PRINCIPLES OF DATA ORGANISATION

Data Storage

MOTIVATION

- We need to store data ..

The block defines the smallest data unit processed. Within any application (file system included), we get blocks not bytes. If you want to read a particular byte, you need to ask for a block and to read this block to get your byte.

- We need to store blocks, what are the options?

OUTLINE

- Memory classification/hierarchy
- Primary storage
- Secondary storage
- Tertiary storage
- Magnetic disk
- Solid State Drive
- Disk interface
- Optical Disk
- Magnetic Tape
- Hierarchical storage management

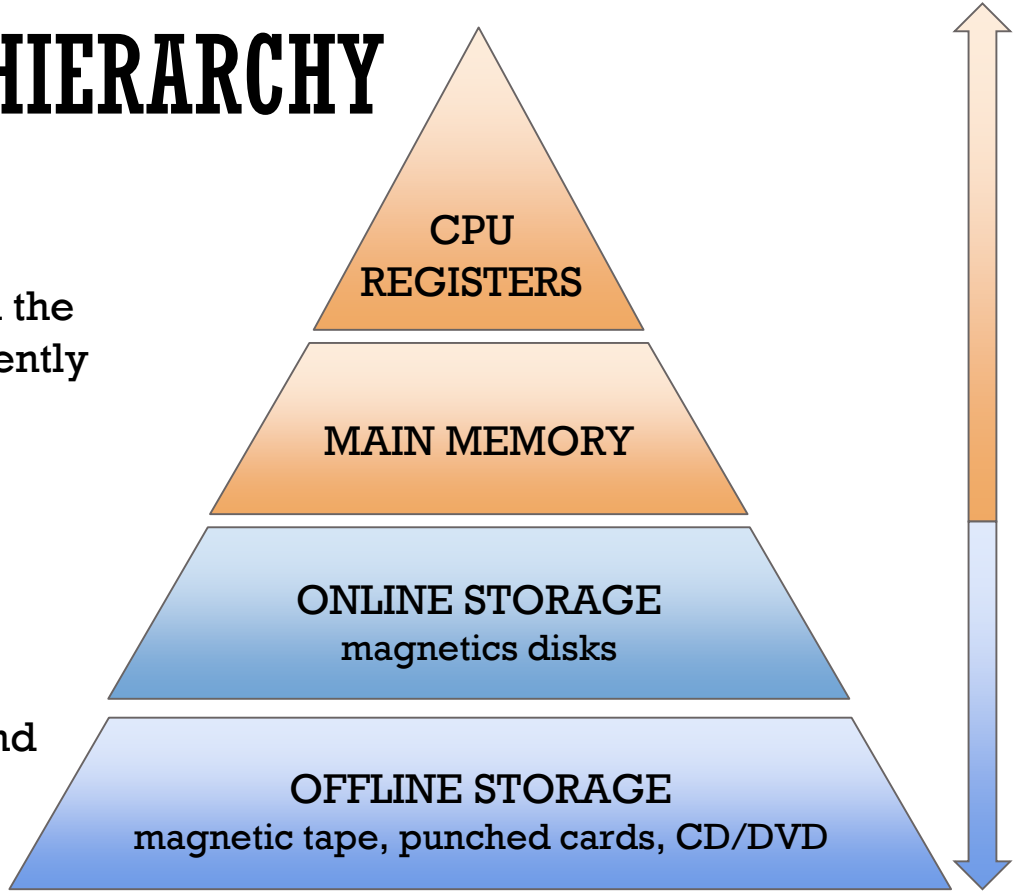
MEMORY CLASSIFICATION

- Mutability
  - read only
  - read/write
  - WORM (Write Once Read Multiple)
  - slow write/fast read
- Accessibility
  - random access
  - sequential access
- Performance
  - Latency
    - Time from request to data
    - Random access – latency is independent of the location
  - Throughput
    - How much data we read per a unit of time
- Cost
  - per data units
    - per GB, TB, …
  - total
- Capacity
- Volatility
  - volatile
    - CPU registers, main memory
  - non-volatile
    - DVD

