Database Access Methods

Data Storage

Motivation

- We need to store data ..

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

IO - Block (Batch on device)

- 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
  - throughput
- Cost
- Capacity
- Volatility
  - volatile
  - non-volatile

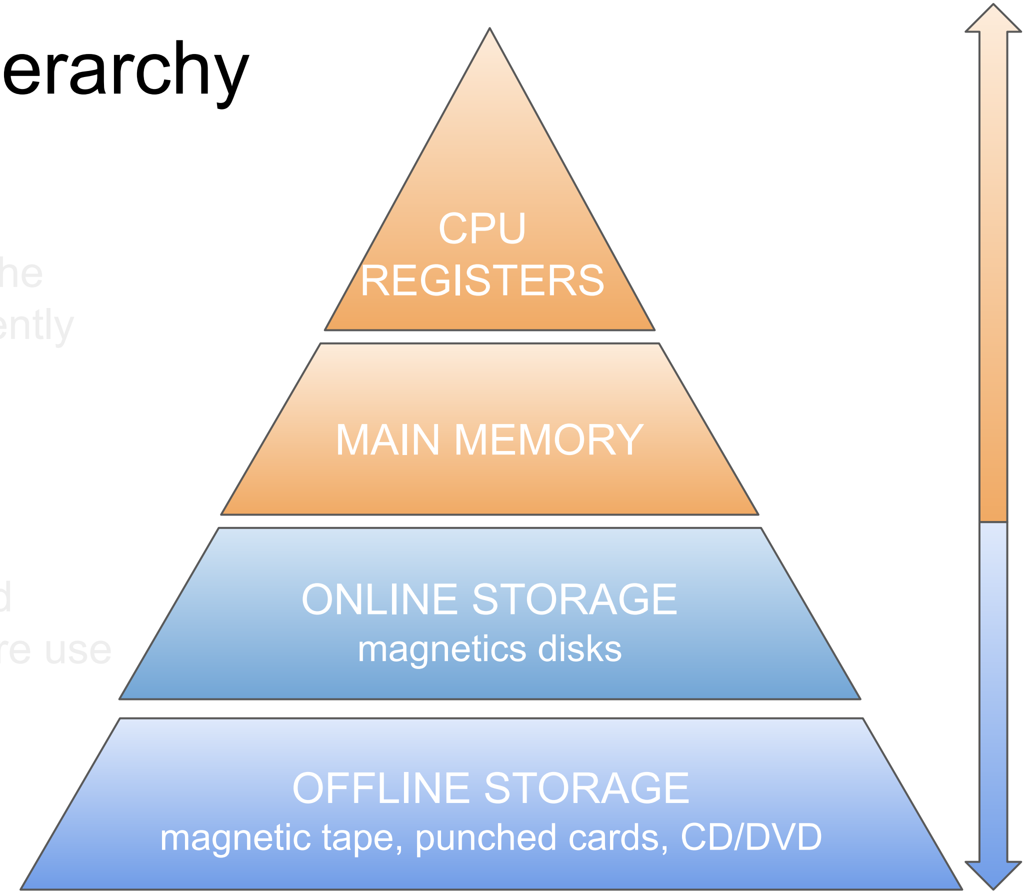
Memory Hierarchy

Memory

hold information the processor is currently using

Storage

preserve data and programs for future use



CPU REGISTERS

MAIN MEMORY

ONLINE STORAGE
magnetics disks

OFFLINE STORAGE
magnetic tape, punched cards, CD/DVD

- 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
- 32/64 bit (word of data)
- fastest and most costly

Cache

- inside processor or disk
- volatile
- most often used data from main memory are stored in a CPU cache
- managed by HW or operating system
- can be hierarchized

