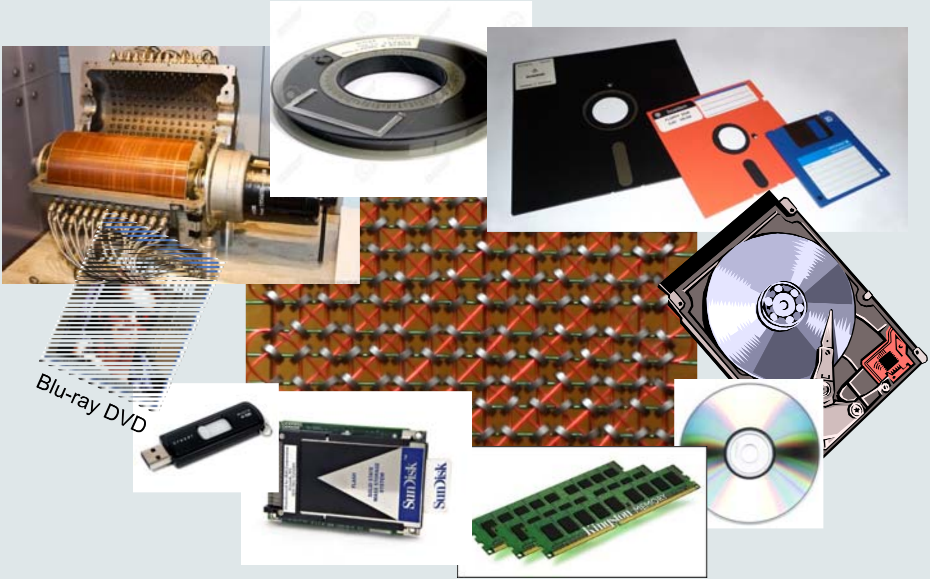
Lecture 2: Memory System

- Main memory

- Secondary memory

- Memory hierarchy

Many Different Technologies



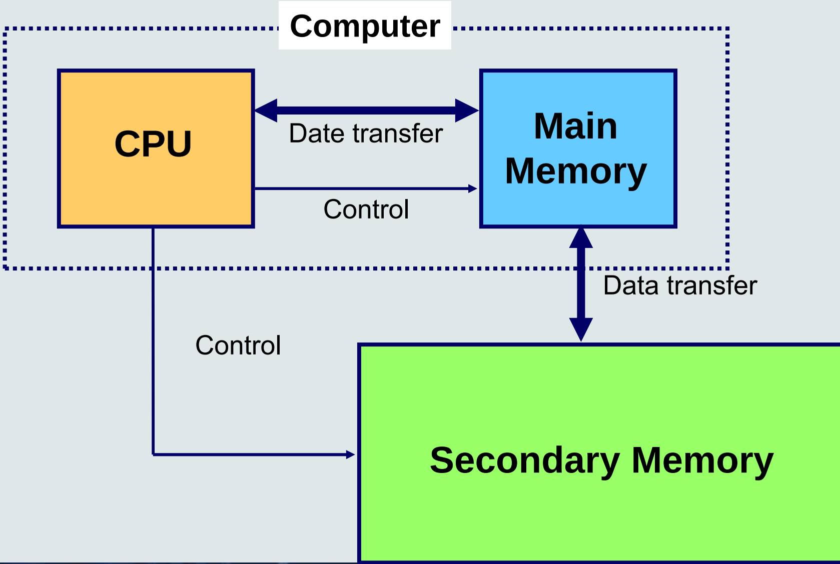
Introduction

- The main memory, also called primary memory, is used to store the program and data which are currently manipulated by the CPU.

