



Lecture 6

- Memory Hierarchy
- Registers
- Main memory

1. Memory Hierarchy:

- ❑ Most computers are built using the Von Neumann model, which is centered on memory. The programs that perform the processing are stored in memory.

The memory is logically structured as a linear array of locations, with addresses from 0 to the maximum memory size the processor can address

- ❑ The design constraints on a computer's memory can be summed up by three questions:

How much? capacity

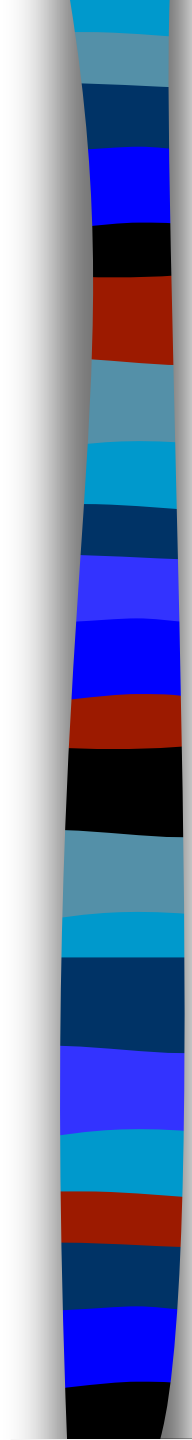
How fast? greatest performance in access time

How expensive? cost

the following relationships hold:

- ■ Faster access time, greater cost per bit;
- ■ Greater capacity, smaller cost per bit;
- ■ Greater capacity, slower access time

The designer would like to use memory technologies that provide for large- capacity memory, both because the capacity is needed and because the cost per bit is low. However, to meet performance requirements, the designer needs to use expensive, relatively lower- 2 capacity memories with short access times.

- 
- ❑ One of the most important considerations in understanding the performance capabilities of a modern processor are the memory hierarchy.
 - ❑ Unfortunately not all memory is created equal, and some types are far less efficient and thus cheaper than others.
 - ❑ Today's computer systems use a combination of memory types to provide the best performance at the best cost. This approach is called hierarchical memory.
 - ❑ Although seemingly simple in concept, computer memory exhibits perhaps the widest range of type, technology, organization, performance, and cost of any feature of a computer system.
 - ❑ As a consequence, the typical computer system is equipped with a hierarchy of memory subsystems, some internal to the system (directly accessible by the processor) and some external (accessible by the processor via an I/O module)

- The base types that normally constitute the hierarchical memory system include internal memory (registers, cache, main memory), and external or secondary memory

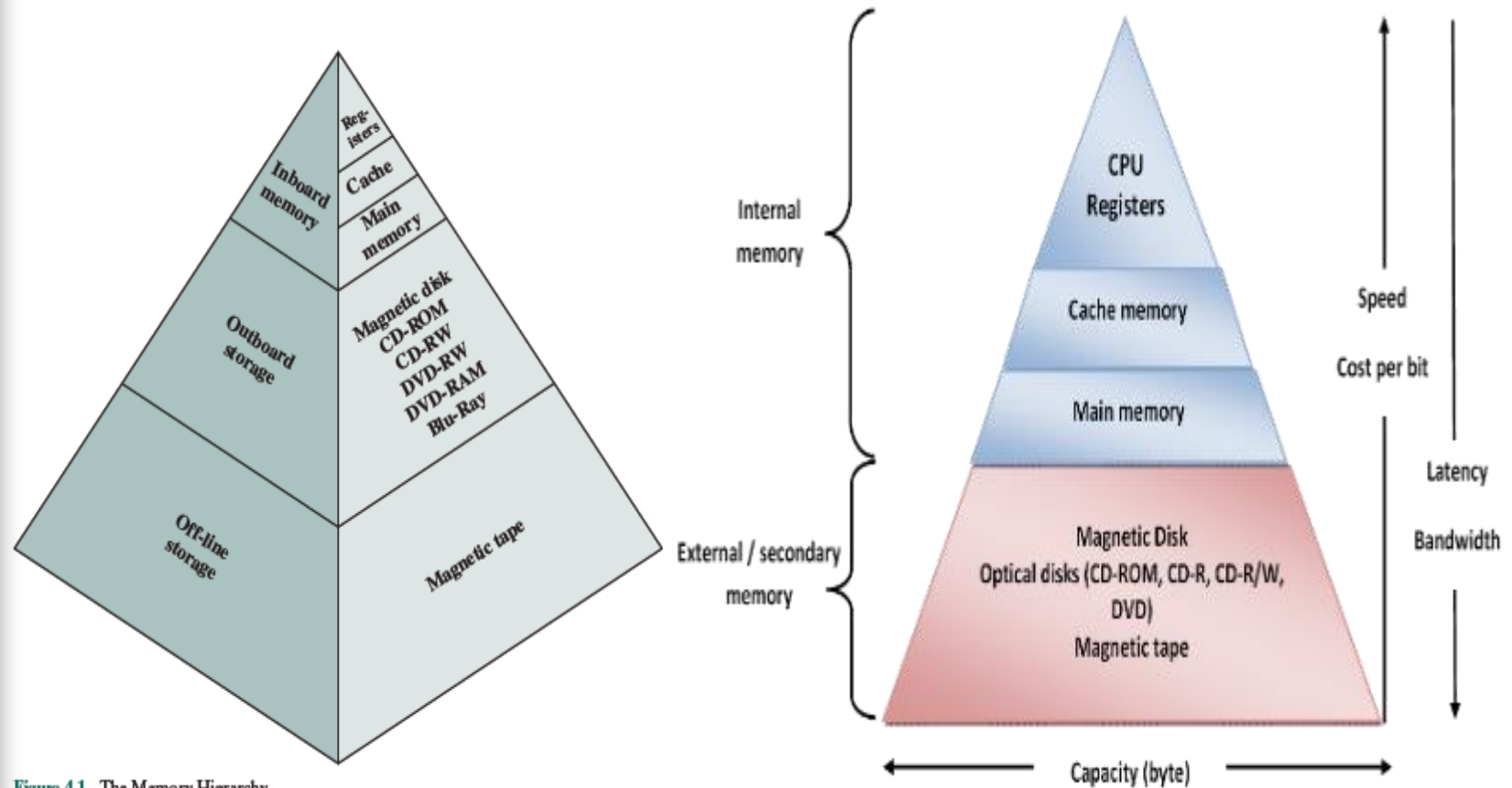
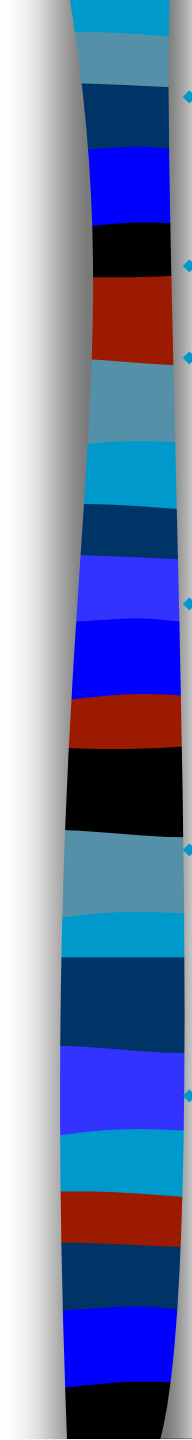