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

Related

High Performance Software Development - NPRG054

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

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

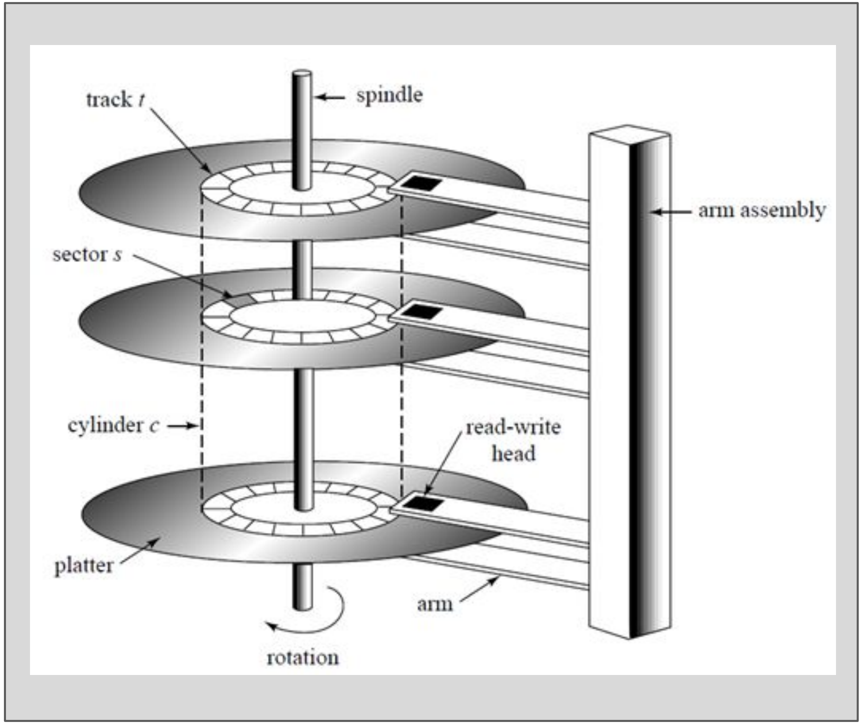
Outline

- Memory classification/hierarchy
- Primary storage
- Secondary storage
- Tertiary storage

- Magnetic disk
- Solid State Drive
- Disk interface
- Optical Disk
- Magnetic Tape
- Hierarchical storage management

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
  - 1 disk - 2 read-write heads (1 head per surface)
- Surface of platters divided into tracks
- Tracks are divided into sectors
- Set of all tracks with the same diameter from a cylinder

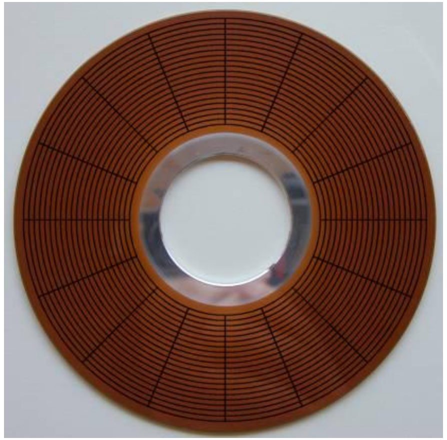
Magnetic disk

Sector

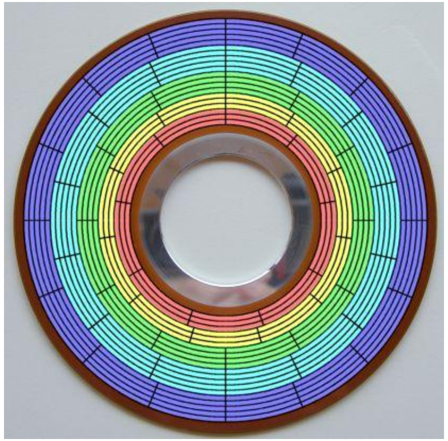
- define minimal amount of information to read or write
- smallest addressable unit
- 512B, 4KB (common nowadays)

Magnetic disk - Zoned Bit Recording

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



- Zoned bit recording
  - tracks grouped into zones
  - each zone assigned a number of sectors per track
  - tracks close to the outer edge contain more sectors per track



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)
- Maximum active primary partition size 2^24 * 512 B = 8 GiB
- 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
- Hides physical details of the storage device
- Cylinder-Head-Sector to Logical block address (LBA)

LBA = (C * number_of_heads + H) * sector_per_track + (S - 1)

Magnetic disk - Parameters

- 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 equals 2r (r is average latency)
- rpm - revolutions per minute
  - 4,200 – 15,000
  - more revolutions → more energetically demanding

