DAAD Summerschool Curitiba 2011

Aspects of Large Scale High Speed Computing Building Blocks of a Cloud

Storage Networks

1: Introduction to Storage systems and Technologies

Christian Schindelhauer

Technical Faculty

Computer-Networks and Telematics

University of Freiburg

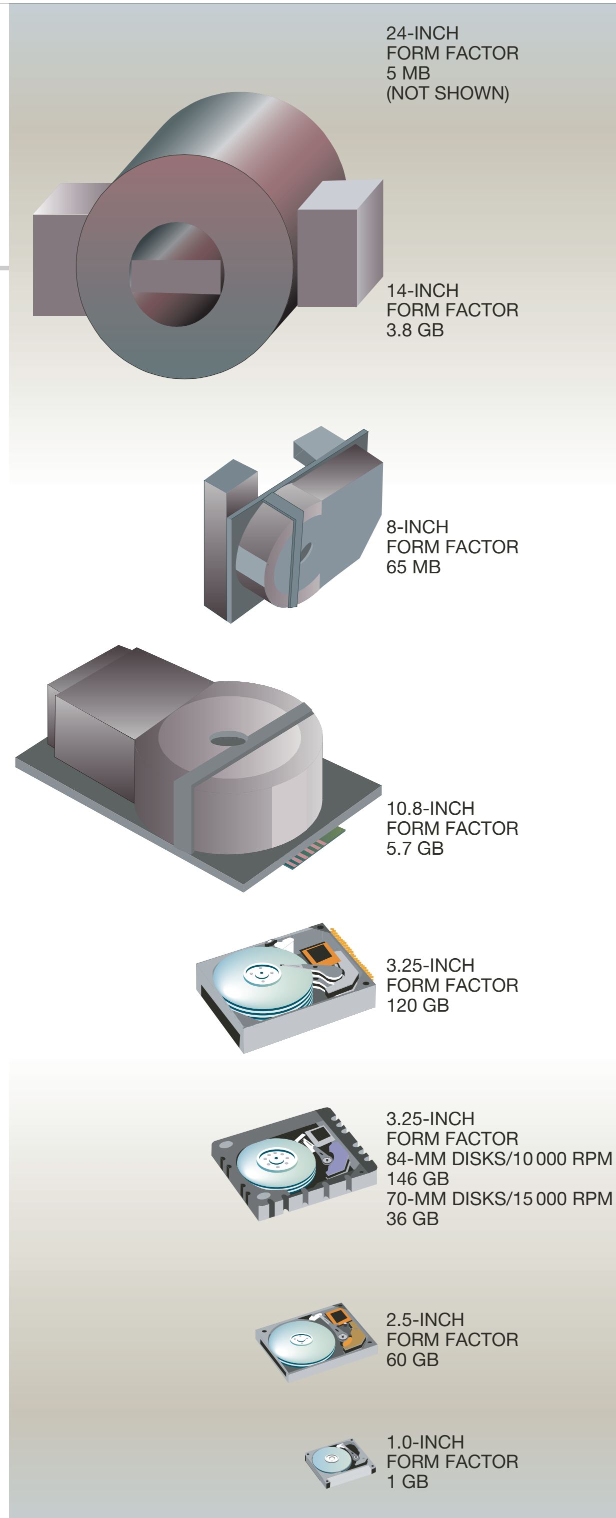
DAAD Summerschool Curitiba 2011

Storage Networks

