ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

Algorithms and Methods for Distributed Storage Networks

2. Hard Disks

Christian Schindelhauer

Albert-Ludwigs-Universität Freiburg
Institut für Informatik
Rechnernetze und Telematik
Wintersemester 2007/08

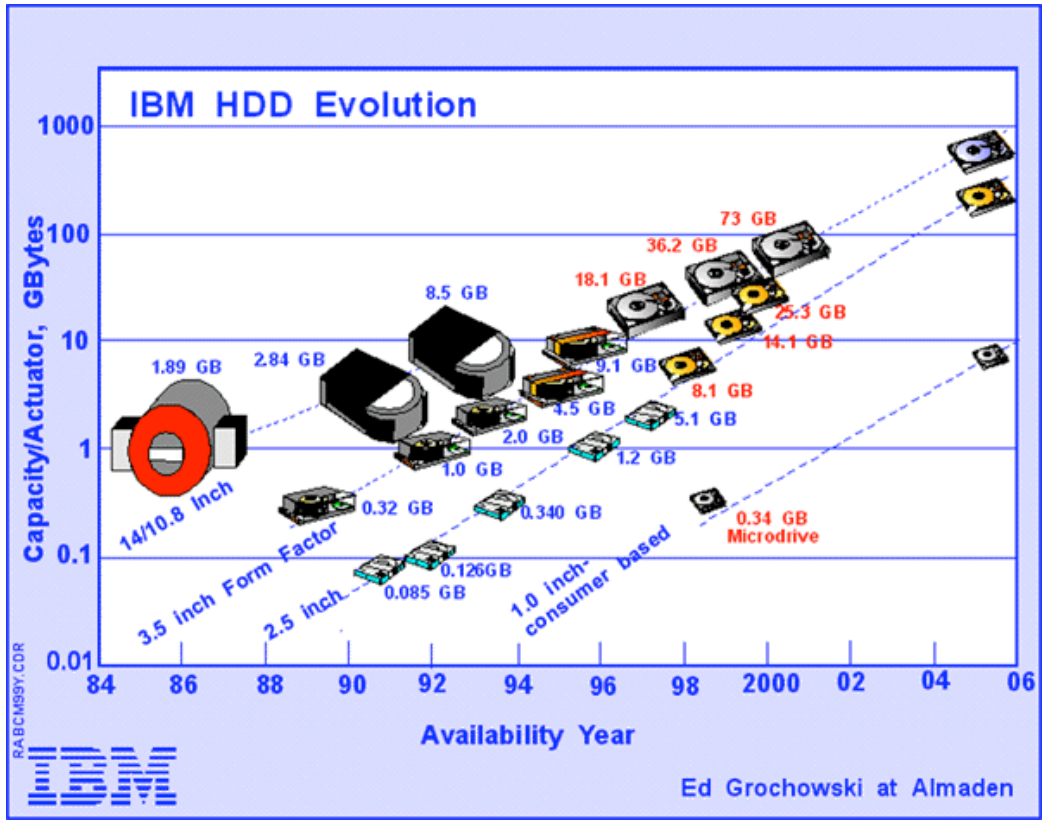
Hard Disks

- **History**
  - Capacity and Access Speed
  - Prices
  - Form factors
- **Construction and Operation**
  - Mechanics
  - Storage technology
- **Low-Level Data Structures**
  - Encoding, Decoding
  - Tracks , Cylinders
  - LBA
- **Interfaces**
  - ATA, SATA
  - SCSI, SAS
  - Fibre-Channel
  - eSATA
- **Lifetime and Disk Failures**
  - Error Management and Recovery
  - Types
  - S.M.A.R.T.
  - Counter methods
- **Special Issues**
  - Sound avoidance
  - Data security