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

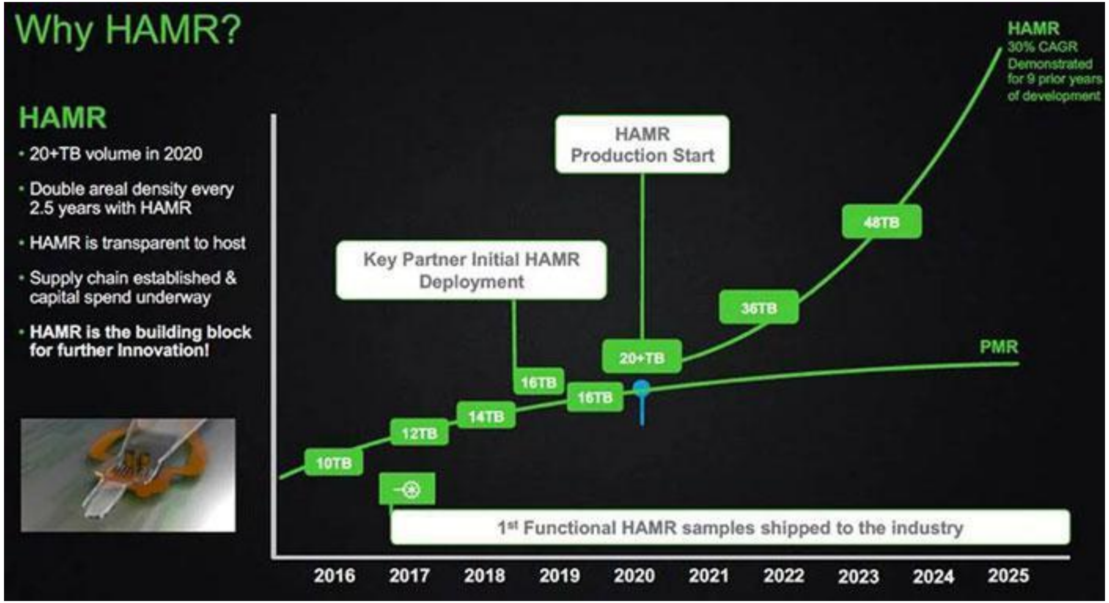
Magnetic disk - Parameters

- (average) media transfer rate
  - speed of reading or writing bits to a single track of one surface of the disk
- (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

Magnetic disk - Future

2018.06.11

New Storage Roadmap shows 100TB HDDs in 2025



Why HAMR?

HAMR

- 20+TB volume in 2020
- Double areal density every 2.5 years with HAMR
- HAMR is transparent to host
- Supply chain established & capital spend underway
- **HAMR is the building block for further Innovation!**

Magnetic disk - Future

2019.01.08

MG08 Series

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
- Interface emulates HDD interface
- Embedded processor
  - data striping
  - data compression
  - caching

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
- limited lifetime
  (writing to the same spot)
- as not an issue with a typical IO load

.. subject to change ...

HDD / SSD - subsystem

Controller

- Interface between disk and the system
- Include logic for checksum, validation, remapping bad sectors

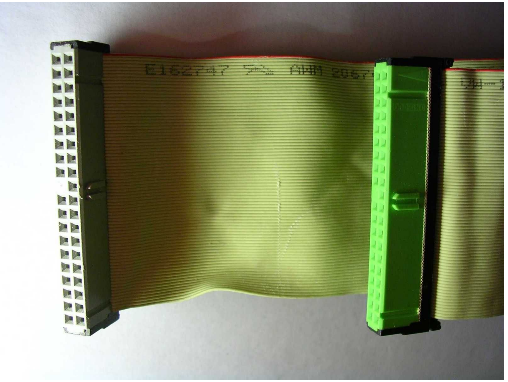
Bus - Disk interface

- Bus is physical and logical infrastructure for transferring data between components
- PATA, SATA, Fiber Channel, SCSI, …

Disk Interface

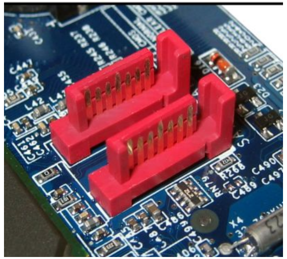
PATA
(Parallel Advanced Technology Attachment)

- originally called ATA
- parallel
- 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 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)

- block-level storage
- ATA, SATA, Fibre Channel, …

NAS
(Network Attached Storage)

- file-level storage
- accessed by mapping (\\NAS\share)
- file system managed by NAS OS
- self-contained solution
- NFS (Unix), SMB/CIFS (Windows)

SAN
(Storage Area Network)

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

Outline

- Memory classification/hierarchy
- Primary storage
- Secondary storage
- Tertiary storage

- Magnetic disk
- Solid State Drive
- Disk interface
- Optical Disk
- Magnetic Tape
- Hierarchical storage management

