Computer Organization & Architecture

3-1 Basic Concepts of Memory Systems

Wang Guohua

School of Software Engineering

Contents of this lecture

- Internal Memory and External Memory
- Memory Locations and Addresses
  - Reference 2.1 in Textbook
- Byte Ordering
  - Reference 2.1 in Textbook
- Main Memory Operations
  - Reference 8.1 in Textbook
- Characteristics of Memory Systems
  - Reference 3.2 in “Computer Organisation and Architecture”, A.P.Godse and D.A.Godse, Technical Publications

Internal Memory and External Memory

- The typical computer system is equipped with a hierarchy of memory subsystems.
- Internal Memory (Primary Storage)
  - Internal to the system, directly accessible by the processor.
  - Example
    - Main Memory, Cache, Processor Registers
- External Memory (Secondary Storage)
  - External to the system, accessible by the processor via an I/O module.

Memory Locations and Addresses (1)

- Main Memory Organization
  - The memory consists of many millions of storage cells, each of which can store a bit of information having the value 0 or 1.
  - We can imagine Main Memory to be organized as a matrix of bits.
  - The memory is organized so that a group of fixed size of bits can be stored or retrieved in a single, basic memory operation.
  - Word: Each group of fixed size of bits is referred to as a word.
  - Word length: The number of bits in each word is referred to as word length. It typically ranges from 16 to 64 bits.

Memory Locations and Addresses (2)

- Main Memory Organization (ctd.)
  - The memory of a computer can be schematically represented as a collection of words.

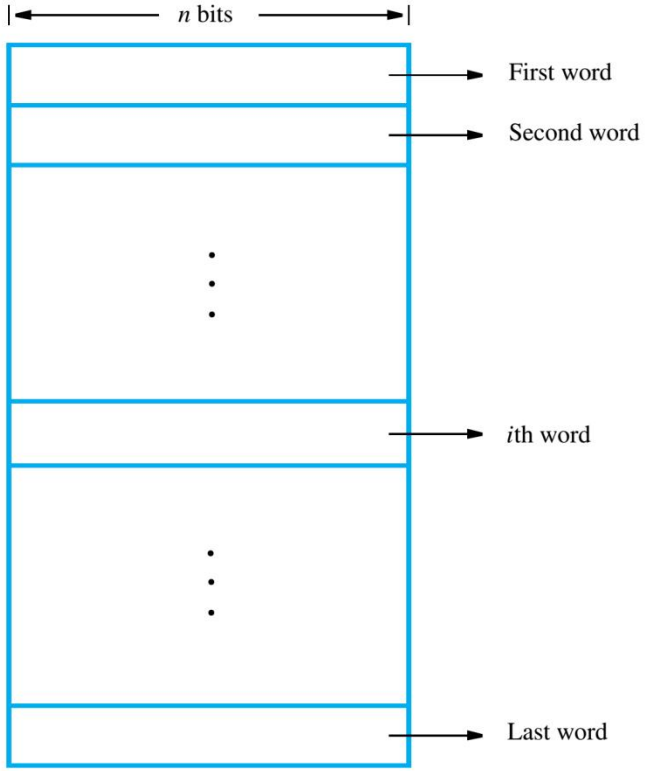


Figure 2.1 Memory words.

Memory Locations and Addresses (3)

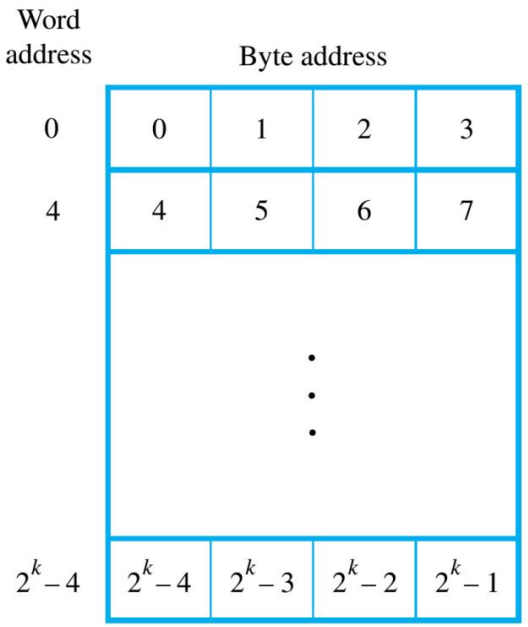
- Address and Address Space
  - Addresses are numbers that identify memory words.
  - It is customary to use numbers from 0 through $2^k - 1$, for some suitable value of k, as the addresses of successive locations in the memory.
  - The 2^k addresses constitute the address space of the computer, and the memory can have up to 2^k addressable locations.
  - Example
    - A 24-bit address generates an address space of 2^{24} or 16M locations.
    - A 32-bit address creates an address space of 2^{32} or 4G locations.