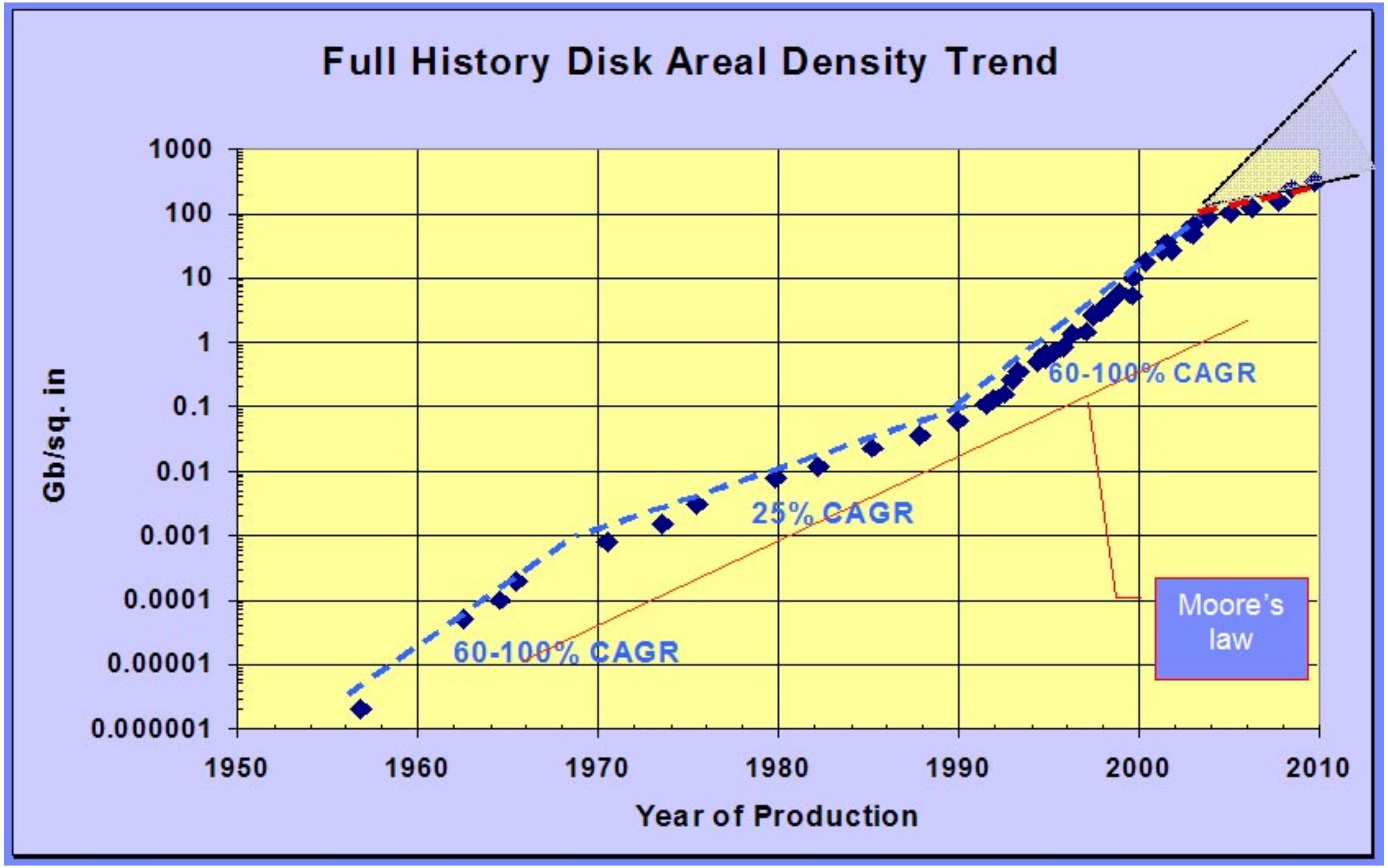
Motivation
Evolution of Disks

Evolution of Disk Form Factors

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

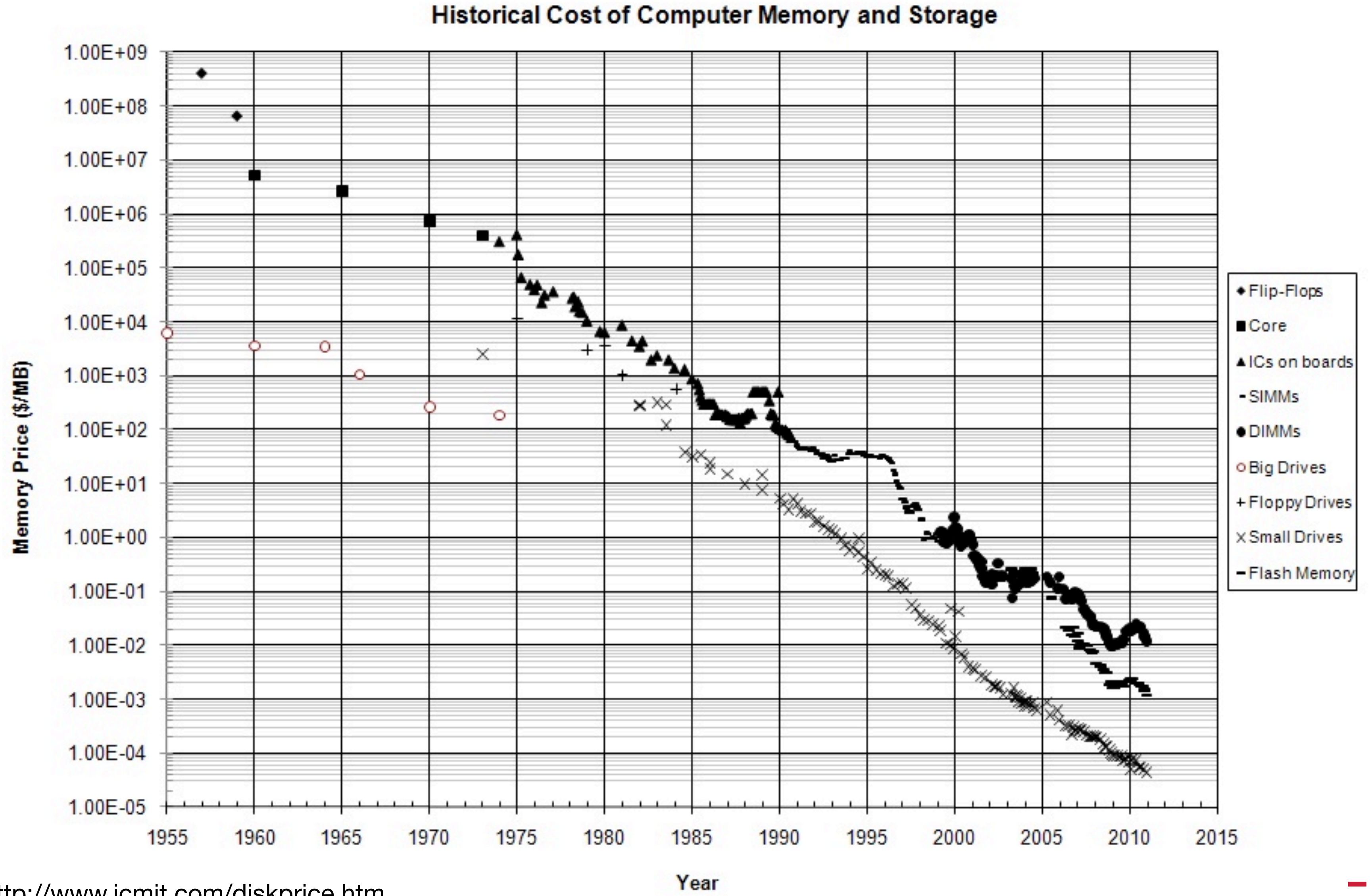


Increase of Density



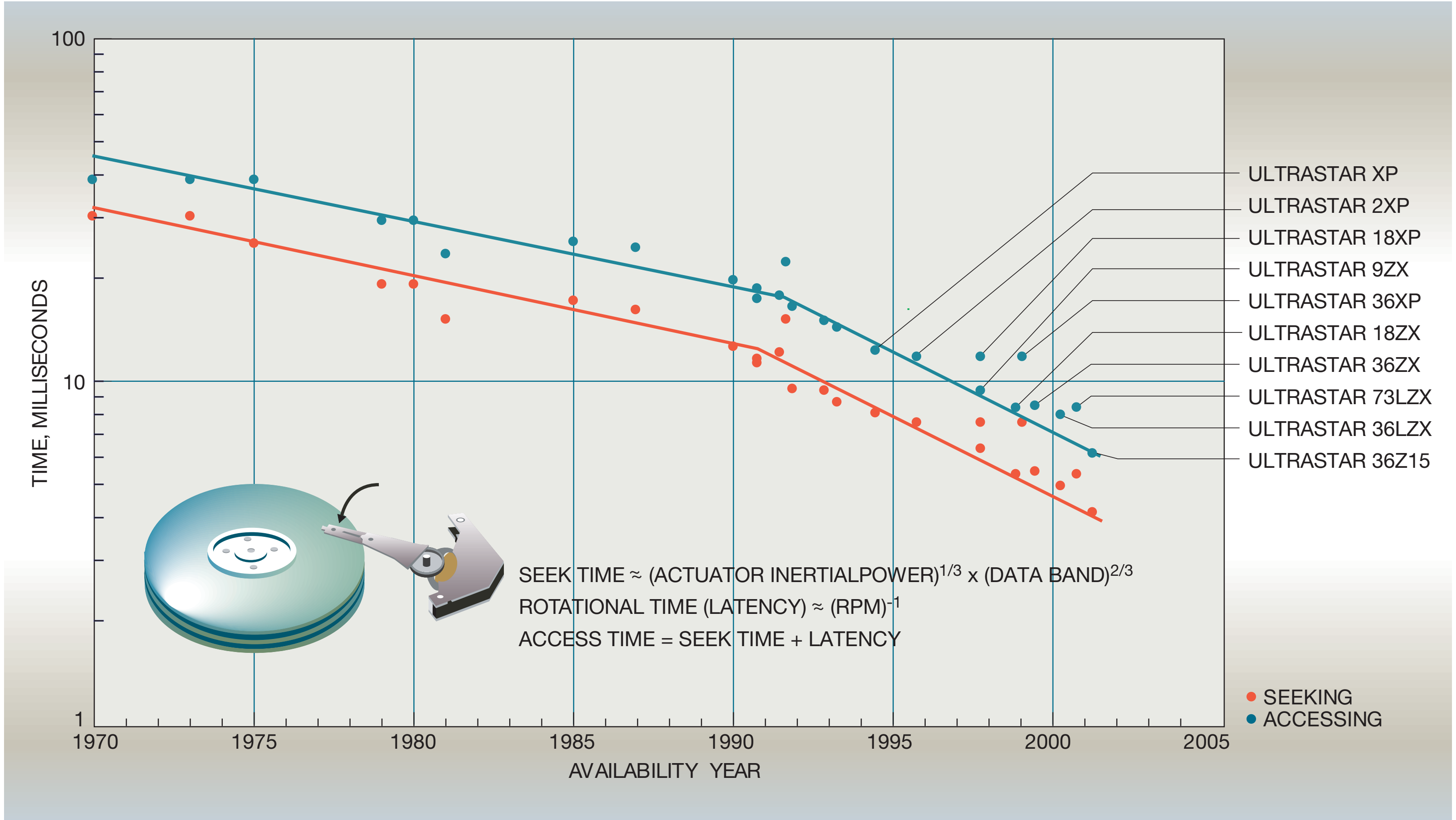
IBM, Barry White, 2011
https://www.ibm.com/developerworks/mydeveloperworks/blogs/storagevirtualization/tags/disk?lang=en