Optical Disk

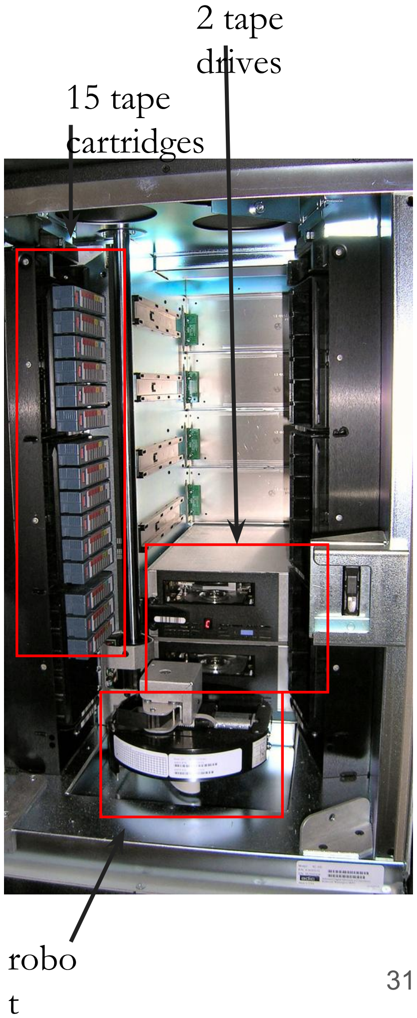
- CD, DVD, Blu-ray
- Based on reflectance (pit/bump ~ 0/1)
- Data stored by laser and read by laser diode when spinning in the optical disc drive

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

Tape Libraries

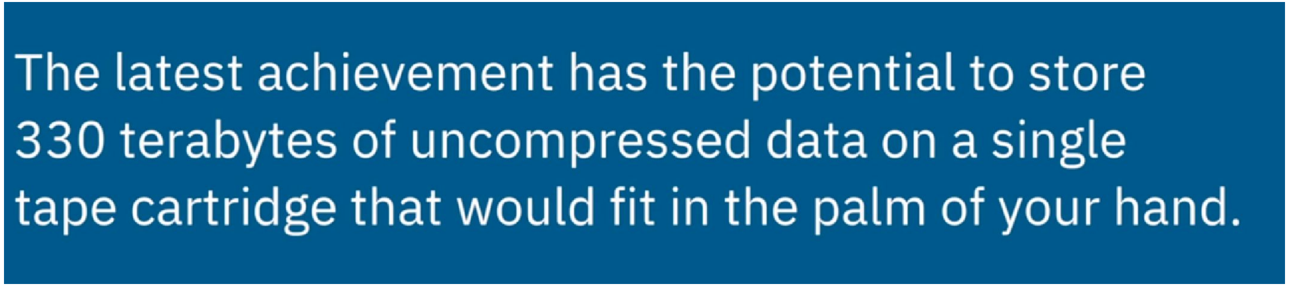
- Capacity up to hundreds of petabytes of data
- Price up to $1 million
- Tape robot, tape jukebox
  - ○ tape drive(s)
  - ○ tape cartridges
  - ○ robot
  - ○ barcode reader



Tape Libraries

2018.08.02

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



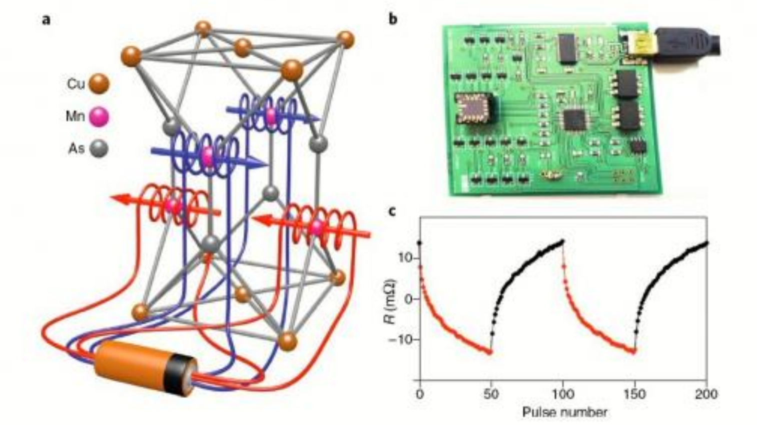
The latest achievement has the potential to store 330 terabytes of uncompressed data on a single tape cartridge that would fit in the palm of your hand.

Future

2018.06.16

Antiferromagnetic CuMnAs multi-level memory cell with microelectronic compatibility

They allow for counting and recording thousands of input pulses and responding to pulses of lengths downscaled to hundreds of picoseconds.
To demonstrate …. we implemented the antiferromagnetic bit cell in a standard printed circuit board managed …. a computer via a USB interface. …



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

Hierarchical storage management - CESNET

- CESNET DU4 ~ 23.341 PB
  - 755 x 10TB disks
  - 1167 x 15TB tape