Memory Locations and Addresses (4)

- Byte-addressable Memory
  - It is impractical to assign distinct addresses to individual bit locations in the memory.
  - The most practical assignment is to have successive addresses refer to successive byte locations in the memory.

Byte Ordering (1)

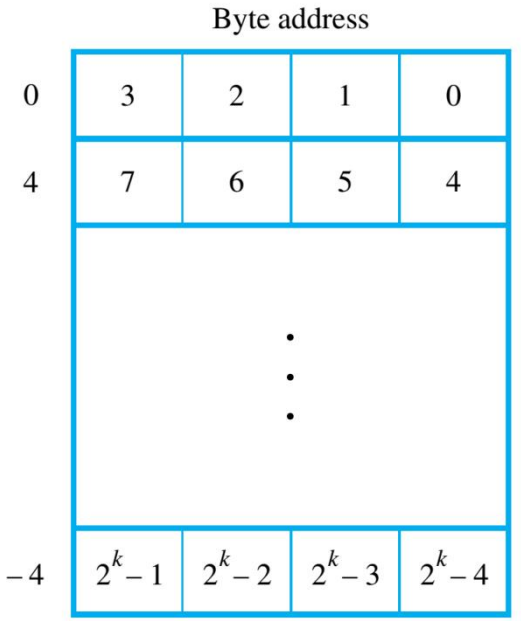
- Big-endian Assignment (Left-to-Right)
  - In this assignment, the lower byte addresses are used for the more significant bytes (the leftmost bytes) of the word.
  - Example: Figure 2.3 (a) Big-endian assignment for 32-bit word-length byte-addressable memory.



(a) Big-endian assignment

Byte Ordering (2)

- Little-endian Assignment (Right-to-Left)
  - In this assignment, the lower byte addresses are used for the less significant bytes (the rightmost bytes) of the word.
  - Example: Figure 2.3 (b) Little-endian assignment for 32-bit word-length byte-addressable memory.



(b) Little-endian assignment

Main Memory Operations (1)

