surface;
  - **Rotate**: wait for sector to rotate under head;
  - **Transfer**: transfer a block of bits (sector) under the head
- Disk Latency = Seek Time + Rotation Time + Transfer Time + Controller Overhead

Actuator Operations and Data Access

- Read/write is a 3-stage process:
  - Seek time: position the arm over proper track
    - Depends on no. of tracks the arm moves, and seek speed of disk (how fast the arm moves)
    - **Average seek time in the range of 8 ms to 12 ms** = (Sum of time for all possible seek) / (total # of possible seeks)
    - Due to locality of disk reference, actual average seek time may only be 25% to 33% of advertised number
  - Rotational latency: wait for desired sector rotate under head
    - Depends on disk rotate speed , and avg. distance a sector is from head
    - 1/2 time of a rotation: 7200 Revolutions Per Minute ⇒ 120 Rev/sec ⇒ **1/2 rotation (revolution) / 4.16 ms**
  - Transfer time: transfer a block of bits under the read-write head
    - Depends on data rate (bandwidth) of disk interface (IDE, SATA, etc.)
    - **30-50 MB/sec; tends to ~70MB/sec**



Example – Two Seagate Disks (large vs. fast)

- Both drives are the same generation (a few years ago) sample drives from the same manufacturer (Seagate)
- Average seek time includes controller overhead

Barracuda 180 (large)	Cheetah X15 (fast)
181.6 GB, 3.5 inch disk	18.4 GB, 3.5 inch disk
12 platters, 24 surfaces	5 platters, 10 surfaces
7,200 RPM	15,000 RPM
26 - 47 MB/s internal media transfer rate	37.4 - 48.9 MB/s internal media transfer rate
Avg. seek: read 7.4 ms, write 8.2 ms	Avg. seek: read 3.9 ms, write 4.5 ms
areal density > 15,000 Mbits/square inch	areal density > 7,500 Mbits/square inch

Example – Disk Performance estimation

- Calculate time to read 1 sector (512B) for **Cheetah X15** using advertised performance, outer track
- Disk latency = average seek time (including controller overhead) + average rotational delay + transfer time

= 3.9 ms + 0.5 * 1/(15000 RPM) + 0.5 kB / (48.9 MB/s)
= 3.9 ms + 0.5 /(250 RPs) + 0.5 kB / (48.9 kB/ms)
= 3.9 ms + 0.5/(0.25 RPms) + 0.5 kB / (48.9 kB/ms)
= 3.9 + 2 + 0.01 = 5.91 ms

- **15,000** RPM
- 37.4 - **48.9** MB/s internal media transfer rate
- Avg. seek: read **3.9** ms, write 4.5 ms
- areal density > 7,500 Mbits/square inch

Example – Disk Performance estimation

- Calculate time to read 1 sector (512B) for **Barracuda 180** using advertised performance, outer track
- Disk latency = average seek time (including controller overhead) + average rotational delay + transfer time

= 7.4 ms + 0.5 * 1/(7200 RPM) + 0.5 kB / (47 MB/s)
= 7.4 ms + 0.5 /(120 RPs) + 0.5 kB / (47 kB/ms)
= 7.4 ms + 0.5/(0.12 RPms) + 0.5 kB / (47 kB/ms)
= 7.4 + 4.17 + 0.01 = 11.58 ms

- **7,200** RPM
- 26 - **47** MB/s internal media transfer rate
- Avg. seek: read **7.4** ms, write 8.2 ms
- areal density > 15,000 Mbits/square inch

Fallacy: Use Data Sheet "Average Seek" Time

- Manufacturers needed standard for fair comparison ("benchmark")
  - "average" = Calculate all seeks from all tracks, divide by number of seeks
- Real average would be based on how data laid out on disk, where seek in real applications, then measure performance
  - Usually, tend to seek to tracks nearby (locality), not to random track
- Rule of Thumb: observed average seek time is typically **about 1/4 to 1/3** of quoted seek time (i.e., 3-4 times faster)
  - Cheetah X15 avg. seek: 3.9 ms ⇒ 1.3 ms

Fallacy: Use Data Sheet Transfer Rate

- Manufacturers quote the speed of the data rate off the surface of the disk
  - Sectors contain an error detection and correction field (can be 20% of sector size) plus sector number as well as data
  - There are gaps between sectors on track
- Rule of Thumb: **disks deliver about 3/4 of internal media rate** (1.3 times slower) for data
  - For example, Cheetah X15 quotes 48.9 to 37.4 MB/s internal media rate
  - Expect 36 to 27 MB/s user data rate