MEMORY HIERARCHY

Memory
holds information the processor is currently using

Storage
preserves data and programs for future use



- Fast
- Expensive
- Small capacity

- Slow
- Inexpensive
- Large capacity

MEMORY HIERARCHY

Primary memory

- fastest
- volatile

- CPU registers
- caches
- main memory

Secondary memory

- moderate access time
- non-volatile
- not accessible by the CPU

- online storage
- magnetic disks
- SSD disks

Tertiary memory

- slow access time
- non-volatile
- offline storage (removable)

- floppy disks
- optical disks
- magnetic tapes

PRIMARY MEMORY

Register

- inside processor
- volatile
- used by arithmetic and logic unit
- usually word-sized
  - 32/64 bit (word of data)
- fastest and most costly

Cache

- inside the processor or disk
  - for instructions, for data, …
  - can be hierarchised
- volatile
- most often used data from main memory are stored in a CPU cache
- managed by HW or an operating system

Main memory

- general-purpose machine instructions operate on data resident in the main memory
- fast access, but generally too small to store the entire data set
- volatile
- connected to the processor

EXAMPLE

caches

main memory

Intel Sandy Bridge	**Registry**	L1	L2	L3	DDR3-1600 D
Latency (cycles)	0	4	12	26-31	~120
GB/s [3 GHz CPU]	480	36-144	96	96	25.6

SECONDARY MEMORY

Magnetic disk

- non-volatile
- data must be moved from disk to main memory for access and written back to storage
- random access
  - not 100% same time, roughly

Flash memory

- non-volatile
- memory cards, USB disks, solid-state drives (SSD)
- random access

TERTIARY MEMORY

Optical disk

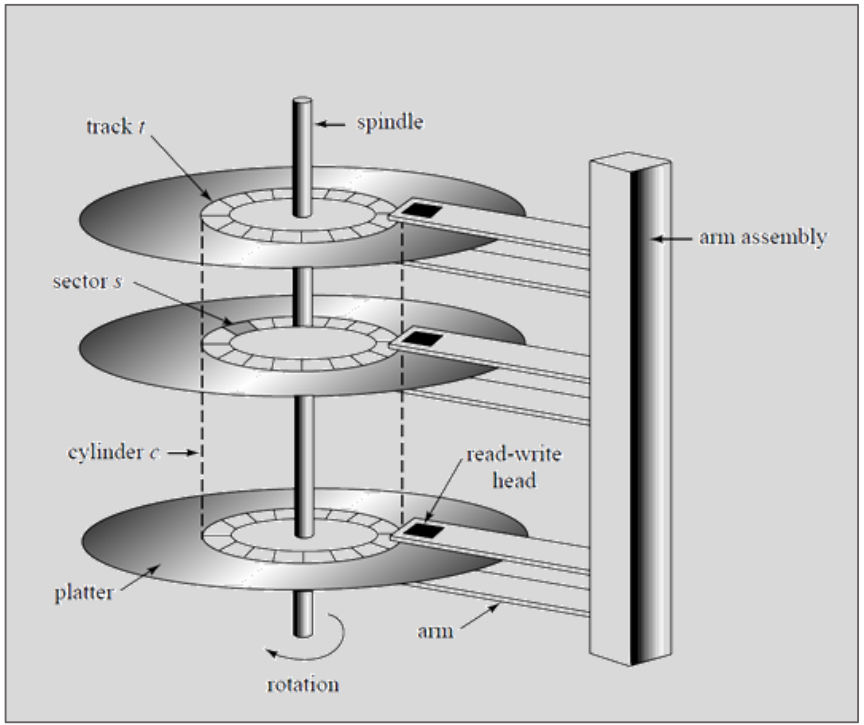
- non-volatile
- CD ROM, DVD ROM, Blu-ray, …

Magnetic tape

- non-volatile
- sequential access
- very high capacity and persistence
- cheap
- used for backup