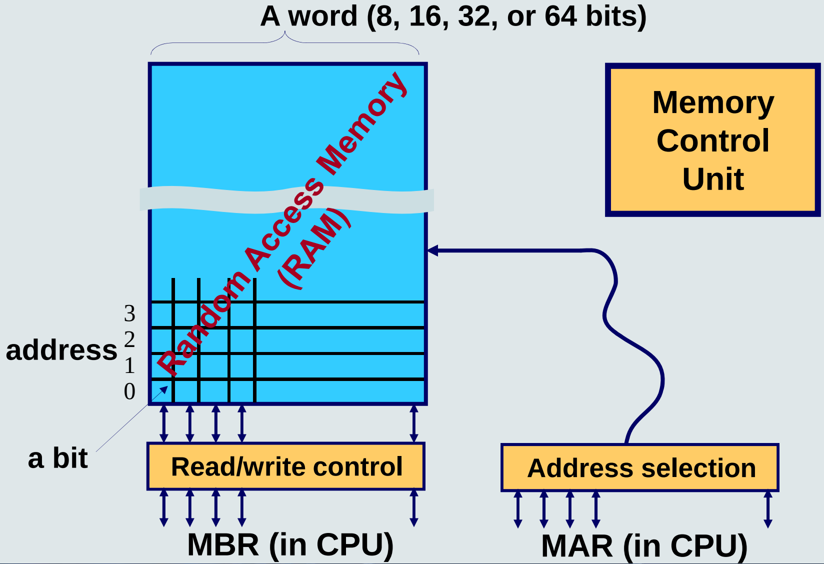
- The secondary memory provides the long-term storage of large amounts of data and program.

- Before the data and program in the secondary memory can be manipulated by the CPU, they must first be loaded into the main memory.

Internal and External Memories



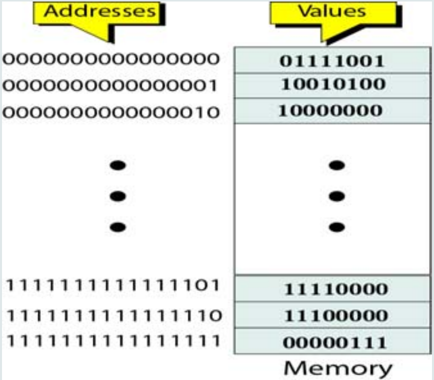
Main Memory Model



Main Memory (MM)

- The MM can be viewed as a set of storage cells, each of which is used to store a word.
- Each cell is assigned a unique address and the addresses are numbered sequentially: 0, 1, 2,...

16-bit address
↓
64 K total size
↓
64 Kbyte memory



Addresses	Values
0000000000000000	01111001
0000000000000001	10010100
0000000000000010	10000000
⋮	⋮
1111111111111101	11110000
1111111111111110	11100000
1111111111111111	00000111

Memory

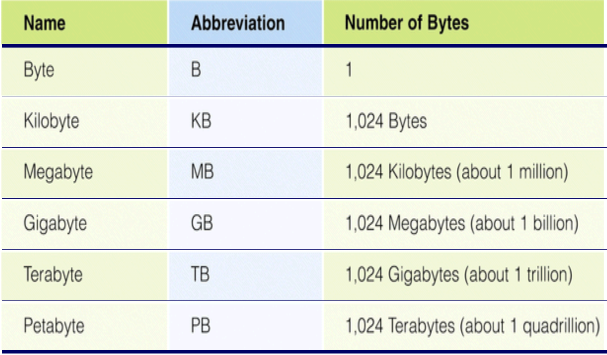
In this example(!)
1 word = 1 byte

This is a special case; very often a word consists of 2, 4, or 8 bytes.

Main Memory Capacity

- A byte is traditionally used to encode a text character, and it is the smallest addressable unit of memory.
- Modern computers are usually byte-addressable, even if a word has 2, 4, or 8 bytes.
- The number of address bits determines the maximal size of the memory.
  - Ex.16 bits—64K, 24 bits—16M, 32 bits—4G.
- There are a read/write controller and an address selection mechanism, which are part of the memory control unit.