Hard Disks

History

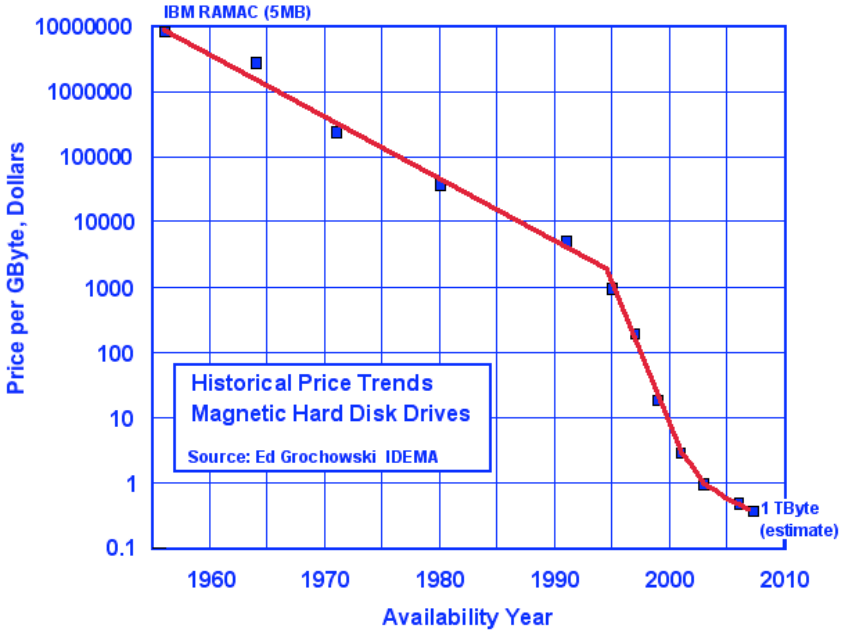
Evolution of Hard Disk Capacity



History

- **1956 IBM invents 305 RAMAC (Random Access Method of Accounting and Control)**
  - 5 MBytes, 24 in
- **1961 IBM invents air bearing heads**
- **1970 IBM invents 8 in floppy disk drives**
- **1973 IBM ships 3340 Winchester sealed hard drives**
  - 30 MBytes
- **1980 Seagate introduces 5.25 in hard disk drive**
  - 5 MBytes
- **1981 Sony ships first 3.25 in floppy drive**
- **1983 Rodime produces 3.25 in disk drive**
- **1986 Conner introduces first 3.25 in voice coil actuators**
- **1997 Seagate introduces 7,200 RPM Ultra hard disk**
- **1996 Fujitsu introduse aero dynamic design for lower flighing heads**
- **1999 IBM develops the smallest hard disk of the World 1in (340 MB)**
- **2007 Hitachi introduces 1 TB hard disk**

History of Hard Disk Prices



Technological impact of magnetic
hard disk drives on storage systems,
Grochowski, R. D. Halem
IBM SYSTEMS JOURNAL, VOL 42, NO 2, 2003

Figure 6 Cost of storage at the disk drive and system level

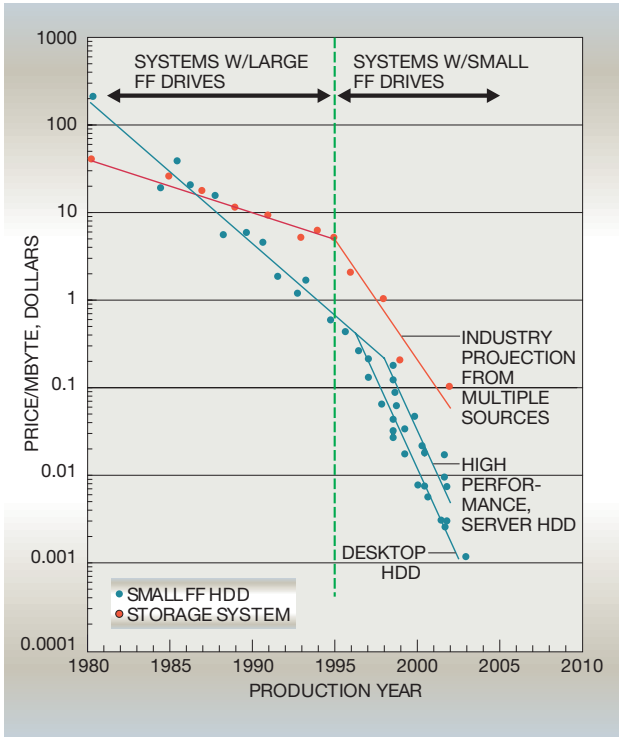
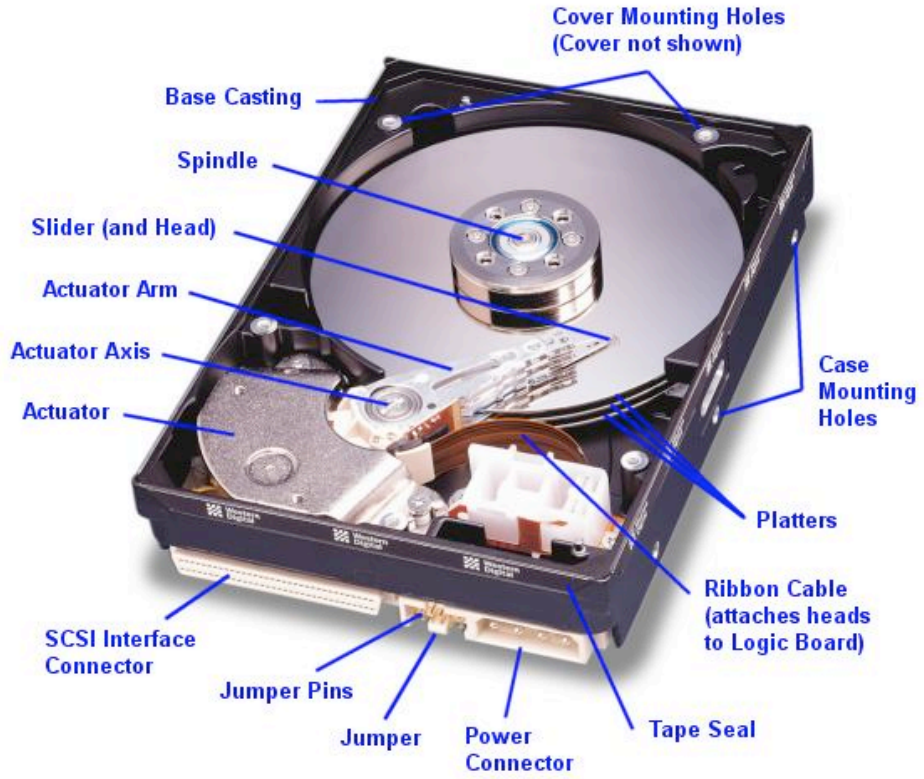