- Connection of the Memory to the Processor

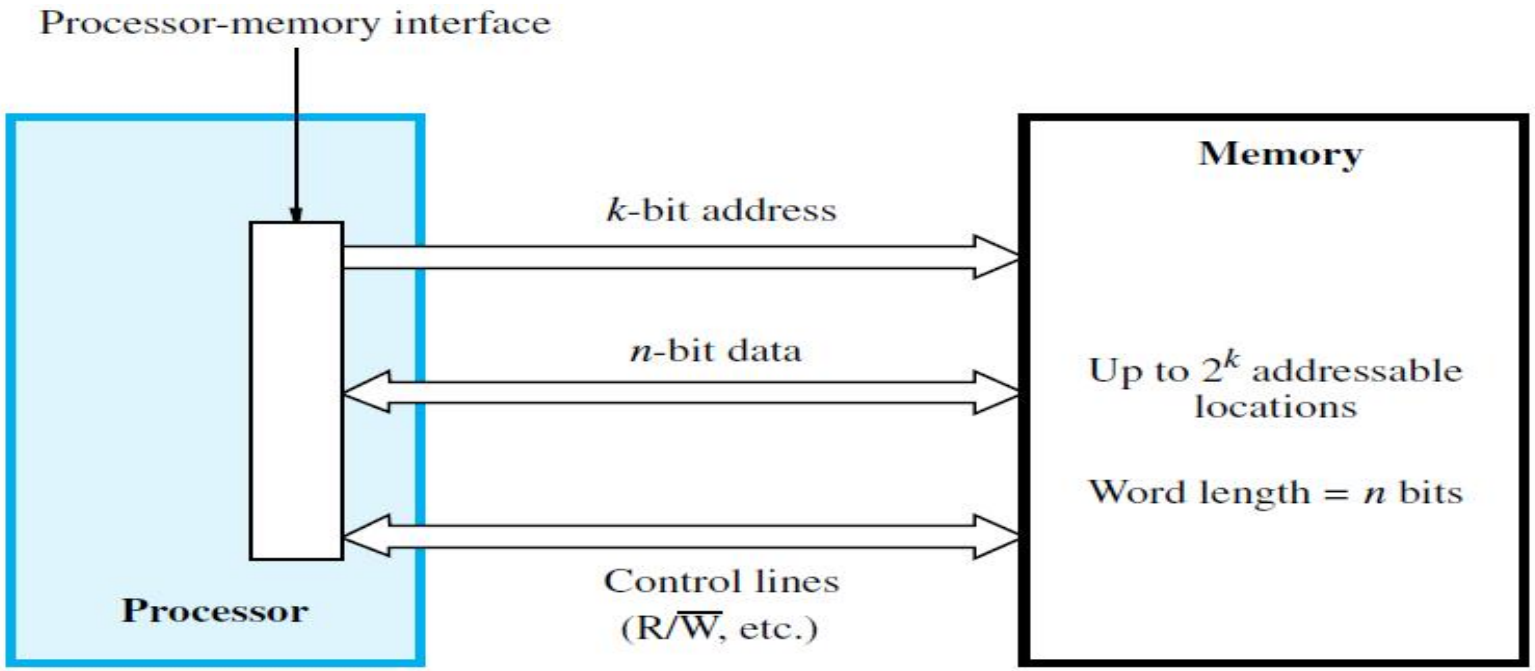
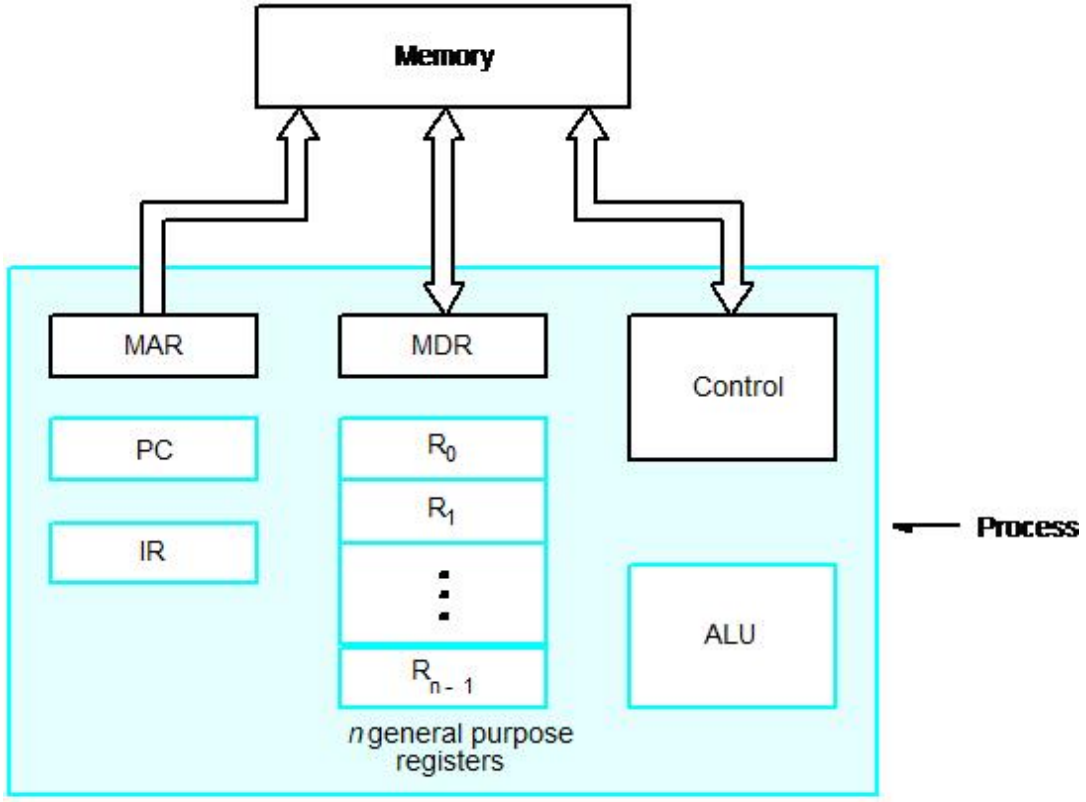


Figure 8.1 Connection of the memory to the processor.