Memory Capacity Units



Name	Abbreviation	Number of Bytes	No. of bytes	Approximation
Byte	B	1	1 byte	1 byte
Kilobyte	KB	1,024 Bytes	2^{10} bytes	10^{3} bytes
Megabyte	MB	1,024 Kilobytes (about 1 million)	2^{20} bytes	10^{6} bytes
Gigabyte	GB	1,024 Megabytes (about 1 billion)	2^{30} bytes	10^{9} bytes
Terabyte	TB	1,024 Gigabytes (about 1 trillion)	2^{40} bytes	10^{12} bytes
Petabyte	PB	1,024 Terabytes (about 1 quadrillion)	2^{50} Bytes	10^{15} bytes

- Exabyte (EB) = $2^{60} = 10^{18}$;
- Zettabyte (ZB) = $2^{70} = 10^{21}$;
- Yottabyte (YB) = $2^{80} = 10^{24}$.

Growing Data Volume

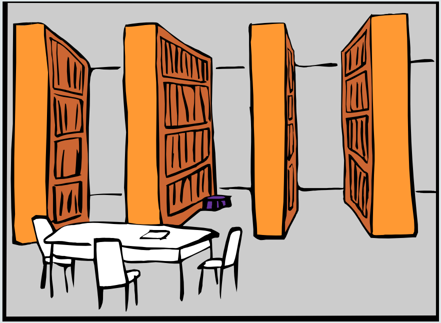
- In a life time, a programmer may generate code of the size:
  - 3 (lines/h) X 8 (hours/d) X 200 (days/y) X 40 (years)
    = 192 000 LOC ≈ 768 000 Bytes < 1 MB.
- In a life time, a writer will at most fill memory of the size:
  - A book page ≈ 2000 characters (2 KB).
  - A 1000-page book ≈ 2 MB.
  - 500 books ≈ 1 GB.
- A digital photo ≈ 20 MB (a iPhone photo ≈ 3 MB).
- A minute of professional video ≈ 0.5 GB.
- A HD movie ≈ 50 GB after compression.

Memory Characteristics

The most important characteristics of a memory:

- Speed — as fast as possible;
- Size — as large as possible;
- Cost — reasonable price.

They are determined by the technology used for implementation.



Your personal library

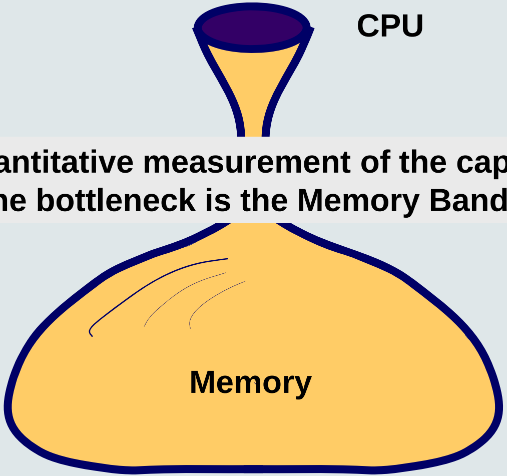
Memory Access

- Reading or writing the content of a memory cell is called an access.
- The time needed to finish one reading or writing operation is called the access time.
- The access time plus any additional time required before a second access can start is called the memory cycle time.
- The memory cycle time is equal to the access time in most technologies.
- Memory access is the bottleneck, because the memory cycle time is much longer than the machine cycle time.
    - Inside a computer, operations are performed in lock-step, each lasting for a clock period.
    - Ex. a memory access can take 10 clock cycles to complete.

Performance Metrics

- Clock frequency
  - Ex. 5GHz = a clock period is 0.2 nano-second long.
- Memory cycle time
  - Ex. A memory access can take 10 clock cycles = 2 nano-second.
- Machine cycle time = Instruction execution time
  - Ex. a computer can have a performance of 500 MIPS (Millions of Instructions Per Second).
  - Since different instructions need different time to execute, the average instruction execution time is often used.
- Very common, FLOPS (FLoating-point Operations Per Second) is used nowadays.
  - Ex. A 4-core PC can have a performance of 20 GigaFLOPS.