Disk Performance estimation

- Calculate time to read 1 sector for **Cheetah X15** again, this time using 1/3 quoted seek time, 3/4 of internal outer track bandwidth; (before we used
- 5.91 ms)Disk latency = average seek time (including controller overhead) + average rotational delay + transfer time

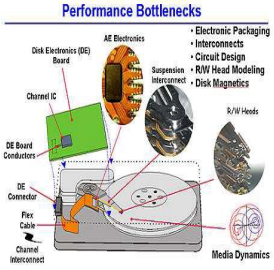
  = (0.33 * 3.9 ms) + 0.5 * 1/(15000 RPM) + 0.5 kB / (0.75 * 48.9 MB/s)

  = 1.3 ms + 2 ms + 0.5 KB / (36.6 kB/ms)

  = 1.73 + 2 + 0.013 = 3.743 ms
- Compare calculated disk latency:
  - When using advertised data: 5.91 ms
  - When using realistic data: 3.743 ms
- Lesson: need to understand how performance numbers are calculated, what they really mean, and how they relate to real scenarios

Some Challenges

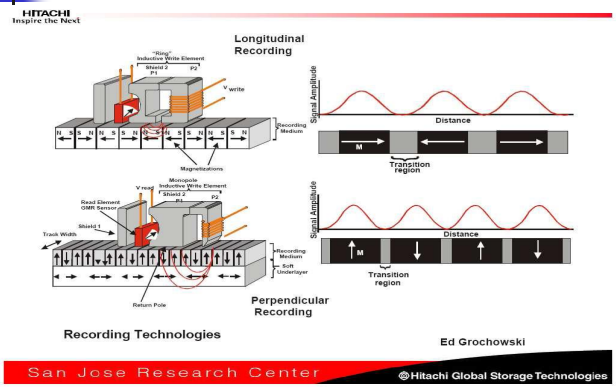
- Some **new technologies** (not covered here due to lack of time):
  - Tunnel-valve effect recording head
  - Patterned magnetic media (as opposed to conventional multigrain media)
  - Drive electronics (signal processing)
  - Disk surface lubricating layer
  - Perpendicular recording
  - …etc.



Perpendicular Magnetic Recording (PMR)

- Promising technology: perpendicular magnetic recording
- Not new, but still difficult to implement
  - Over 20 years from theory to first commercial products
  - First commercial drives: 2006
- Now Perpendicular Magnetic Recording (PMR) hard drives can deliver up to **10 times** the storage density of longitudinal hard drives
- Potential to achieve 20 times increase in areal density in the next 5(?) years
- Perpendicular vs. longitudinal magnetic recording:
  - PMR stacks bits vertically on the surface rather than laid out horizontally

Perpendicular Magnetic Recording (PMR)



Disk Performance Model /Trends

- New technologies + fewer chips + increased areal density
- Capacity
  - + 60%/year (2X / 1.5 yrs)
- Transfer rate (BW)
  - + 40%/year (2X / 2.0 yrs)
- Rotation + Seek time
  - + 8%/ year (2X in 10 yrs)
- MB/$
  - > 60%/year (2X / <1.5 yrs)
- Size?

State of the art trends: 2003

- Speed: [ref. Seagate Cheetah 15K.3]
  - Capacities: 18.4/36.7/73.4 GB, 3.5 inch disks, **15,000 RPM**
  - Interface: SCSI (improved Ultra320 SCSI protocol, max. theoretical transfer rate of 320 MB/sec)
  - 57-86 MB/s internal media transfer rate
  - Some models have 16MB cache buffer (still max in 2006!)
- Capacity: [ref. Western Digital Caviar "Drivezilla"]
  - **200GB**, IDE interface, 7,200 RPM, 8MB cache buffer
- Notes:
  - capacity / RPM / interface compromise (we can't have all!)
  - More about SCSI protocols, incl. Ultra320 on: www.adaptec.com
  - Hard disk news, tests, articles, etc. on: www.storagereview.com

State of the art trends: 2006-2007

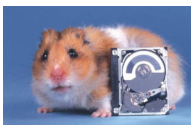
- More commercial products use **P**erpendicular **M**agnetic **R**ecording
- The largest **capacity drives**: **1TB** (**PMR**)
- Max. GB/platter increased to over **250GB/platter**
- **2.5 inch** and **1.8 inch** drives (laptop, video recorders) up to **160GB** and **7,200rpm**, use perpendicular recording
- More Serial ATA drives with 10,000RPM spindle speed, but usually with smaller capacity
  - But: common "standard" spindle speed remains **7,200RPM**
- Large cache drives: 16MB and 32MB
  - Is larger cache always = faster data transfers???
- Every manufacturer now offers 'very fast' drives.
  - Typical models: Ultra320 SCSI interface, **15,000rpm**, up to **1TB capacity** (though smaller capacities are more common), around 3ms seek time, 8-16MB cache.