Historical Cost of Computer Memory and Storage



http://www.jcmit.com/diskprice.htm
Data Last Updated on 2010 Dec 10

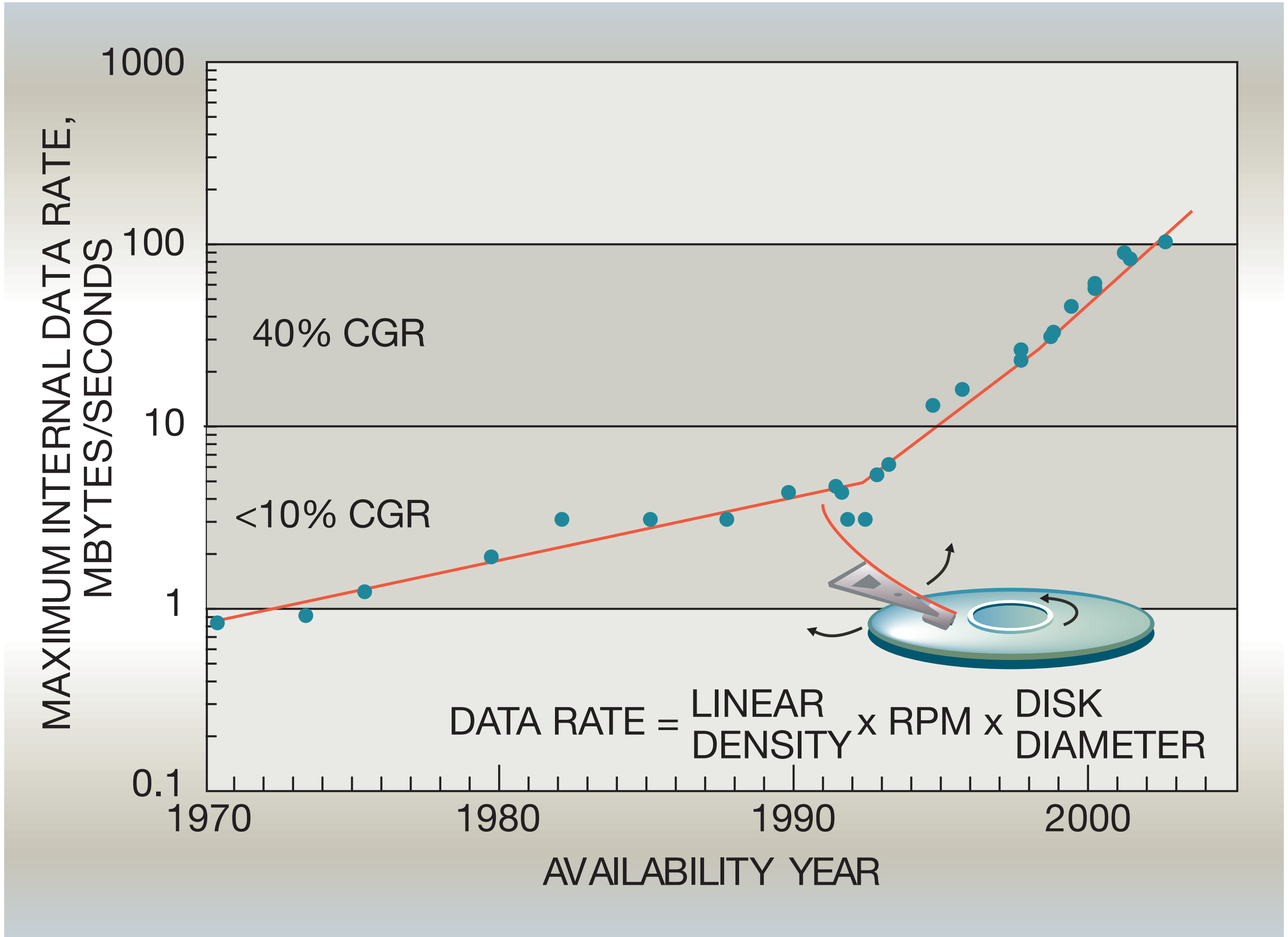
Increase of Speed



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

Increase of Speed

Figure 10 Hard disk drive maximum internal data rate for enterprise/server drives



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

Motivation
Consumer
Behavior

Consumer Usage

- Consumer Survey on Digital Storage in Consumer Electronics 2008, Coughlin Associates (Dec. 2007)
  - 51% said that 1 TB disk would be useful
  - Most storage of content was on hard disk
  - 46% backup data less than once per year
    - except pictures most of them do not backup
    - but most think it is important to have backups out of their homes
  - Most people want to store entire TV series, copies of their entire music collection
- Projection
  - by 2013 average home has 9 Terabyte
  - by 2015 user content sums up to 650 Exabyte

Peta, Exa, Zetta, Yotta

Prefixes for bit and byte multiples

Decimal			Binary				
Value	**SI**		**Value**	**IEC**		**JEDEC**	
1000	k	kilo	1024	Ki	kibi	K	kilo
1000^2	M	mega	1024^2	Mi	mebi	M	mega
1000^3	G	giga	1024^3	Gi	gibi	G	giga
1000^4	T	tera	1024^4	Ti	tebi		
1000^5	P	peta	1024^5	Pi	pebi		
1000^6	E	exa	1024^6	Ei	exbi		
1000^7	Z	zetta	1024^7	Zi	zebi		
1000^8	Y	yotta	1024^8	Yi	yobi		

Storage Hierarchy

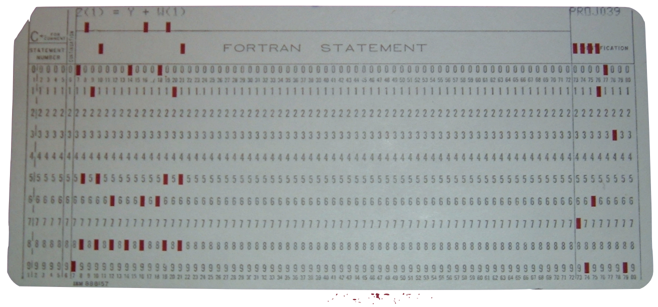
- Primary storage
  - Processors registers
  - Processor cache
  - RAM
- Secondary storage
  - Hard disks
  - Solid state disks
  - CD, DVD
- Tertiary storage
  - tape libraries
  - optical jukeboxes