Memory Access Bottleneck



Quantitative measurement of the capacity of the bottleneck is the Memory Bandwidth

Memory Bandwidth

- Memory bandwidth denotes the amount of data that can be accessed from a memory per second:

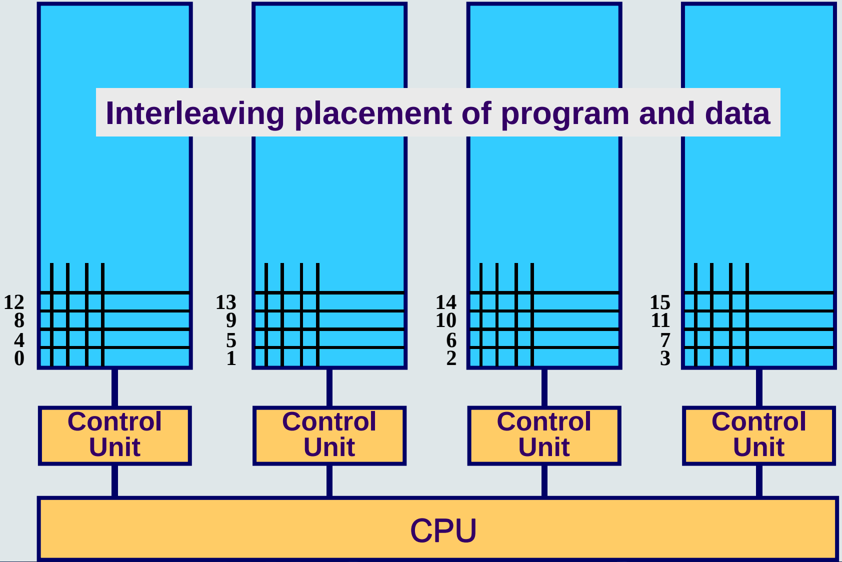
$$\text{M-Bandwidth} = \frac{1}{\text{memory cycle time}} \bullet \text{amount of data per access}$$

Ex. MCT = 10 nano-second and 4 bytes (a 32-bit word) per access:

M-Bandwidth = 400 megabytes per second.

- There are two basic techniques to increase the bandwidth of a given memory:
  - Reduce the memory cycle time
    - Expensive
    - Memory size limitation
  - Divide the memory into several banks, each of which has its own control unit (using parallelism).

Memory Banks



Semiconductor Memories

- The most widely used technology to implement main memories is semiconductor memories.
  - Ex. CMOS technology: high density, low power consumption, and relatively cheap.
- They use flip-flops made of transistors as the basic unit to represent 0 and 1.
- They are random access memory (RAM).
  - Static RAM (SRAM): fast but expensive, no refreshing needed.
  - Dynamic RAM (DRAM): small and cheap, but need refreshing.
- The information stored in a semiconductor memory will be lost when electrical power is removed (volatile)!

Read Only Memory (ROM)

- It is a permanent memory which can only be read but not written.
  - Hold instructions that start the computer when it is first switched on (BIOS).
  - A part of the main memory.
- Since ROM is usually very fast, they can also be used when fast reading of program/data is required.
  - Store library subroutines (e.g., for division operation).
  - Store dictionaries for spell checking.
- Programmable ROM (PROM) - A ROM which can be programmed once, by either the vendor or a customer.

Read-Mostly Memory

- Erasable PROM (EPROM) - To erase the contents of the memory, the chip is exposed to extra-violet radiation.
  - It can take up to 20 minutes to do this.
- Electrically Erasable PROM (EEPROM) - Electronic impulses (higher voltages) are used for write, but takes much longer time than the read operation.
- Flash Memory - An entire memory or part of it can be erased electrically, based on floating-gate transistor technology.

Memory Classification

- **Random-access memory (RAM)**:
  - The time taken to read or write data does not depend on where the data is stored inside the memory.
  - All main memories are of the RAM type.
  - Ex. semiconductor memories.

- **Sequential-access memory (SAM)**:
  - If a data item is to be read, all data items before it must also be read.
  - Ex. magnetic tape.