Figure 7 Cost of storage for disk drive, paper, film, and semiconductor memory

Hard Disks

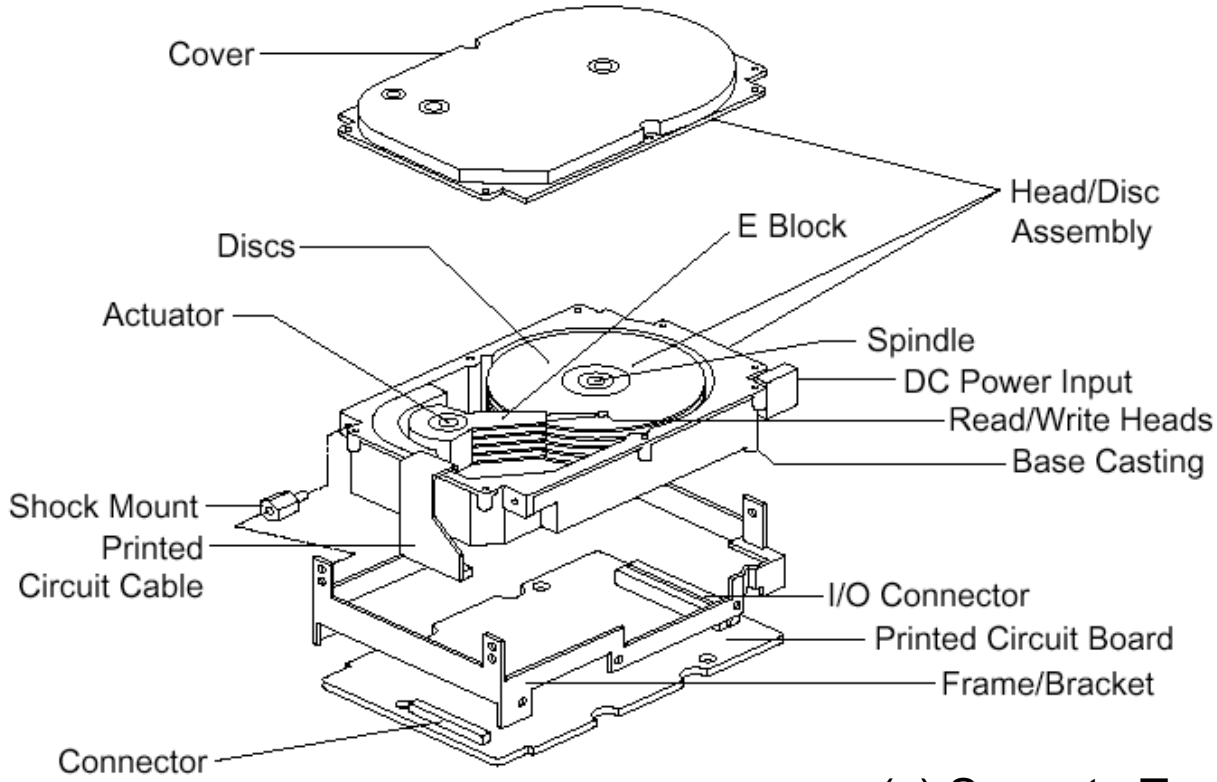
Construction and Operation

Construction of a Hard Disk



(c) Western Digital Corporation

Construction of a Hard Disk



(c) Seagate Technology

Physical Components

- **Platters**
  - round flat disks with special material to store magnetic patterns
  - stacked onto a spindle
  - rotate at high speed
- **Read/Write Devices**
  - usually two per platter
  - Actuator
    - old: stepper motor
      - mechanic adjusts to discrete positions
      - low track density
      - still used in floppy disks
    - now: voice coil actuator
      - servo system dynamicall positions the heads directly over the data tracks
    - Head arms
      - are moved by the actuator to choose the tracks
    - Head sliders
      - are responsible to keep the heads in a small defined distance above the platter
      - heads „fly“ over the platter on an air cushion
    - Read/write heads mounted on top of arms