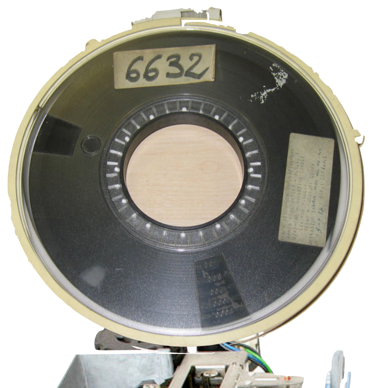
Characteristics of Storage

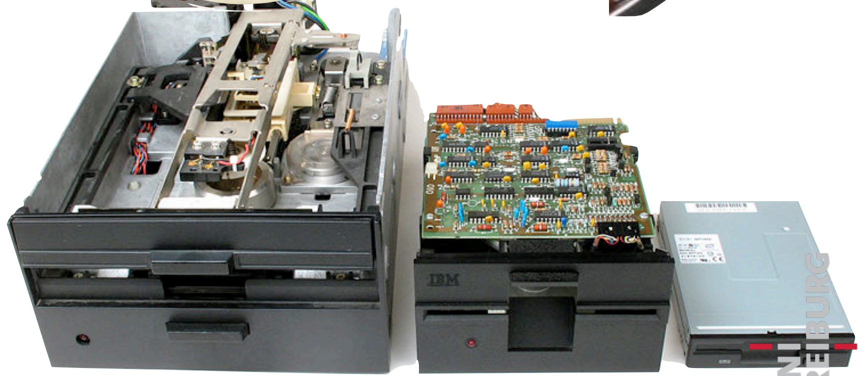
- Volatile — non-volatile memory
  - non-volatile: dynamic or static
- Read & write — Read only — Slow write, fast read
- Random access – Sequential access
- Addressability
  - location addressable
  - file addressable
  - content addressable
- Capacity
- Performance
  - Latency
  - Throughput

Non-volatile Storage Technologies

- Punch cards (Hollerith) 1886-1950s
- Magnetic tape data storage 1951-today
- Hard disk drive 1956-today
- Floppy disks 1970s-1990s
- EEPROM (Electrically Erasable Programmable Read-Only Memory) 1980-today
  - Flash memory
- Optical disc drive (read/write) 1997-today







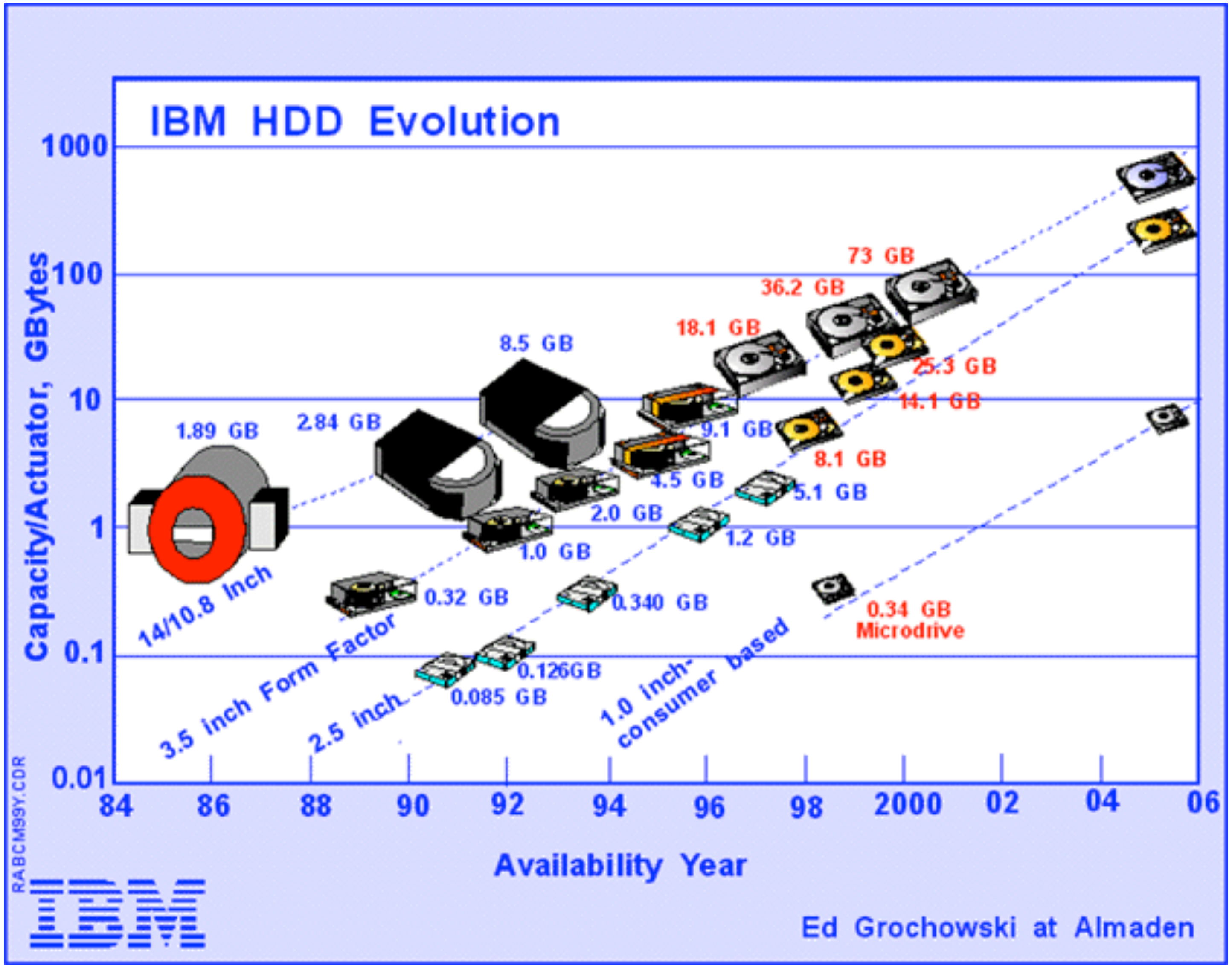
Network Storage Types

- Direct attached storage (DAS)
  - traditional storage
- Network attached storage (NAS)
  - storage attached to another computer accessible at file level over LAN or WAN
- Storage area network (SAN)
  - specialized network providing other computers with storage capacity with access on block-addressing level
- File area network (FAN)
  - systematic approach to organize file-related storage systems
  - organization wide high-level storage network