State of the art trends: 2011-

- Capacity: Terabyte era
  - 1TB **platter** capacity or even **areal** capacity
  - 1TB **drive**s in laptops
  - Toshiba breakthrough: 4TB/in^2 (not per 3.5" platter)
- Size: HDD form-factor
  - Switch from 3.5-inch to 2.5-inch
  - 2.5-inch hard drive technology is energy efficient
- Speed: Hard drive performance
  - State-of-the-art is 15,000 rpm
  - Size and capacity are emphasized over performance - less attractive to increase HDD speeds
- Perpendicular Magnetic Recording (PMR); Heat-Assisted Magnetic Recording (HAMR); Microwave-Assisted Magnetic Recording (MAMR); Bit-Patterned Media (BPM);

Smallest disk drives

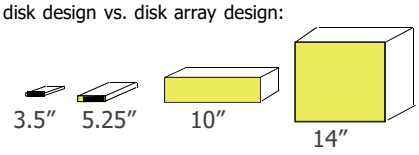
- 1.7" : Yr2000, IBM introduced Microdrive:
  - 1.7" x 1.4" x 0.2",1 GB, 3600 RPM, 5 MB/s, 15 ms seek
- 1.0" : Yr2005: 1", reached 8 GB
  - Smaller form: Compact Flash Type II
  - Applications: digital cameras, handheld devices, mobile phones, MP3 players, game consoles, etc.
  - BUT: solid state flash memory reached the same and higher capacities, and became more popular
- 1.8" : Yr2008, Hitachi GST (2003 IBM sold hard disk operation to Hitachi Global Storage Technologies)
  - 1.8" drive, up to 80GB, 3600 RPM, 14 ms seek, PATA interface



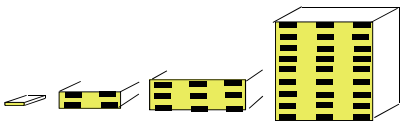
Arrays of Small Disks Concept

- Randy Katz and David Patterson asked in 1987:
  - Can smaller disks be used to close gap in performance between disks and CPUs?
- Conventional 4 disk design vs. disk array design:

Conventional:
4 disk designs



Disk Array:
1 disk design



RAID

- "RAID"
  - Originally: Redundant Array of Inexpensive Disks (as opposed to: SLED or Single Large Expensive Disk).
  - Today all disks are relatively inexpensive, thus "I" was changed to: "Independent"
- Files are "striped" across multiple disks
- Redundancy yields high data availability
  - Availability: service still provided to user, even if some components failed
- Disks will still fail
- Contents reconstructed from data redundantly stored in the array
  - Capacity penalty to store redundant info
  - Bandwidth penalty to update redundant info

Berkeley History, RAID-I

- RAID-I (1989)
  - Consisted of a Sun 4/280 workstation with 128 MB of DRAM, four dual-string SCSI controllers, 28 5.25-inch SCSI disks and specialized disk striping software
- Today RAID is around $30 billion dollar industry, 80% non- PC disks are sold in RAID configurations

Three Basic RAID Concepts

- **P**arity (next slide)

- **S**triping

- **M**irroring

Parity Bits

- Parity is used as data redundancy technique in RAID. [Also used for error detection in communication (modems, etc), memory, etc.]
- Created by using logical operation XOR (exclusive OR):
- For example:
  - 111111 XOR 000000 = 111111
  - 101010 XOR 111111 = 010101
  - 111111 XOR 010101 = 101010

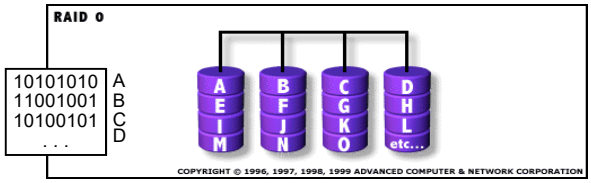
How Parity Bits Provide Fault Tolerance?

- Assume a setup with two separate groups of disk drives and a disk controller.
  - One group of disks for "data", and
  - Another group for "parity" bits (as an example see specific configuration for RAID level 3).
  - The disk controller writes data as 0's and 1's to a "data" disk
- Data bits are added up [the total for the row of data was odd or even], and the controller records parity bit 1 or 0 onto a "parity" disk depending upon whether odd or even
- If a disk fails, the controller rechecks all of the rows of data and writes 0 or 1 that "disappeared", but should be present on a new, "replacement" drive
- The reconstruction of "failed" disk and details of parity bits role varies between different implementations of **RAID levels** – see next slides.