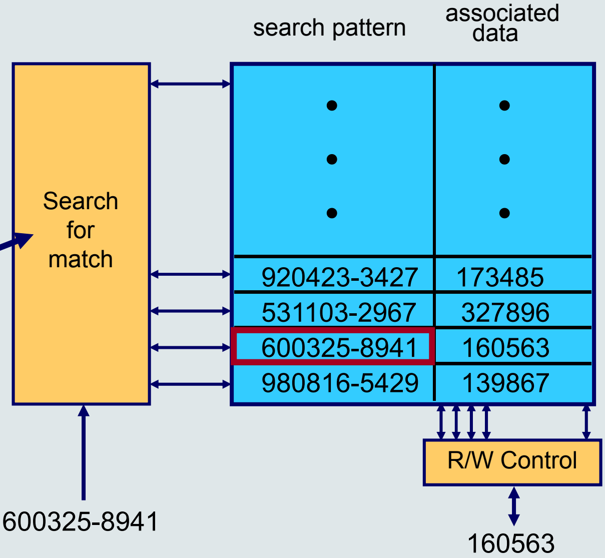
Memory Classification (Cont'd)

- **Direct-access storage devices**:
  - Data can be obtained without having to read through masses of other data.
  - Ex. hard disk.
- **Associative memory**:
  - A RAM where a word is retrieved based on a portion of its content rather than its address.
  - Comparison of the given bits of a word with a specified pattern is made for each access, and this is performed for all words simultaneously.

Associative Memory Example

Very complex:

Parallel comparisons are needed for many worlds at the same time!



search pattern	associated data
•	•
•	•
•	•
920423-3427	173485
531103-2967	327896
600325-8941	160563
980816-5429	139867

Lecture 2: Memory System

- Main memory

- Secondary memory

- Memory hierarchy

Magnetic Tape

- Magnetic tape is made up from a layer of plastic that is coated with iron oxide.
  - The oxide can be magnetized in two different directions to represent 0/1.
- Its operation uses a similar principle as in the case of a tape recorder.
- Main features:
  - Sequential access (access time about 1-5 s).
  - Relatively high capacity of storage (ca. 80 MB per tape).
  - Inexpensive.
- It is often used for backup or archive purpose.

Diskette

- Data are recorded on the surface of a floppy disk made of polyester coated with magnetic material.
- A special diskette drive must be used to access data stored in the floppy disk.
  - It works much like a record turntable of a gramophone.
- Main features:
  - Direct-access memory
  - Cheap
  - Portable, convenient to use

- Main standards:
  - 5 1/4-inch. Capacity ≈ 360 KB/disk
  - 3 1/2-inch. Capacity ≈ 1.44 MB/disk (about 700 pages of A4 text)

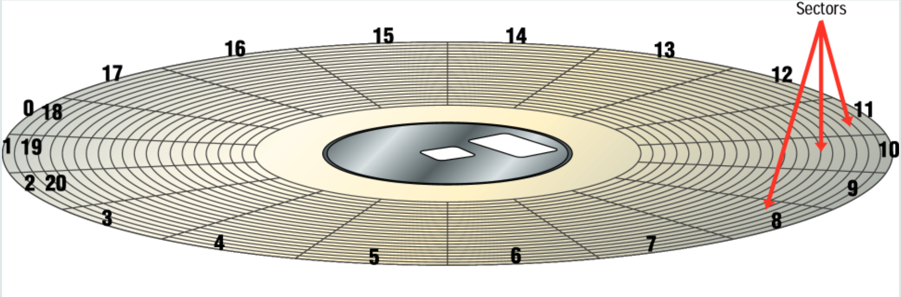
Disk Performance

- Access time:
  - Seek time — the time required to spin the disk to a constant rotation speed and to position the read/write head at the right track.
  - Rotational delay — the time required for the read/write head to position at the beginning of the sectors where data are stored.
- Read/write time — the time required to read/write a basic unit of data. Often, the data transfer rate is given instead:

$$\text{DTR} = \frac{1}{\text{read/write time}}$$

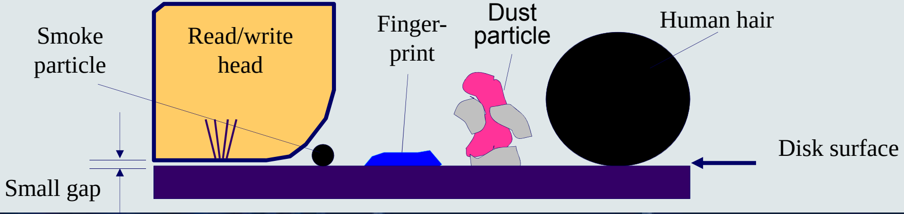
Accessing Secondary Memories

- A secondary memory is usually divided into large blocks (of the size of a few kilobytes, for example).
- Each block has a unique address and can be individually addressed.
- Data are moved between the secondary memory and the main memory one block at a time (not one word!).



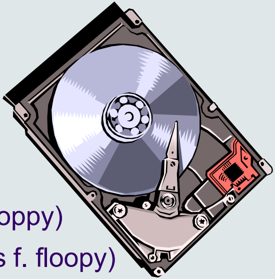
Hard Disk

- Data are recorded on the surface of a hard disk made of metal coated with magnetic material.
- A hard disk spins constantly to reduce seek time.
- The disk spins also at a very high speed (up to 10,000 rpm) to reduce the rotational delay and read/write time.
- The disks and the drive are usually built together and encased in a air tight container.
  - It protects the disks from pollutants, such as smoke particle and dust.



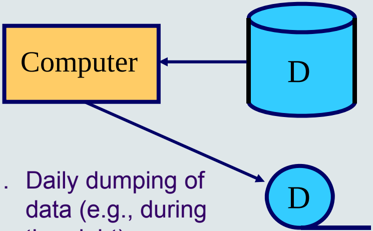
Hard Disk (Cont'd)

- Several disks are usually stacked on a common drive shaft with each disk having its own read/write head.
- Main features:
  - Direct access (not random access!).
  - Fast access:
    - seek time ≈ 8 ms (vs. 100 ms for floppy)
    - rotational delay ≈ 3 ms (vs. 100 ms for floppy)
    - data transfer rate ≈ 1 Gbits/s (0,5 Mbits/s f. floopy)
  - Huge storage capacity (ca. 500 GB for a compact unit)

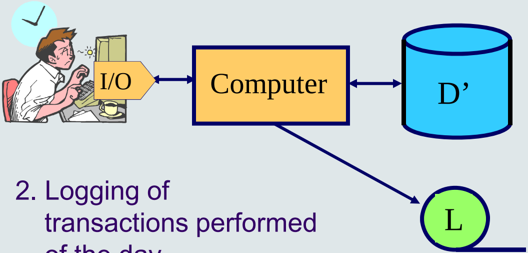


- The huge amount of data stored in hard disks must be backed up regularly.

Backup Procedure

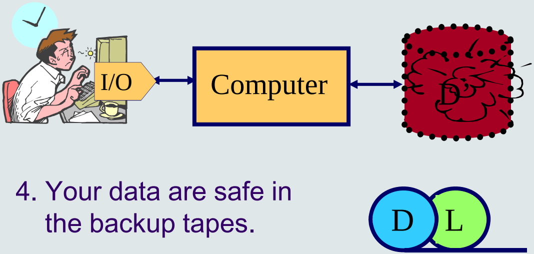


1. Daily dumping of data (e.g., during the night).



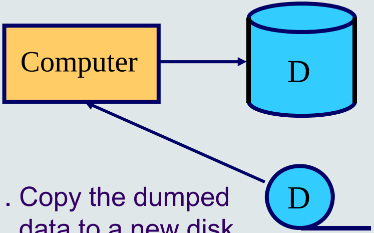
2. Logging of transactions performed of the day.