Hard Disks

History

Evolution of Hard Disk Capacity



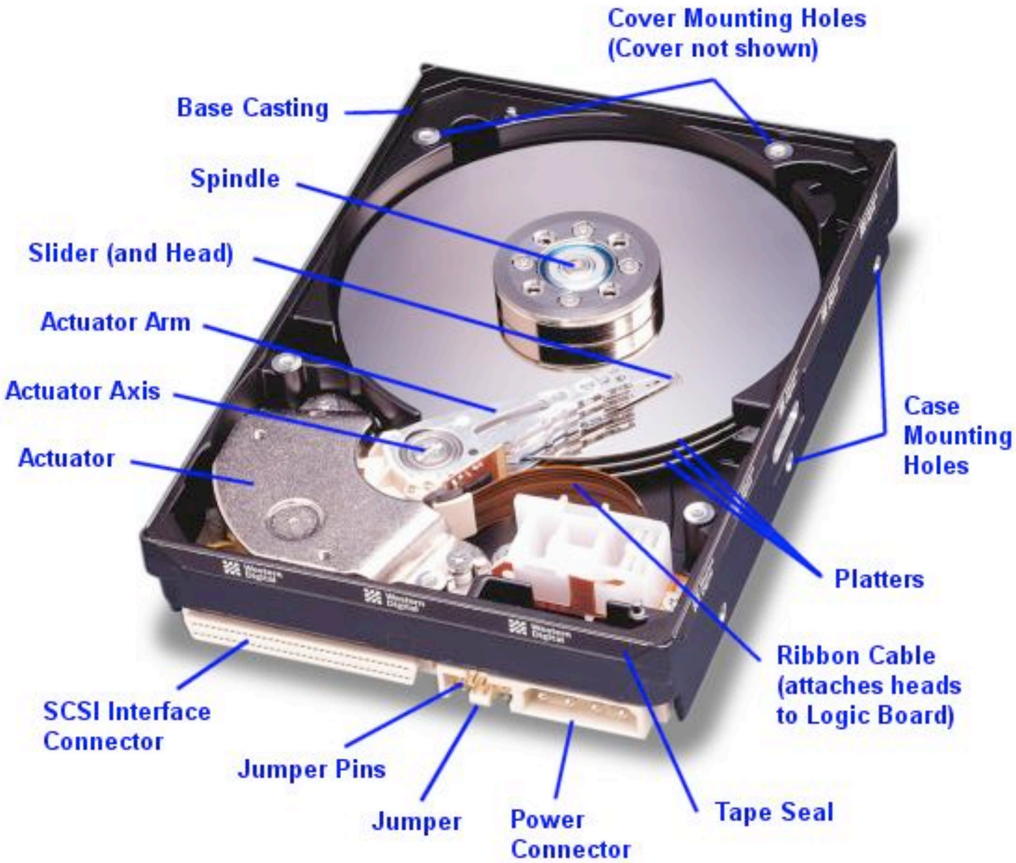
History

- 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

Hard Disks

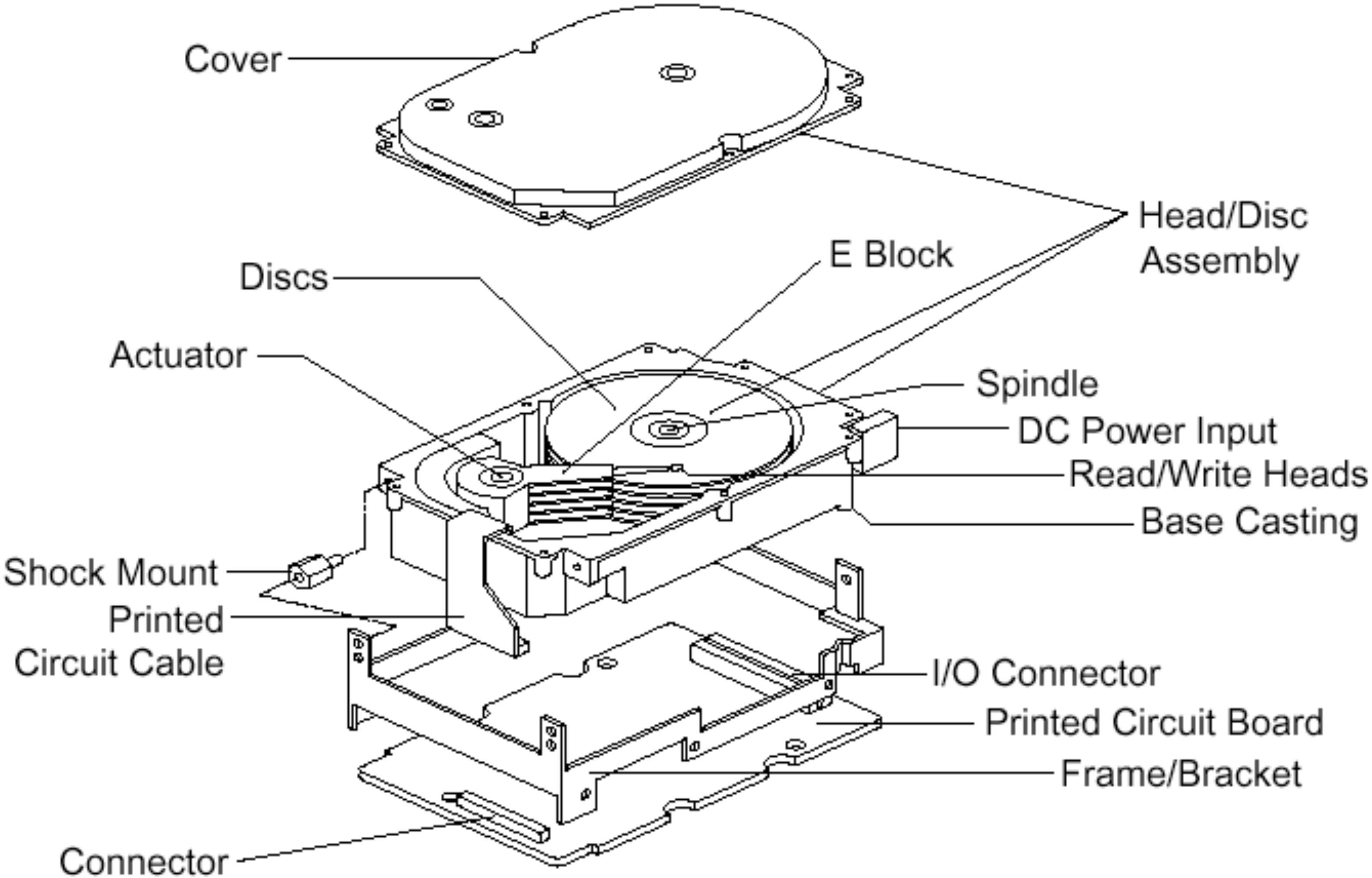
Construction and Operation

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

CoNe Freiburg

Slider

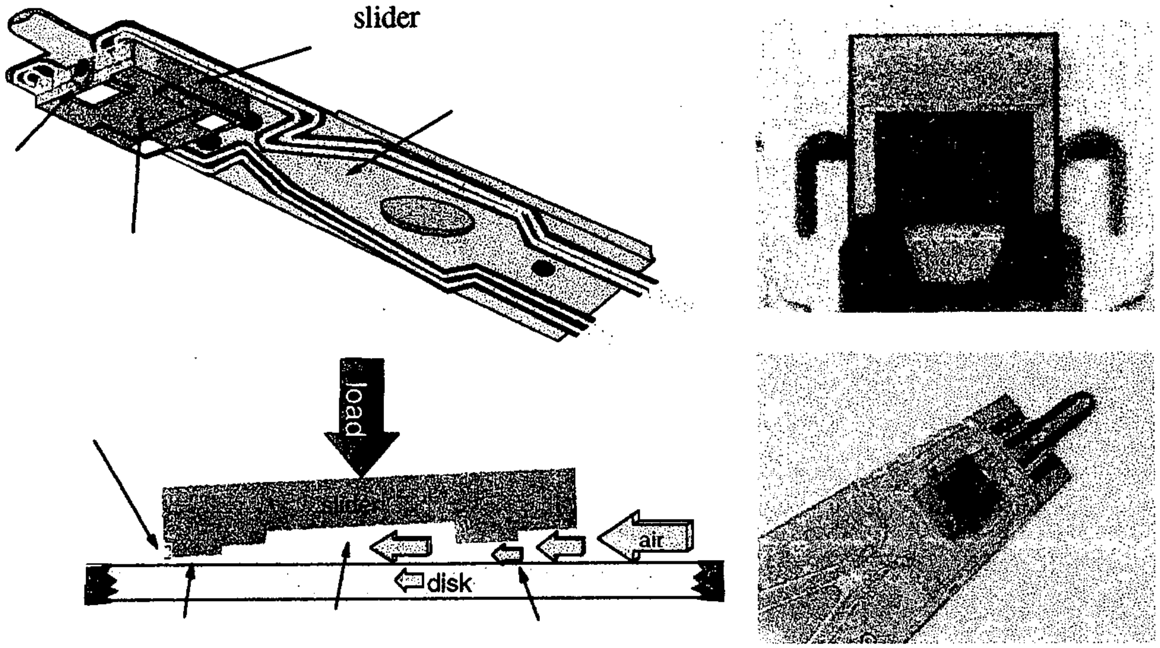


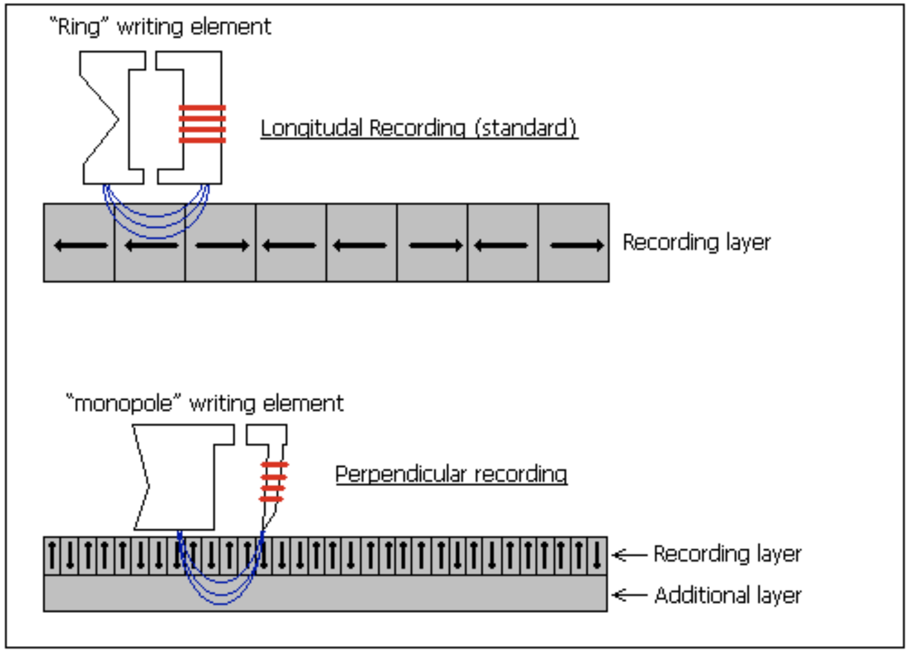
Figure 6. Illustration of suspension and slider. Left: schematic. Right: photograph. (Source: Tom Albrecht, IBM)