Figure 4.1 The Memory Hierarchy

- The memory hierarchy can be characterized by a number of parameters. Among these parameters are the (access type, capacity, cycle time, latency, bandwidth, and cost).

- 
- ❖ The term **access** refers to the action that physically takes place during a read or writes operation.
 - ❖ The **capacity** of a memory level is usually measured in bytes.
 - ❖ The **cycle time** is defined as the time elapsed from the start of a read operation to the start of a subsequent read.
 - ❖ The **latency** is defined as the time interval between the request for information and the access to the first bit of that information.
 - ❖ The **bandwidth** provides a measure of the number of bits per second that can be accessed.
 - ❖ The **cost** of a memory level is usually specified as costs per megabytes.

2. Registers:

A register is a group of flip-flops capable of storing one bit of information.

An n-bit register has a group of n flip-flops and is capable of storing any binary information of n bits.

- ❑ Registers are with different sizes and functions.
- ❑ Two basic types of registers are commonly used:

1- parallel registers:

consists of a set of 1-bit memories that can be read or written simultaneously. It is used to store data.

2- shift registers.

flip- flops: accepts and/or transfers information serially. can be used to interface to serial I/O devices. In addition, they can be used within the ALU to perform logical shift and rotate functions.

- ❑ The simplest form of sequential circuit is the flip- flop. There are a variety of flip- flops, all of which share two properties:
- ❑ ■■ The flip- flop is a bistable device. It exists in one of two states and, in the absence of input, remains in that state. Thus, the flip- flop can function as a 1-bit memory.

3. Main memory.

The main memory is the central storage unit in a computer system. It is a relative large and fast memory used to store programs and data during the computer operation. There are only two basic types of main memory:

1. RAM (random access memory):

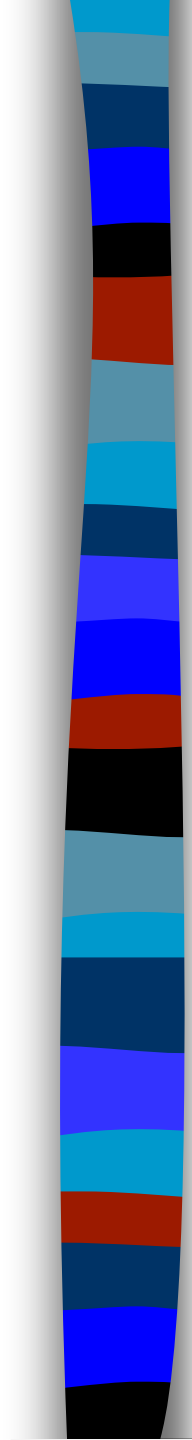
- ❖ A more appropriate name is read–write memory. RAM is the memory to which computer specifications refer,
- ❖ RAM is used to store programs and data that the computer needs when executing programs; but RAM is volatile, and loses this information once the power is turned off.