3. Disk crash happens.

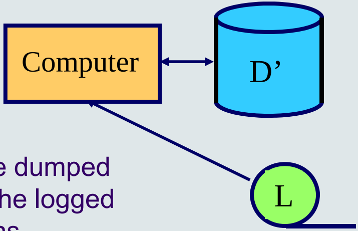


4. Your data are safe in the backup tapes.

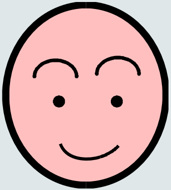
Recovery from Backup

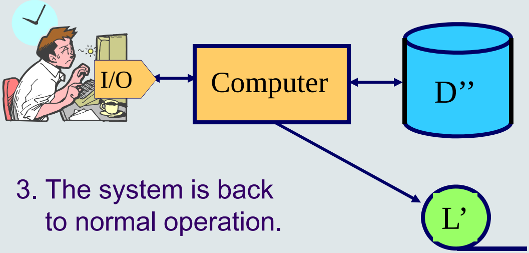


1. Copy the dumped data to a new disk.



2. Update the dumped data with the logged transactions.

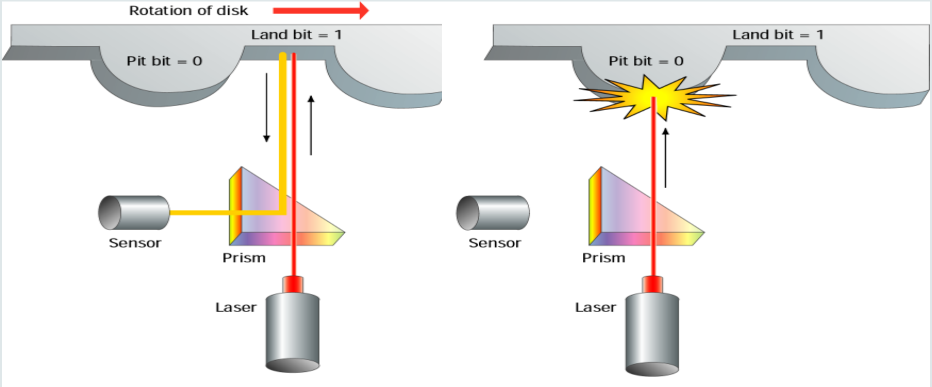




3. The system is back to normal operation.

Optical Memory

- An optical disk's surface is imprinted with microscopic holes which record digital information.
- When a low-powered laser beam shines on the surface, the intensity of the reflected light changes, representing 0 or 1.



Optical Memory Devices

- **CD-ROM (Compact Disk ROM)**:
  - Large capacity: > 650 MB/disk (ca. 460 diskettes).
  - Inexpensive replication, cheap for mass production.
  - Removable.
  - Long access time (could be half a second).
  - Read-only.

- **CD-Recordable (CD-R)**:
  - Write-once read-many (WORM).
  - A laser beam of modest intensity is used to imprint holes.
  - Good for archival storage by providing a permanent record of large volumes of data.

Optical Memory Devices (Cont'd)

- **CD-Rewritable (CD-RW)**:
  - Based on different reflectivities in two different phase states.
  - Erasable (500,000 to 1,000,000 erase cycles possible).
  - Getting cheaper.
- **Digital Versatile Disk (DVD)**:
  - Huge capacity: 17 GB per disk (ca. 12,000 diskettes).
  - Full length movie on a single disk (with MPEG compression).
  - Read-only, but DVD Recordable and DVD Rewritable are coming.
- **High Definition Optical Disks (HD-DVD)**:
  - Much higher capacity than DVD: 50 GB per disk.
  - Shorter wavelength laser (blue-violet range, called Blue-ray DVD).

USB Flash Drive

- A small, portable flash memory card.
- It plugs into a computer's USB port.
  - USB (Universal Serial Bus): an industry standard that defines the cables, connectors and communications protocols for computers and electronic devices.
  - Off-line memory, disconnected or removable memory.
- It functions as a portable hard drive.
- Large capacity is possible: e.g., 256 GB.
- Convenient to store and transfer data.