MAGNETIC DISK

MAGNETIC DISK



- Disk pack consists of multiple platters on a spindle
  - Platters are usually double-sided
- Data read by read-write head
  - Kept on an arm
  - Arms kept on the arm assembly
  - 2 read-write heads (1 head per surface)
- Surface of platters divided into tracks
- Tracks are divided into sectors
  - Smallest unit to be read/written
- Set of all tracks with the same diameter foRm a cylinder

MAGNETIC DISK

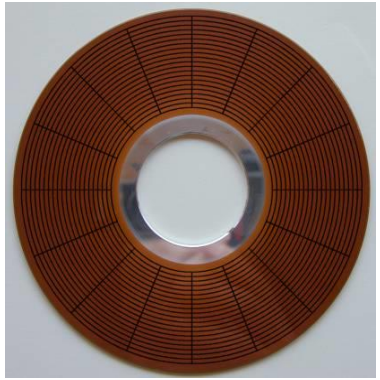
Sector

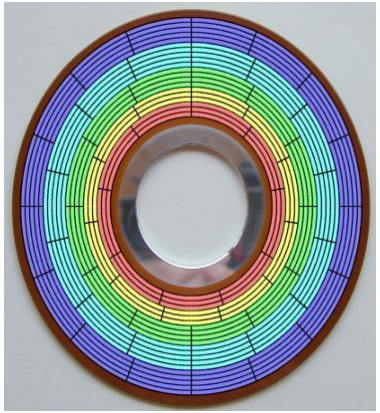
- define a minimum amount of information to read or write
  - Not a bit or byte
- smallest addressable unit
- 512B, 4KB (standard nowadays)

MAGNETIC DISK – ZONE BIT RECORDING

- Earlier disks had the same number of sectors per track
  - inner tracks as dense as possible
  - outer tracks underutilised by reducing bit density
    - Wasting space

- Zone bit recording
  - tracks grouped into zones
  - each zone is assigned several sectors per track
  - tracks close to the outer edge contain more sectors per track
  - example: 13 zones, 600-1200 tracks in the zone, 400-800 sectors per track, speed 189 – 372 MBits/s





MAGNETIC DISK – ADDRESSING

- Using the physical build-up of early drives ~ geometry-based addressing
- Cylinder-Head-Sector address
  - 10 bits – cylinder (C)
  - 8 bits – head (H)
  - 6 bits – sector (S)
- Drawbacks:
  - 24 bits = maximum active primary partition size 2^24 * 512 B = 8 GiB
    - Not enough today
  - Does not map well to other devices like tape, SSD disk.

MAGNETIC DISK – ADDRESSING

- Logical block address
- Linear addressing space starting with 0
- Each sector has unique number
- Must be supported by disk, BIOS, OS
  - Nowadays common
- Drawback:
  - Hides physical details of the storage device (cannot be used)
- Cylinder-Head-Sector to Logical Block Address (LBA):

$$LBA = (C * number\_of\_heads + H) * sector\_per\_track + (S - 1)$$

MAGNETIC DISK – PARAMETERS

- How fast we can read/write blocks?
- s – seek
  - average seek time from one random track (cylinder) to any other
  - 3ms – 15 ms, usually between 8 and 12 ms
- r – rotational delay (latency)
  - one revolution = 2r (r is average latency)
- RPM – revolutions per minute
  - 4,200 – 15,000
  - more revolutions → more energetically demanding
- btt = block transfer time
  - Reading a block = seek the cylinder and wait for the rotation (latency)