Slider

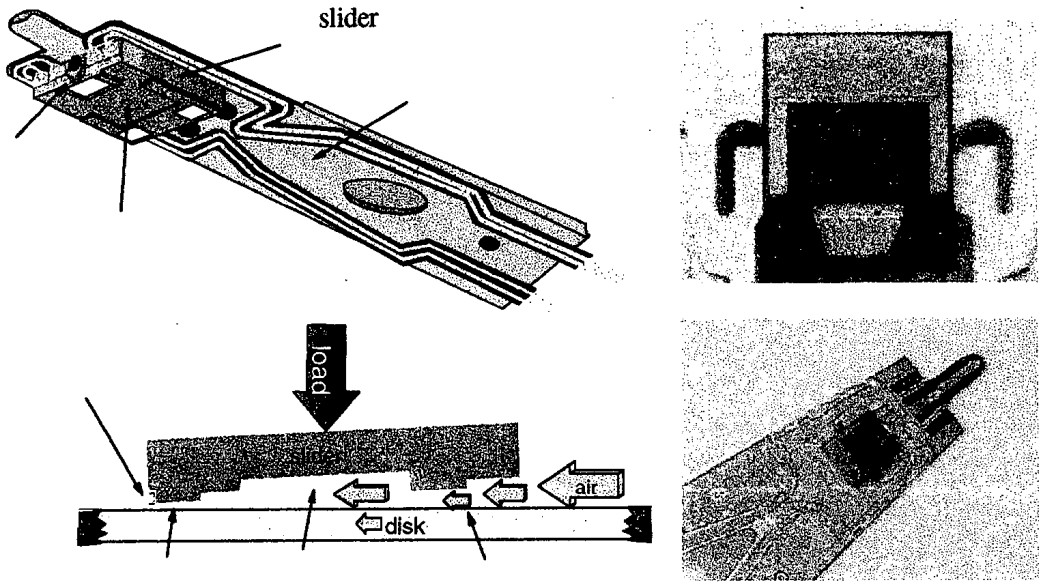
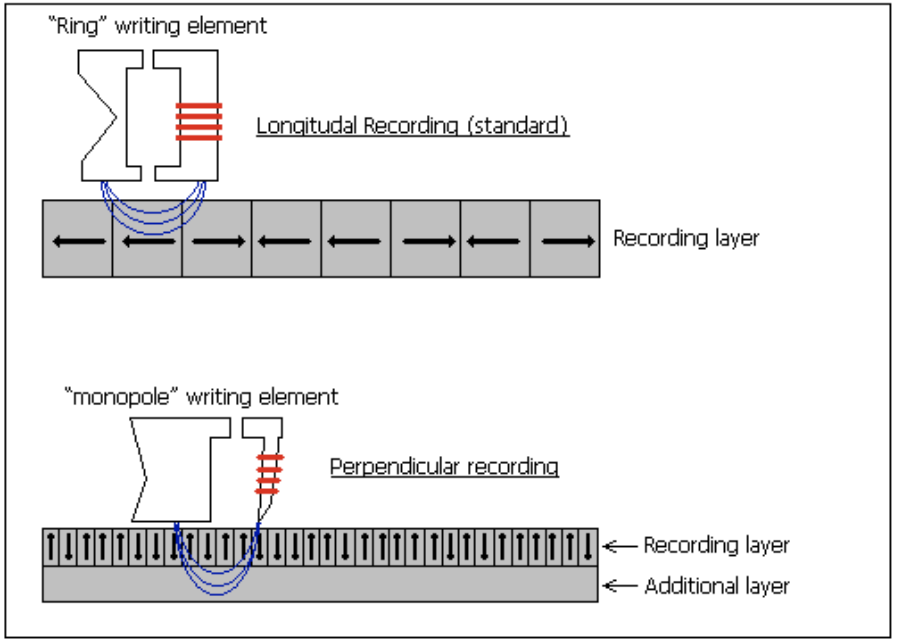


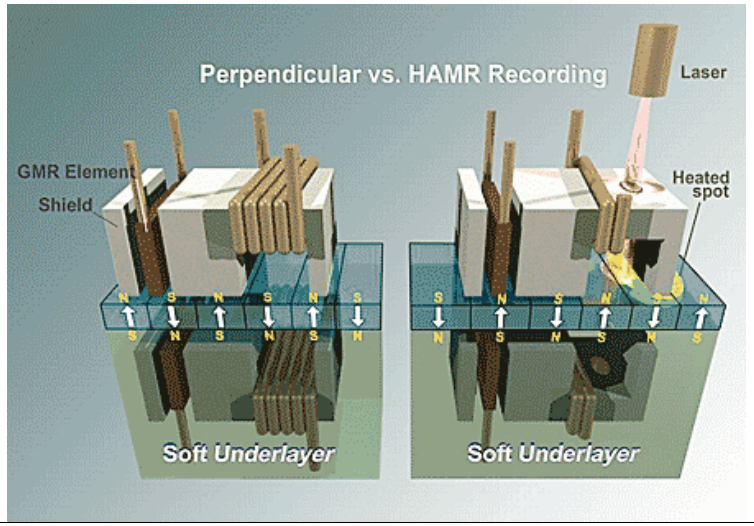
Figure 6. Illustration of suspension and slider. Left: schematic. Right: photograph. (Source: Tom Albrecht, IBM)