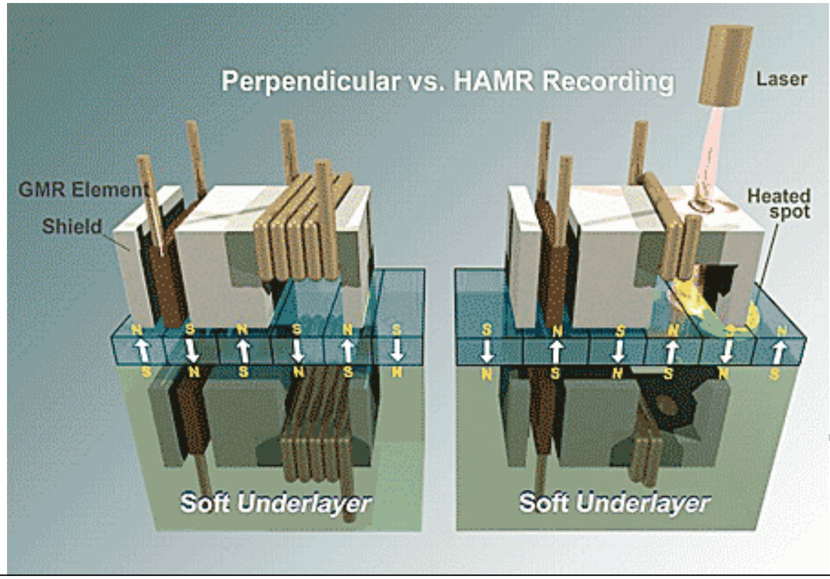
Proceedings of the American Control Conference ,Arlington, VA June 25-27, 2001
A Tutorial on Controls for Disk Drives William Messner , Rick Ehrlich

UNI FREIBURG

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

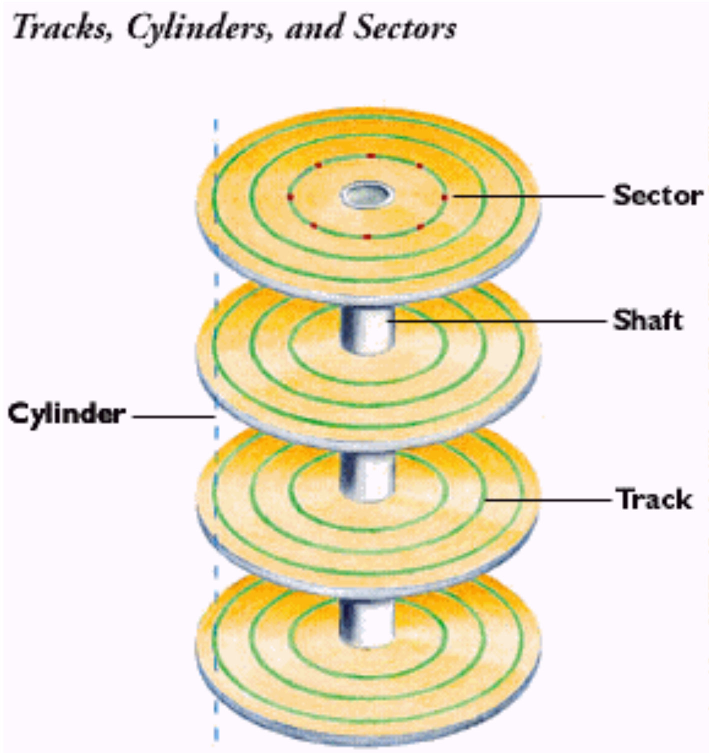
- Magnetized Surface on platter
- Read/Write-Head
- Embedded controller
- Disk buffer (disk cache)
  - store bits going to and from the platter
  - read-ahead/read-behind
  - speed matching
  - write acceleration
  - command queueing
- Interface