There are two general types of chips used to build the bulk of RAM memory in computers:

(a) SRAM (static random access memory).

(b) DRAM (dynamic random access memory).

- DRAM is constructed of tiny capacitors that leak electricity. DRAM requires a recharge every few milliseconds to maintain its data.
- SRAM technology, in contrast, holds its contents as long as power is available.



<u>Properties</u>	<u>DRAM</u>	<u>SRAM</u>
Power	requires a periodic power to maintain its data	requires a continuous power to maintain its data
Speed	slow	fast
Cost	cheap	expensive
Size	large	small
Usage	main memory	cache memory

2. ROM (read-only memory):

Most computers contain a small amount of ROM (read-only memory) that stores critical information necessary to operate the system, such as the program necessary to boot the computer (BIOS). ROM is not volatile and always retains its data. This type of memory is also used in embedded systems or any systems where the programming does not need to change.

Many appliances use ROM chips to maintain information when the power is shut off.

There are five basic different types of ROM.

1. ROM.

2. PROM, PROM (programmable read-only memory) is a variation on ROM. PROMs can be programmed by the user with the appropriate equipment. Whereas ROMs are hardwired, PROMs have fuses that can be blown to program the chip. Once programmed, the data and instructions in PROM cannot be changed.

3. EPROM, EPROM (erasable PROM) is programmable with the added advantage of being reprogrammable (erasing an EPROM requires a special tool that emits ultraviolet light). To reprogram an EPROM, the entire chip must first be erased.

4. EEPROM, EEPROM (electrically erasable PROM) removes many of the disadvantages of EPROM: no special tools are required for erasure (this is performed by applying an electric field) and you can erase only portions of the chip, one byte at a time.

5. Flash memory, is essentially EEPROM with the added benefit that data can be written or erased in blocks, removing the one-byte-at-a-time limitation. This makes flash memory faster than EEPROM.

Characteristics of Memory Systems

- ❑ The complex subject of computer memory is made more manageable if we classify memory systems according to their key characteristics.
- ❑ The most important of these are listed in Table 4.1.
- ❑ location: refers to whether memory is internal or external to the computer.
- ❑ Internal memory is often equated with main memory, but there are other forms of internal memory. The processor requires its own local memory, in the form of registers (e.g., see Figure 2.3). Further, as we will see, the control unit portion of the processor may also require its own internal memory. Cache is another form of internal memory.
- ❑ External memory consists of peripheral storage devices, such as disk and tape, that are accessible to the processor via I/O controllers.
- ❑ An obvious characteristic of memory is its capacity. For internal memory, this is typically expressed in terms of bytes (1 byte = 8 bits) or words. Common word lengths are 8, 16, and 32 bits.
- ❑ External memory capacity is typically expressed in terms of bytes.

Table 4.1 Key Characteristics of Computer Memory Systems

Location

Internal (e.g., processor registers, cache, main memory)

External (e.g., optical disks, magnetic disks, tapes)

Capacity

Number of words

Number of bytes

Unit of Transfer

Word

Block

Access Method

Sequential

Direct

Random

Associative

Performance

Access time

Cycle time

Transfer rate

Physical Type

Semiconductor

Magnetic

Optical

Magneto-optical

Physical Characteristics

Volatile/nonvolatile

Erasable/nonerasable

Organization

Memory modules

unit of transfer. For internal memory.

is equal to the number of electrical lines into and out of the memory module. This may be equal to the word length, but is often larger, such as 64, 128, or 256 bits.

The three related concepts for internal memory.

■ ■ **Word:** The “natural” unit of organization of memory. The size of a word is typically equal to the number of bits used to represent an integer and to the instruction length.

■ ■ **Addressable units:** In some systems, the addressable unit is the word. However, many systems allow addressing at the byte level. In any case, the relationship between the length in bits A of an address and the number N of addressable units is $2^A = N$.

■ ■ **Unit of transfer:** For main memory, this is the number of bits read out of or written into memory at a time. The unit of transfer need not equal a word or an addressable unit. For external memory, data are often transferred in much larger units than a word, and these are referred to as blocks. Another distinction among memory types is the method of accessing units of data. These include the following:

■ ■ **Sequential access:** Memory is organized into units of data, called records. Access must be made in a specific linear sequence.



■ ■ Direct access: As with sequential access, direct access involves a shared read– write mechanism.

■ ■ Random access: Each addressable location in memory has a unique, physically wired– in addressing mechanism. The time to access a given location is independent of the sequence of prior accesses and is constant. Thus, any location can be selected at random and directly addressed and accessed.

□ Main memory and some cache systems are random access.

□ ■ ■ Associative: This is a random access type of memory that enables one to make a comparison of desired bit locations within a word for a specified match, and to do this for all words simultaneously.