Proceedings of the American Control Conference ,Arlington, VA June 25-27, 2001
A Tutorial on Controls for Disk Drives William Messner , Rick Ehrlich

Magnetization Techniques

- **Longitudinal recording**
  - magnetic moments in the direction of rotation
  - problem: super-paramagnetic effect
  - 100-200 Gigabit per square inch
- **Perpendicular**
  - magnetic moments are orthogonal to the rotation direction
  - increases the data density
  - 1 Terabit per square inch
- **HAMR (Heat Assisted Magnetic Recording)**
  - upcoming technology
  - Laser heats up area to keep the necessary magnetic field as small as possible





Electronic Components

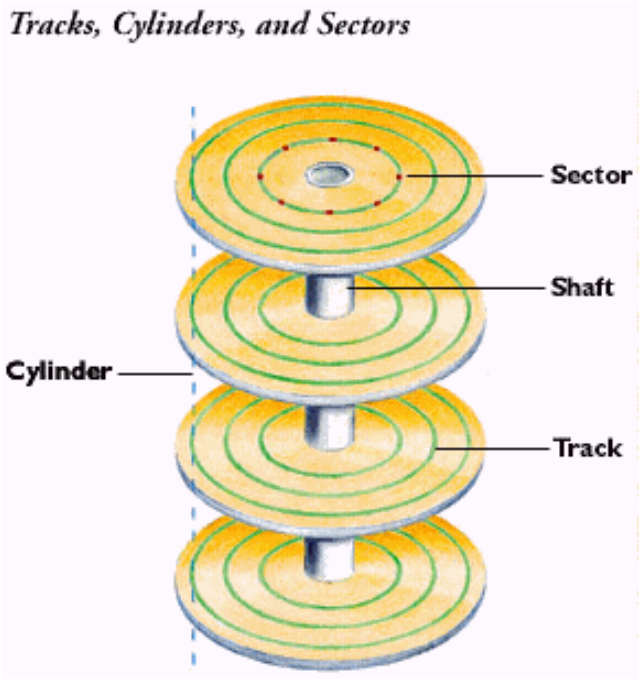
- **Magnetized Surface on platter**
- **Read/Write-Head**
- **Embedded controller**
- **Disk buffer (disk cache)**
  - store bits going to and from the platter
  - read-ahead/read-behind
  - speed matching
  - write acceleration
  - command queueing
- **Interface**

Hard Disks

Low Level Data Structure

Tracks and Cylinders

- **Tracks**
  - is a circle with data on a platter
- **Cylinder**
  - is the set of tracks on all platters that are simultaneously accessed by the heads
- **Sector**
  - basic unit of data storage
  - angular section of a circle



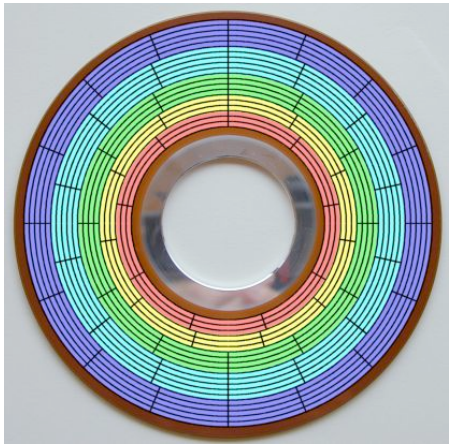
(c) Quantum Corporation

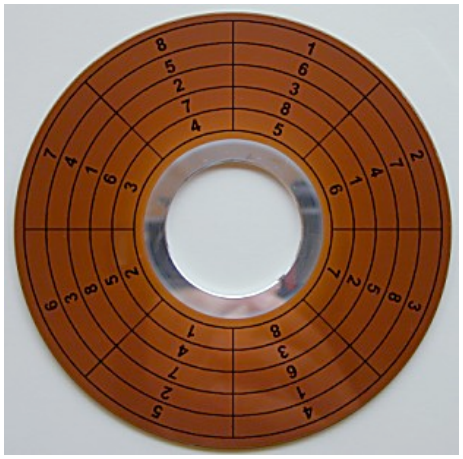
Addressing

- **CHS (cylinder, head, sector)**
  - each logical unit is addressed by the cylinder
    - set of corresponding tracks on both sides of the platters
  - head
  - sector (angular section)
  - old system
- **LBA (Logical Block Addressing)**
  - simpler system all logical blocks are number
  - the translation to CHS is