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

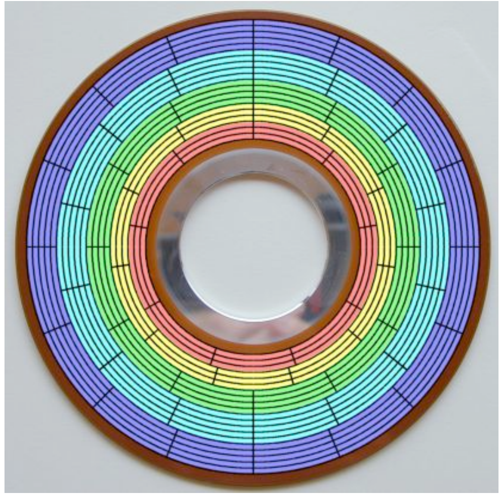


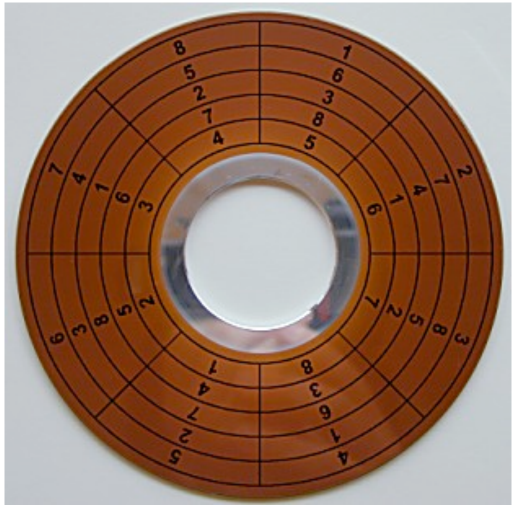
(c) Quantum Corporation

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

Adapting Sectors





- Zoned bit recording
  - adapt the sector size to the bit density
  - 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

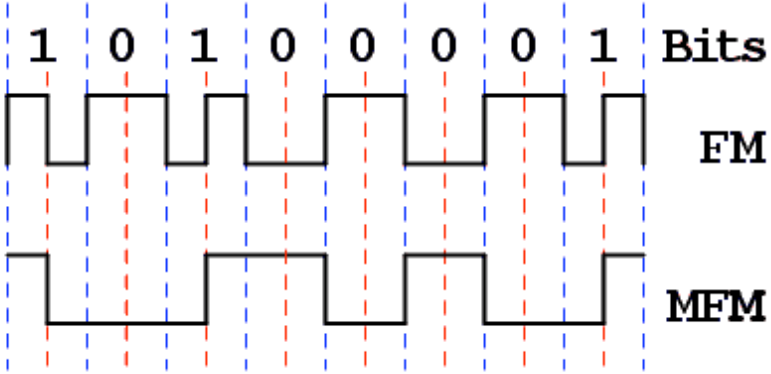
http://www.storagereview.com/guide2000/ref/hdd/geom/tracksZBR.html

Encoding

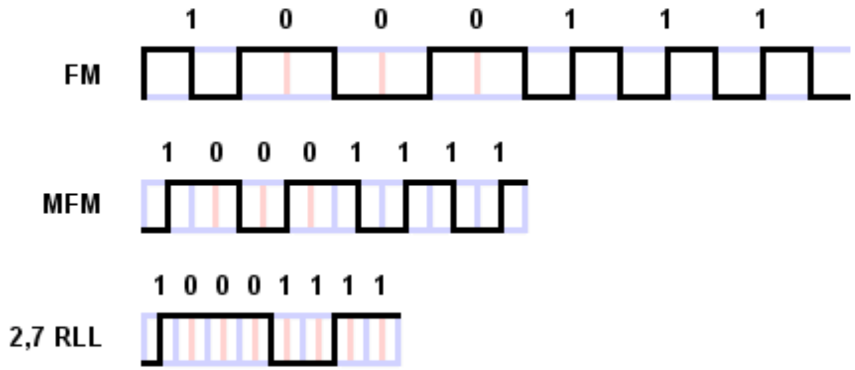
- Problem
  - Only the difference of orientation can be measured
  - Because of the para-magnetic effect orientation changes need a minimum distance
  - Long sequences of same orientation lead to errors
- Encoding
  - must have long, but not too long flux reversals

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)

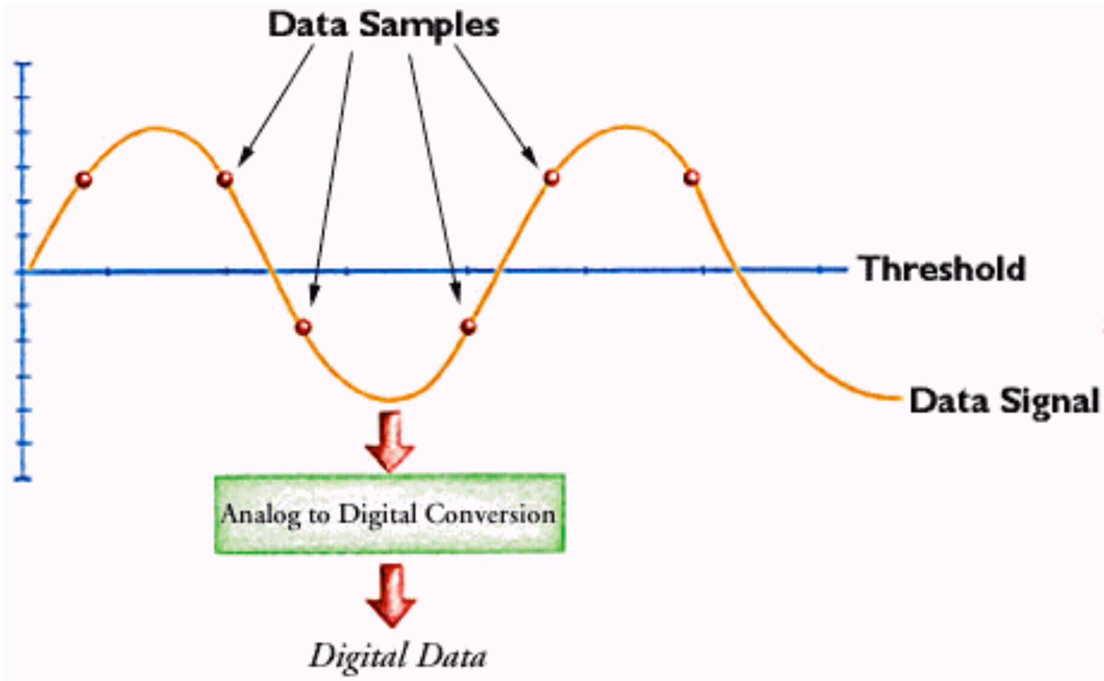


Bit Pattern	Encoding Pattern	Flux Reversals Per Bit	Bit Pattern Commonality In Random Bit Stream
11	RNNN	1/2	25%
10	NRNN	1/2	25%
011	NNRNNN	1/3	12.5%
010	RNNRNN	2/3	12.5%
000	NNNRNN	1/3	12.5%
0010	NNRNNRNN	2/4	6.25%
0011	NNNNRNNN	1/4	6.25%
Weighted Average		0.4635	100%