Main Memory Operations (2)

- Processor-Memory Interface



Main Memory Operations (3)

- Read (Load) Operation
  - Transfers a copy of the contents of a specific memory location to the processor.
  - Processor: Load the address of the required memory location into the MAR register and set the R/W line to 1.
  - Memory: Place the data from the addressed location onto the data lines and confirm this action by asserting the MFC signal.
  - Processor: Upon receipt of the MFC signal, the processor loads the data on the data lines into the MDR register.

Main Memory Operations (4)

- Write (Store) Operation
  - Transfer an item of information from the processor to a specific location, destroying the former contents of that location.
  - Processor: Load the address of the specific location into MAR and load the data into MDR register. It also set the R/W line to 0.
  - Memory: When the data have been written, it responses processor with MFC signal.

Characteristics of Memory Systems

- Physical Types
- Capacity
- Unit of Transfer
- Access Methods
- Performance
- Physical Characteristics

Physical Types

- Semiconductor
  - Example
    - Main Memory
- Magnetic Surface
  - Example
    - Magnetic disk and Magnetic tape
- Optical
  - Example
    - CD, CD-R, CD-RW, DVD, Blu-Ray

Capacity

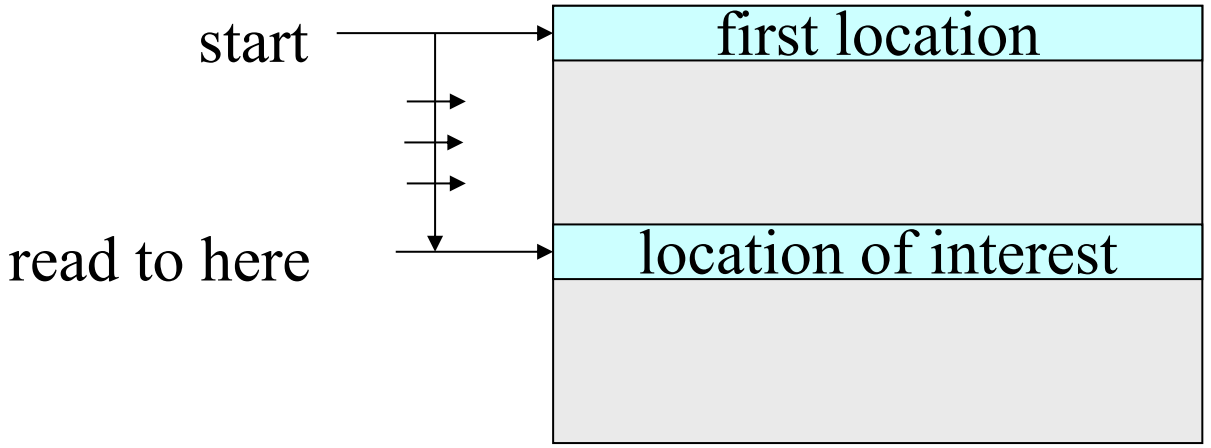
- For main memory, it is expressed using
  - Word size
    - Number of bits in natural unit of organization
    - The common word size is 8-bits, 16 bits, 32 bits and 64 bits.
  - Number of Word
    - Specifies the number of words available in the particular memory device.
- Example
  - If memory capacity is 4K×8, then its word size is 8, and the number of word is 4K.

Unit of Transfer

- Number of bits read from, or written into memory at a time.
- Internal Memory
  - Usually equal to the number of data lines into and out of the memory module.
  - It is often equal to the number of the word length, but it may not be.
- External Memory
  - Data are often transferred in much larger units than a word, and these are referred to as blocks.