Adapting Sectors

- **Zoned bit recording**
  - adapt the sector size to the bit sensity
  - different number of sectors depending from the distance from the center
- **Sector interleaving**
  - for cylinder switch
  - when the arm moves then the disk continues spinning
  - to avoid waiting times the numbering of the sectors has an offset





http://www.storagereview.com/guide2000/ref/hdd/geom/tracksZBR.html

Sector Format

- **A sector is the atomic data unit of an hard disk**
- **No absolute position**
  - must be identified from its contents
- **Contents**
  - ID Information (number and location)
  - Synchronization fields
  - Data
  - ECC: Error correcting codes
  - Gaps
- **Specific contents varies from hard disk type**

Formatting

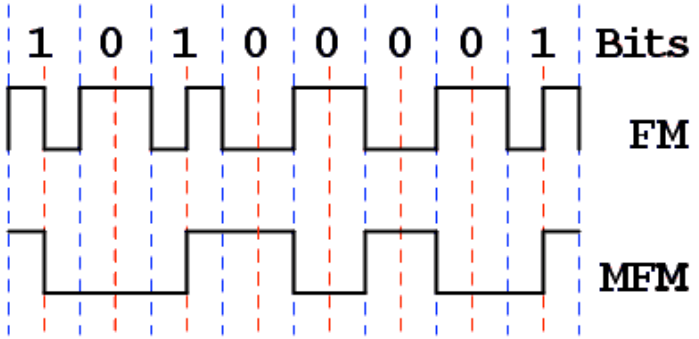
- **Low-level formatting**
  - creates the physical structures (tracks, sectors, control information)
    - starts from empty platter
    - map out bad sectors
- **Partitioning**
  - divides the disk into logical pieces (i.e. hard disk volumes)
- **High-level formatting**
  - logical structures for the operating-system level components

Encoding

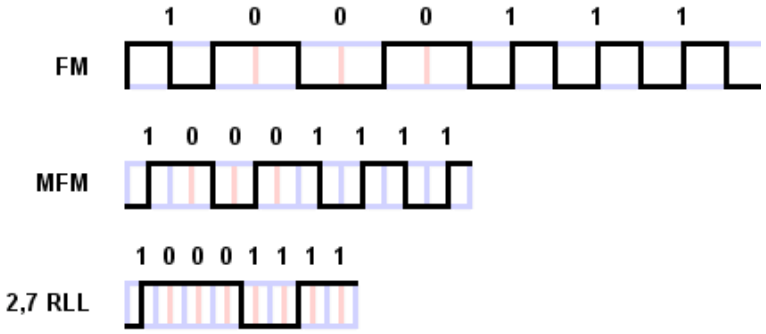
- **Problem**
  - Only the difference of orientation can be measured
  - Because of the para-magnetic effect orientation changes need a minimum distance
  - Long sequences of same orientation lead to errors
- **Encoding**
  - must have long, but not too long flux reversals

MFM

- **R: Flux reversal**
- **N: no flux reversal**
- **FM (Frequency Modulation)**
  - 0 -> RN
  - 1 -> RR
- **MFM (Modified Frequency Modulation)**
  - 0 (preceded by 0) -> RN
  - 0 (preceded by 1) -> NN
  - 1 -> NR



Run Length Limited (RLL)