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

Lifetime and Disk Failures

Disk Failure Rates

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

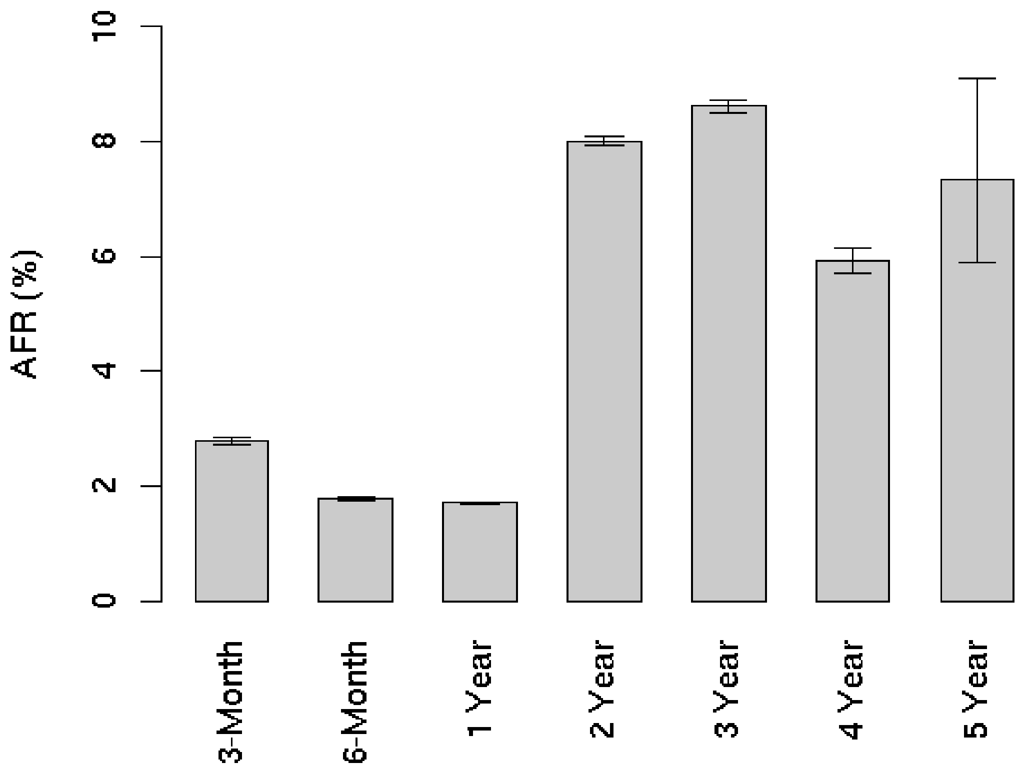


Figure 2: Annualized failure rates broken down by age groups

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
-

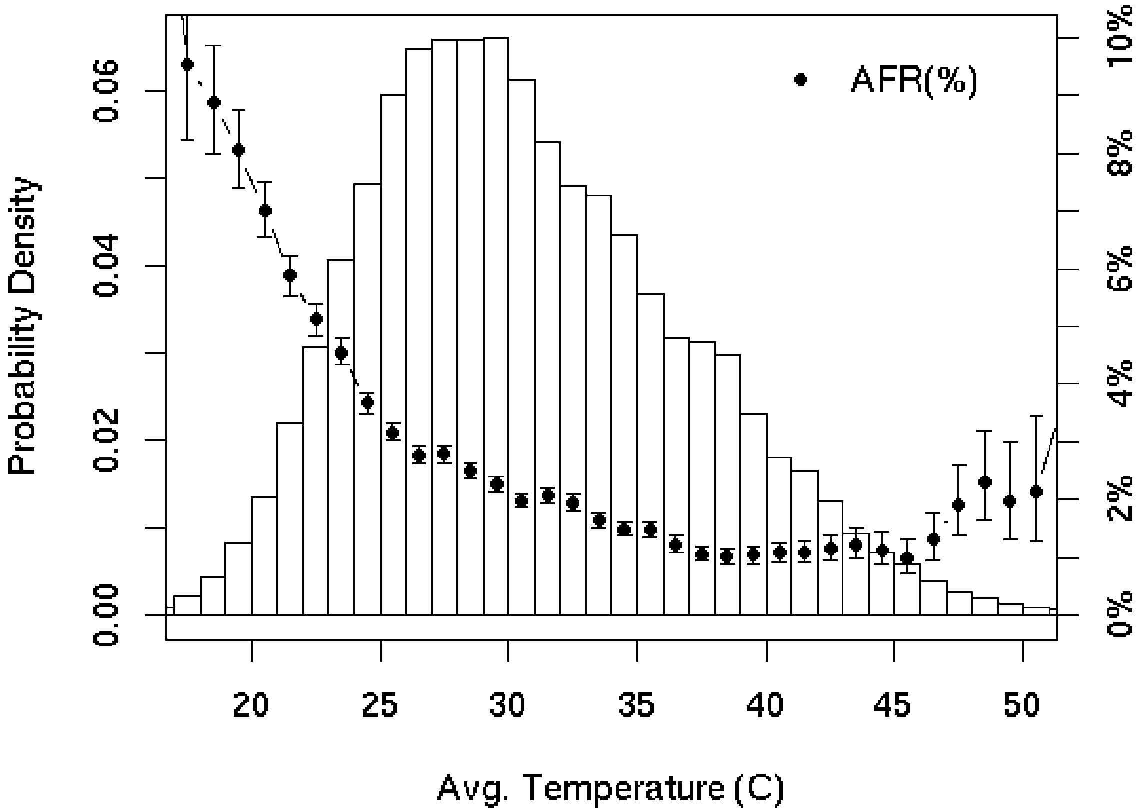


Figure 4: Distribution of average temperatures and failures rates.

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)
  - Throughput 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

DAAD Summerschool Curitiba 2011

Aspects of Large Scale High Speed Computing Building Blocks of a Cloud

Storage Networks

1: Introduction to Storage systems and Technologies

Christian Schindelhauer
Technical Faculty
Computer-Networks and Telematics
University of Freiburg