Speed (RPM)	Average latency
15,000	2 ms
10,000	3 ms
7,200	4.16 ms
5,400	5.55 ms

MAGNETIC DISK – PARAMETERS

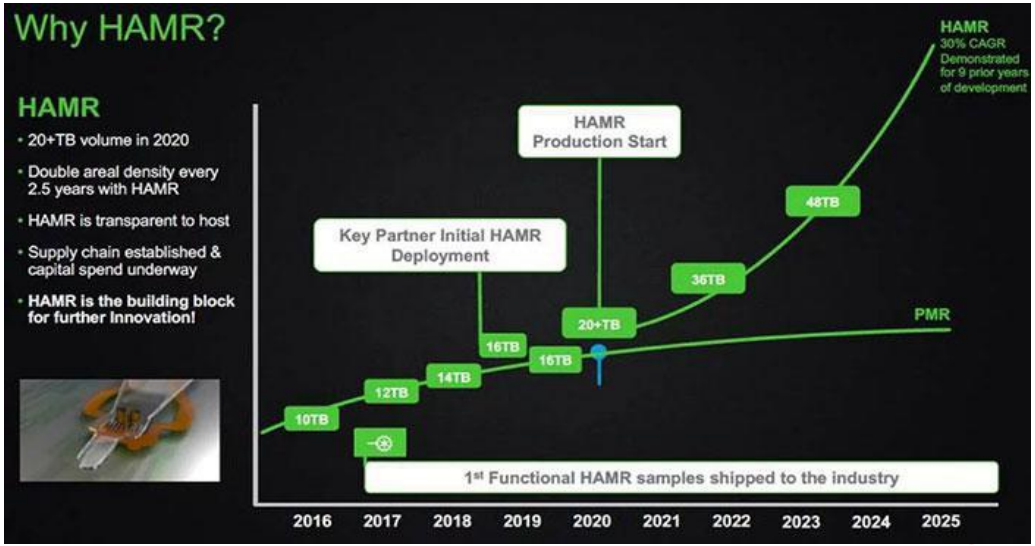
- (average) media transfer rate
  - speed of reading/writing bits from/to a single track of one surface of the disk
    - Data smaller than one track
    - Tracks have different sizes
- interface/external transfer rate
  - the speed with which the bits can be moved to/from the hard disc platters from/to the hard disc's integrated controller
  - purely electronic operation = much faster than the mechanic ones
- (**average**) sustained/sequential transfer rate
  - real-world transfer rate when a file spans multiple platters and cylinders
  - media transfer rate + head switch time (electronic operation) + cylinder switch time
  - ~ 100-200MB/s.

MAGNETIC DISK – FUTURE?

2018.06.11
New Storage Roadmap shows **100TB HDDs in 2025**

HAMR = Heat-Assisted Magnetic Recording

Idea: increase data density



MAGNETIC DISK – FUTURE?

2019.01.08
MG08 Series

Idea: HDD is filled with helium, thus it can fit more plates

Formatted Capacity	16 TB
Buffer Size	512 MiB
Data Transfer Speed (Sustained)	262 MiB/s
Rotation Speed	7,200 rpm
Sector	4K native 512 emulation

SOLID STATE DRIVE

- Does not contain moving mechanical components
- Flash memory
  - Data is stored in an array of unipolar floating gate transistors, called "cells", each typically holding 1 bit or today 3 bits or more of information
- Interface emulates HDD interface
- Embedded processor
  - data striping
  - data compression
  - caching

- **separate lecture later**

SOLID STATE DRIVE

Advantages of SSDs

- silent
- lower consumption
- more resistant to shock and vibration
- lower access time
  - no need to move heads
- higher transfer rates
  - up to 500MB/s or even higher in enterprise-level solutions
- does not require cooling

Disadvantages of SSDs

- lower (affordable) capacity
- higher cost
  - for larger storage capacity
- limited lifetime
  (writing to the same spot)
- as not an issue with a typical IO load

.. subject to change ...