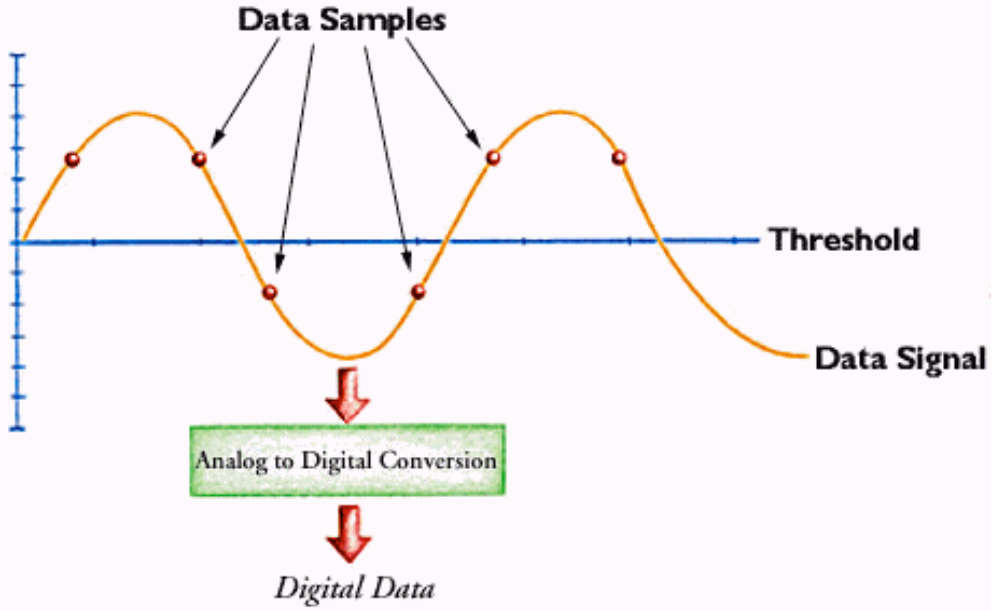


Bit Pattern	Encoding Pattern	Flux Reversals Per Bit	Bit Pattern Commonality In Random Bit Stream
11	RNNN	1/2	25%
10	NRNN	1/2	25%
011	NNRNNN	1/3	12.5%
010	RNNRNN	2/3	12.5%
000	NNNRNN	1/3	12.5%
0010	NNRNNRNN	2/4	6.25%
0011	NNNNRNNN	1/4	6.25%
Weighted Average		0.4635	100%

http://www.storagereview.com/guide2000/ref/hdd/geom/dataRLL.html

Partial Response, Maximum Likelihood (PRML)

- **Peak detection by analog to digital conversion**
  - use multiple data samples to determine the peak
  - increase areal density by 30-40% to standard peak detection
- **Extended PRML**
  - further improvement of PRML



http://www.storagereview.com/guide2000/ref/hdd/geom/dataPRML.html

Hard Disks

Interfaces

ATA (AT Attachment)

- **Parallel connection standard, a.k.a. P-ATA**
- **evolves since 1994**
  - ATA-1: 1994-99, up to 8.3 MB/s
  - ATA-2: 1996-01
    - PCMCIA connector
  - ATA-3: 1997-02
    - introduces S.M.A.R.T.
  - ATA-4: 1998-,
    - supports CD-ROM, tape, etc.
    - features for solid state drives,
    - hidden protected area
      - hidden against OS
  - ATA-5: 2000-, up to 66 MB/s
  - ATA-6: 2002-, up to 100 MB/s, automatic acoustic management
  - ATA-7: 2005-, SATA 1.0, up to 150 MB/s
  - ATA-8, in progress

SATA – Serial Advanced Technology Attachmnet

- **Serial ATA**
- **Computer bus designed for transfer of data between motherboard and mass storage device**
  - faster transfers than P-ATA, designed as successor
  - allows removing and adding devices while operating (hot swapping)
- **Evolution:**
  - SATA 1.5 Gbit/s
  - SATA 3.0 Gbit/s
  - SATA 6.0 Gbit/s

SCSI

- **Small Computer System Interface**
  - offers higher data rates than SATA
  - hides the complexity of physical format
  - peripheral interface: 8/16 devices can be attached to a single bus
  - buffered interface
- **Evolution of Parallel SCSI**
  - SCSI: 1986, 5 MB/s
  - Fast SCSI: 1994, 10 MB/s
  - ...
  - Ultra SCSI: 1999, 160 MB/s
  - Ultra-320 SCSI: 2002, 320 MB/s
  - Ultra-640 SCSI: 2003, 640 MB/s
- **Evolution of Serial SCSI**
  - SSA: 1990 40 MB/s
  - SAS: Serial Attached SCSI, 300 MB/s
- **SCSI-Fibre Channel interface**
  - FC-AL 1Gb: 1993 Fibre Channel 100 MB/s
  - FC-AL 2Gb: Fibre Channel 200 MB/s
  - FC-AL 4Gb: Fibre Channel 400 MB/s
  - length 500m / 3km

Other Interfaces

- **eSATA (since 2004)**
  - variant of SATA for consumer market
  - maximum cable length of 2m
- **USB (Universal Serial Bus)**
  - allows hot swapping
  - 12 or 480 MBit/s
- **Firewire (IEEE 1394 interface)**
  - serial bus interface standard
  - Firewire 400: ~100/200/400 MBit/s half-duplex
  - Firewire 800: 786 MBit/s full-duplex

Hard Disks

Lifetime and Disk Failures

Disk Failure Rates

- **Failure Trends in a Large Disk Drive Population, Pinheiro, Weber, Barroso, Google Inc. FAST 2007**