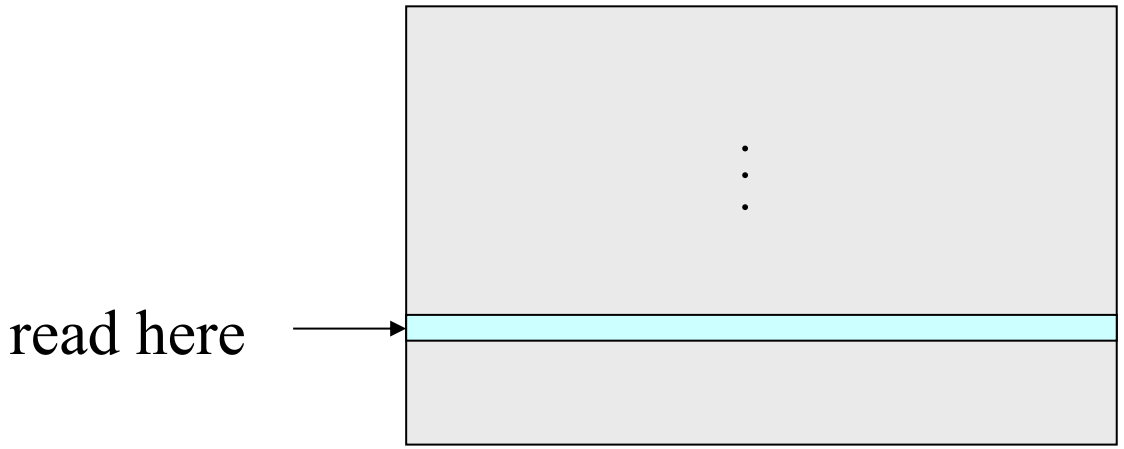
Access Methods (1)

- Sequential Access
  - Memory is organized into units of data, called records.
  - If current record is 1, then in order to read record N, it is necessary to read records 1 through N-1.
  - The time to access an arbitrary record is highly variable.
  - Example: Magnetic Tape



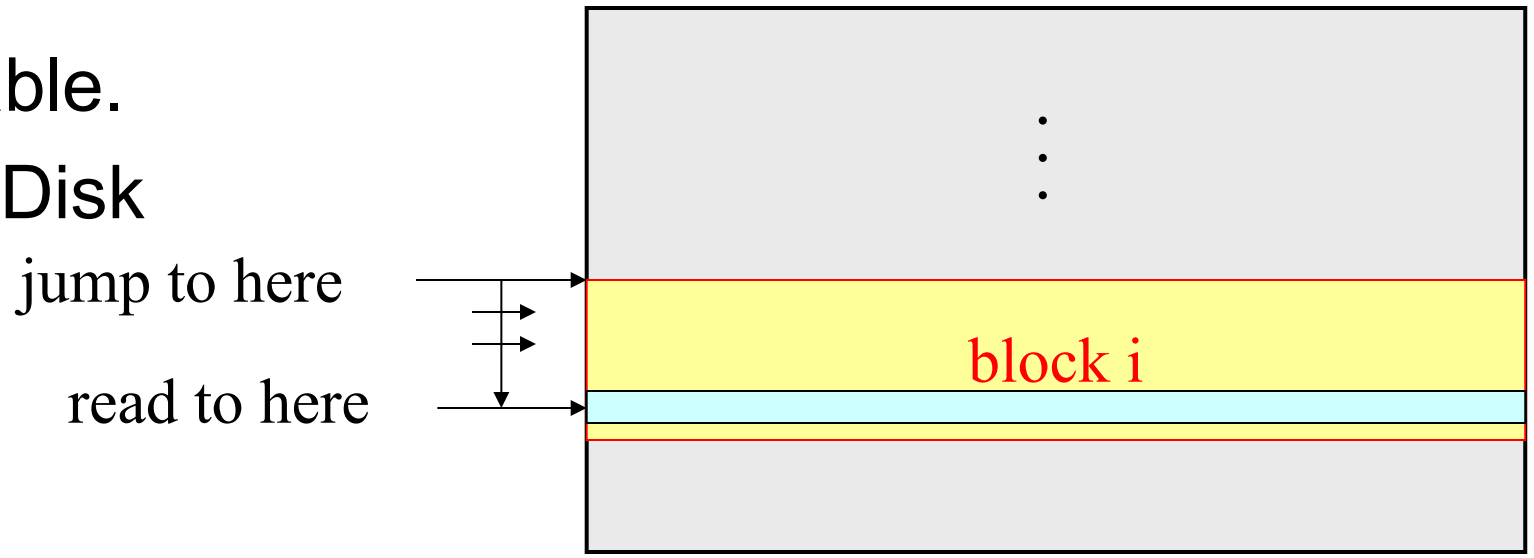
Access Methods (2)