HDD / SSD – SUBSYSTEMS

Controller

- The interface between disk and the system
- Accepts instructions to read/write data
- Multiple speaking with each other
  - On the side of the motherboard
  - On the side of the disk
- Include logic for checksum, validation, and remapping bad sectors

Bus – disk interface

- Bus is a physical and logical infrastructure for transferring data between components
- How we connect the disk to a motherboard
- PATA, SATA, Fiber Channel, SCSI, …

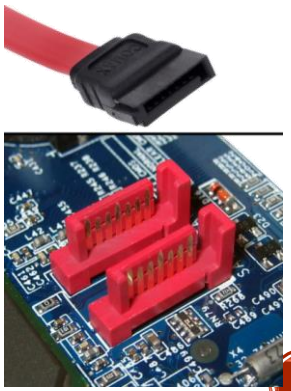
DISK INTERFACE

PATA
(Parallel Advanced Technology Attachment)

- originally called ATA
- **parallel**
- Can transfer up to 167 MB/s

SATA
(Serial ATA)

- enables hotplug
- **serial**
- modifications for different device types
  - eSATA
  - mSATA
- up to 600 MB/s



DISK INTERFACE

SCSI (Small Computer System Interface)

- set of standards for transferring data between computer and devices
  - magnetic disks, optical drives, printers, …
- allows to connect up to 16 devices to a single bus
- up to 640 MB/s

Fiber Channel

- mainly for storage networking (SAN – storage area network)
- Fiber Channel Protocol
- up to 12800 MB/s (128 Gigabit)

DISK ATTACHMENT STRATEGIES

DAS (Direct Attached Storage)

- disk inside a computer
- **block-level** storage
- ATA, SATA, Fibre Channel, …

NAS
(Network Attached Storage)

- uses a network
- **file-level** storage
- accessed by mapping (\\NAS\share)
- file system managed by NAS OS
- for data backup
- self-contained solution
- NFS (Unix), SMB/CIFS (Windows)

SAN
(Storage Area Network)

- enterprise solution
- **block-level** storage
- iSCSI, Fibre Channel, FCoE
- usually only server accesses SAN (not clients)
- OS sees it as a local hard drive

OPTICAL DISK

- CD, DVD, Blu-ray
- Based on reflection (pit/bump ~ 0/1)
- Data stored by laser and read by laser diode when spinning in the optical disc drive

- On a decline nowadays

MAGNETIC TAPE

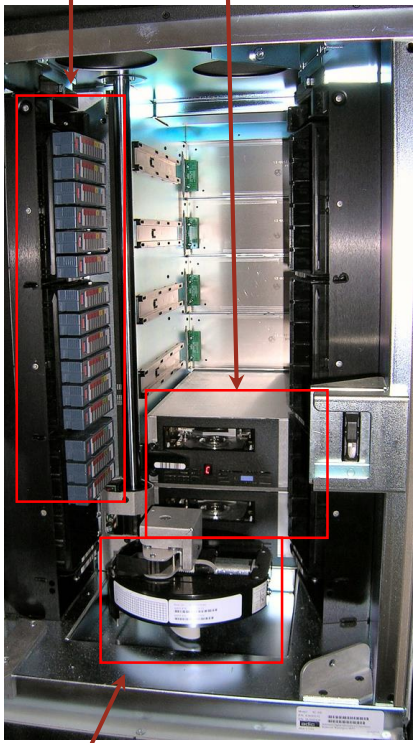
- Magnetizable coating on a long, narrow strip of plastic film
- **Sequential access**
- Low cost per bit – available surface area on a tape is far greater than for HDD
- Originally main secondary storage
- Transfer rate comparable to magnetic disks
- Automatic change of tapes

- Still popular

TAPE LIBRARIES

1,000,000,000,000,000 bytes

- Capacity up to hundreds of petabytes of data
- Price up to $1 million
- Tape robot, tape jukebox
  - tape drive(s)
  - tape cartridges
    - barcode reader to identify



TAPE LIBRARIES

2018.08.02

IBM Achieves the World's Highest Areal Recording Density for Magnetic Tape Storage

The latest achievement has the potential to store 330 terabytes of uncompressed data on a single tape cartridge that would fit in the palm of your hand.

HIERARCHICAL STORAGE MANAGEMENT

- Using various types of storages to increase usable capacity with limited costs
- Less often used data moved to cheaper storages with higher capacity → tiers
- Conceptually analogous to the (multi-level) cache
- Moving of data is managed by a migration policy
- May and may not require special commands