RAID Levels: RAID level 0

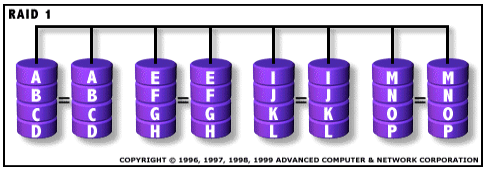
- RAID level 0
  - Disk **striping** only, data interleaved across multiple disks for better performance. No safeguard against failure.

Below: 4 disks, faster access as transfer from 4 disks at once.



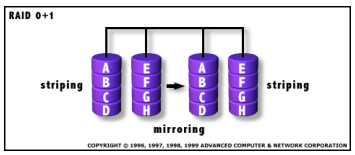
RAID Levels: RAID level 1

- RAID level 1
  - Disk **mirroring** and duplexing. 100% duplication of data. Highest reliability, but double cost. Minimum two disks to implement.
  - For highest performance controller performs two concurrent reads and writes per mirrored pair.



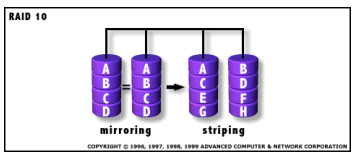
RAID Levels: RAID level 0+1

- RAID level 0+1
  - mirrored array whose segments are RAID 0 arrays
  - high data transfer performance
  - not high reliability: single drive failure will cause the whole array to become level 0 array
  - requires minimum 4 drives, expensive



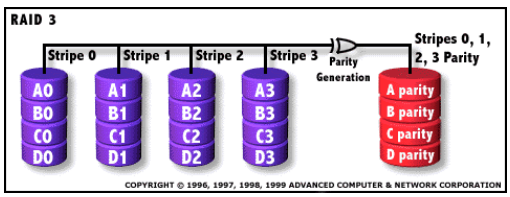
RAID Levels: RAID level 10

- RAID level 10
  - combination of level 0 and level 1 (striping and mirroring). Stripped array whose segments are RAID 1 arrays.
  - very high reliability and performance
  - requires minimum 4 drives; expensive



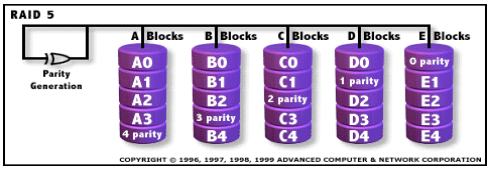
RAID Levels: RAID level 3

- RAID level 3
  - Data **striped** across three or more drives. Highest data transfer, drive operate in parallel. **Parity** provides fault tolerance, parity bits are stored on separate drives. If one drive fails, the controller reconstructs data.



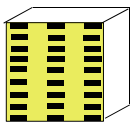
RAID Levels: RAID level 5

- RAID level 5
  - most widely used. Data **striped** across three or more drives for performance, **parity** bits used for fault tolerance. Different drives hold the parity bits.



Array Reliability

- Reliability - whether or not a component has failed
  - Measured as Mean Time To Failure (MTTF)
- Reliability of N disks
  - Reliability of 1 Disk ÷ N
  - 50,000 Hours ÷ 70 disks = 700 hour
- Disk system MTTF:
  - Drops from 6 years to 1 month!
- Large and simple arrays too unreliable to be useful!
- But: disks are becoming more and more reliable!



Revision and quiz

- Abstraction helps create a model and understand the reality, that we focus on most significant aspects and throw away unnecessary details. For hard-drives, we utilise the Platter-Actuator model:

- We use the following equation to estimate the Disk Latency performance:

  Disk Latency = Seek Time + Rotation Time + Transfer Time + Controller Overhead

  1) `True` 2) `False`
- With level 1 RAID, the parity data is stored in a dedicated hard disk.

  1) `True` 2) `False`

Recommended readings

General Data	UnitOutline | LearningGuide | Teaching Schedule | Aligning Assessments |
Extra Materials	ascii_chart.pdf | bias_representation.pdf | HP_AppA.pdf | instruction decoding.pdf | masking help.pdf | PCSpim.pdf | PCSpim Portable Version | Library materials

PH4: §6.3, P575: Disk storage

PH6: §5.11, P488 [§5.11-1 to §5.11-8]: RAID

PH5: §5.11, P470 [§5.11-1 to §5.11-8]: RAID

PH4: §6.9, P599: RAID

Text readings are listed in Teaching Schedule and Learning Guide

PH6 (PH5 & PH4 also suitable): check whether eBook available on library site

PH6: companion materials (e.g. online sections for further readings)

https://www.elsevier.com/books-and-journals/book-companion/9780128201091

PH5: companion materials (e.g. online sections for further readings)
http://booksite.elsevier.com/9780124077263/?ISBN=9780124077263