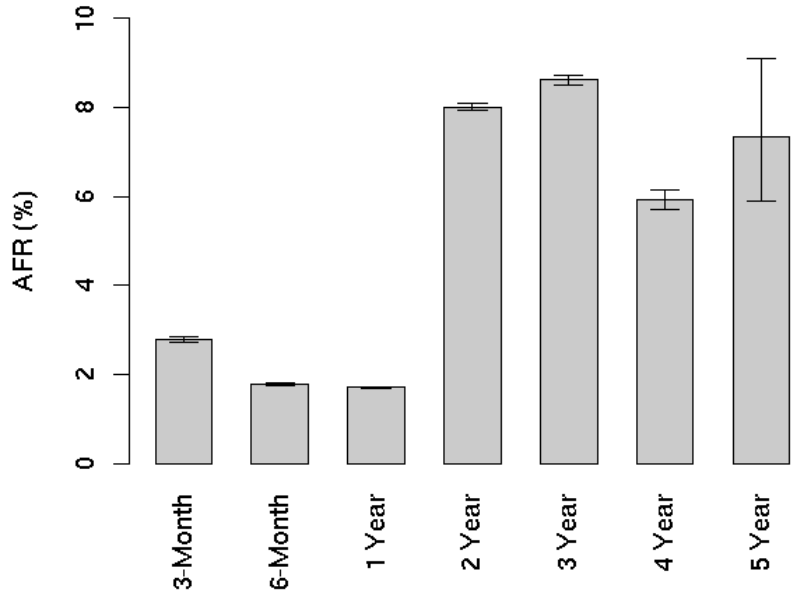


Figure 2: Annualized failure rates broken down by age groups

Reasons for Failures

- **From: www.datarecorvery.org**
- **Physical reasons**
  - scratched platter
  - broken arm/slider
  - hard drive motor failed
  - humidity, smoke in the drive
  - manufacturer defect
  - firmware corruption
  - bad sectors
  - overheated hard drive
  - head crash
  - power surge
  - water or fire damage
- **Logical Reasons**
  - failed boot sector
  - master boot record failure
  - drive not recognized by BIOS
  - operating system malfunction
  - accidentally deleted data
  - software crash
  - corrupt file system
  - employee sabotage
  - improper shutdown
  - disk repair utilities
  - computer viruses
  - ...

Reasons for Failure

- **Failure Trends in a Large Disk Drive Population, Pinheiro, Weber, Barroso, Google Inc. FAST 2007**

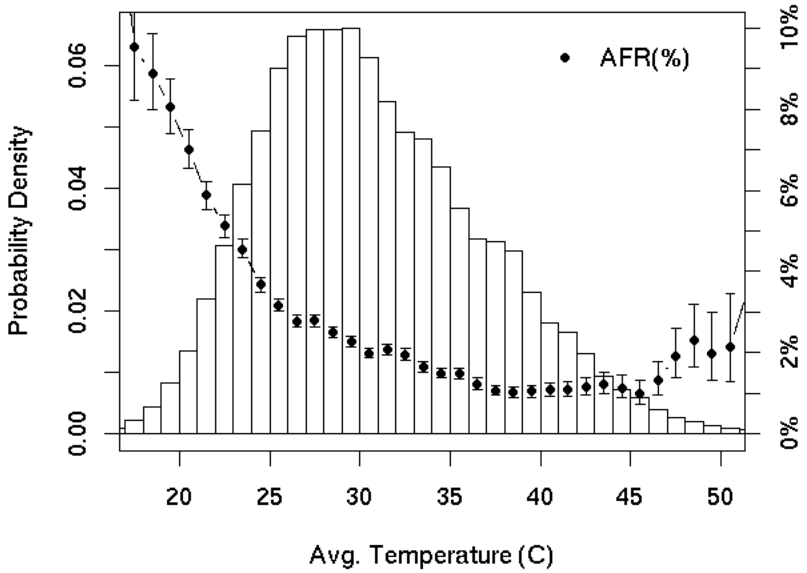


Figure 4: Distribution of average temperatures and failures rates.

S.M.A.R.T.

- **Self-Monitoring, Analysis and Reporting Technolgoy**
- **Relevant Parameters**
  - Seek error rate
    - track was not hit
  - Raw read error rate
    - problems in the magnetic medium
  - hardware ECC recovered
    - recovered bits by error correction (not really alarming)
  - Scan error rate
    - at periodic check non repairable error occurs (problems in the magnetic medium)
  - Throughout performance
    - spinning rate problem
  - Spin up time
    - startup time
  - Reallocated sector count
    - number of used reserve sectors
  - Drive temperature
- **Informative parameters**
  - Start/stop count
  - Power on hours count
  - Load/unload cycle count
  - Ultra DMA CRC Error Count

Hard Disks

Special Issues

Landing Zones

- **Problem**
  - When a hard disk is switched off the head can damage the platter
  - Power loss
  - Sudden movements
    - can lead to permanent damage
- **Solution**
  - Extra landing zones
    - in the middle of the disk
    - outside of the disk at extra park situation

Sound Control

- **In desktop computers the working sound can irritate and disturb**
- **Solution**
  - Extra mode with
    - slower actuator arm movement
    - slower rotation time
- **Problem**
  - Performance slows down

Data Safety

- **Problem**
  - Resold or disposed hard disk still carry sensible data
  - Deleting data does not overwrite data
  - Overwriting does not completely erase the information
- **Solution**
  - Extra hardware (strong magnets, physical destruction)
  - Cryptographic algorithms for storing
  - Sophisticated overwriting algorithms

ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

Algorithms and Methods for Distributed Storage Networks

2. Hard Disks

Christian Schindelhauer

Albert-Ludwigs-Universität Freiburg
Institut für Informatik
Rechnernetze und Telematik
Wintersemester 2007/08