- Random Access
  - Any memory location can be selected at random and directly addressed and accessed.
  - The time to access a given location is independent of the location's address and is constant.
  - Example: Semiconductor RAM Memory



Access Methods (3)

- Direct Access
  - Individual blocks or records have a unique address based on physical location.
  - Access is accomplished by direct access to reach a general vicinity plus sequential searching, counting or waiting to reach the final location.
  - Access time is variable.
  - Example: Magnetic Disk



Performance (1)

- Memory Access Time
  - Time between presenting the address and getting the valid data (memory or other storage)
  - Random-Access Memory
    - The time it takes to perform a read or write operation.
  - Non-Random-Access Memory
    - The time it takes to position the read-write mechanism at the desired location.

Performance (2)

- Memory Cycle Time
  - Used only for random access memory.
  - Time from a memory access to the next memory access.
  - It consists of the access time plus any additional time required before a second access can commence.
  - Memory Cycle Time = Memory Access Time + Recovery Time
  - Access time and memory cycle time are all measures of the speed of memory units.

Performance (3)

- Transfer Rate
  - It is defined as the rate at which data can be transferred into or out of a memory unit.

Physical Characteristics (1)

- Volatile/Nonvolatile
  - Volatile
    - Information decays naturally or is lost when electrical power is switched off.
  - Nonvolatile
    - Information once recorded remains without deterioration until deliberately changed. No electrical power is needed to retain information.
  - Example
    - Magnetic surface memories are nonvolatile.
    - Semiconductor memory may be either volatile or nonvolatile.

Physical Characteristics (2)

- Erasable/Non-erasable
  - Erasable
    - The contents of the memory can be altered.
  - Non-erasable
    - The contents of the memory can not be altered, except by destroying the storage unit.

Quiz (1)

1. The memory is organized so that a group of *n* bits can be stored or retrieved in a single, basic operation. *n* is called the ____.

   A. word B. word length C. address D. cell

2. The 32-bit value 0x30A79847 is stored to the location 0x1000. If the system is little endian, the value of the byte ____ is stored in address 0x1002.

   A. 0x30 B. 0xA7 C. 0x98 D. 0x47

Because the higher byte addresses are used for the more significant bytes (the rightmost bytes) of the word, so 0x30 is stored in 0x1003, 0xA7 is stored in 0x1002, and so on.

Quiz (2)

3. In a main memory, its word size is 16, the number of word is 8K, what is the capacity of this main memory?

   A. 16K×16　　B. 16K×8　　C. 8K×16　　D. 8K×8

4. For random access memory, the time it takes to perform a read or write operation is called ____.

   A. memory cycle time　　B. memory hit time

   C. memory recovery time　　D. memory access time

Quiz (3)

5. Memory in which any location can be reached in a short and fixed amount of time after specifying its address is called ____.

   A. direct access memory
   B. sequential access memory
   C. associative access memory
   D. random access memory

6. If you turn off the power to the computer, items stored on ____ device will be lost.

   A. RAM
   B. disk
   C. DVD
   D. CD-ROM

Quiz (4)

7. *True or False?* For internal memory, data are often transferred in much larger units than a word, and these are referred to as blocks.

8. *True or False?* Memory access time is longer than memory cycle time.

9. What is memory access time? What is memory cycle time? Explain the relationship between them.

   Solution: Memory access time is time between presenting the address and getting the valid data. Memory cycle time is time from a memory access to the next memory access. Their relationship is Memory Cycle Time = Memory Access Time + Recovery Time.