Lecture 2: Memory System

- Main memory
- Secondary memory
- Memory hierarchy

Motivation

- What do we need?
  - A memory to store very large programs/data and to work at a speed comparable to that of the CPU.

- The reality is:
  - The larger a memory, the slower it will be;
  - The faster a memory, the greater the cost per bit.

- A solution:
  - To build a composite memory system which combines a small and fast memory with a large and slow memory, and behaves, <u>most of the time</u>, like a large and fast memory.
  - This two-level principle can be extended to a hierarchy of many levels.

Memory Hierarchy

CPU

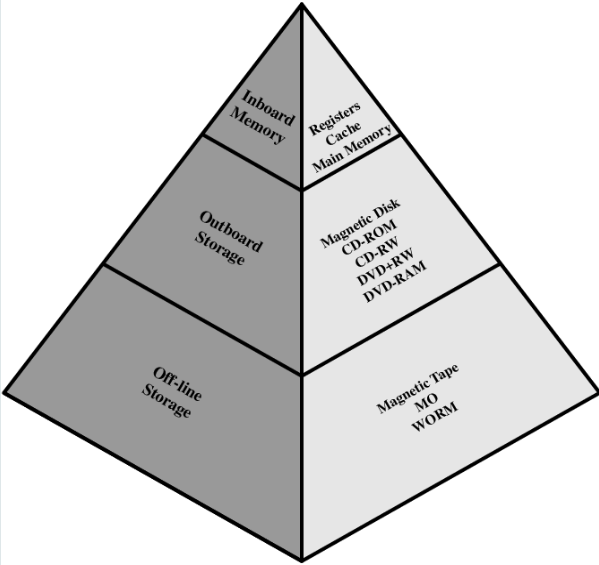
Registers

Cache

Main Memory

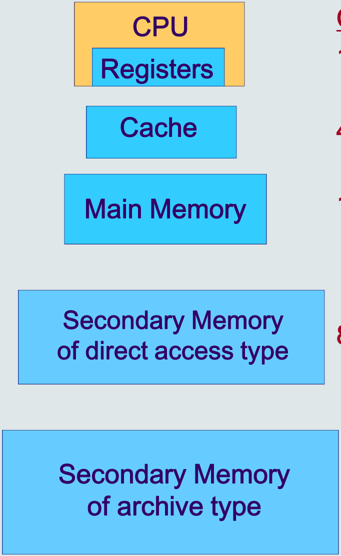
Secondary Memory of direct access type

Secondary Memory of archive type



Technical data are always out of date!

Memory Hierarchy



Access time example:		Capacity example:
0.25 ns	CPU Registers	1 KB
1 ns	Cache	4 MB
8 ns	Main Memory	16 GB
1 ms (4KB)	Secondary Memory of direct access type	8 TB
100 ms (100KB)	Secondary Memory of archive type	(100 MB/tape)

As one goes down the hierarchy, the following occurs:

- Increasing capacity.
- Increasing access time.
- Decreasing cost/bit.
- Decreasing frequency of access by the CPU.

Locality of Reference

- Programs access a small portion of their address space at any short period of time.
- **Temporal locality:** If an item is accessed, it will tend to be accessed again soon.
- **Spatial locality:** If an item is accessed, items whose addresses are close by will tend to be accessed soon.
- This access pattern is an intrinsic features of the von Neumann architecture:
  - Sequential instruction storage and execution.
  - Loops and iterations (e.g., subroutine calls).
  - Sequential data storage (e.g., array).

Summary

- A memory system has to store very large programs and a huge amount of data and still provide fast access.
- No single type of memory can provide all such need for a computer system.
- Therefore, several different storage mechanisms are organized in a layer hierarchy.
  - The main memory stores the program and data which are <u>currently manipulated</u>.
  - The secondary memory provides the long-term storage of large amounts of data and program.
- The layer structure works very well due to the <u>locality of reference</u> principle.