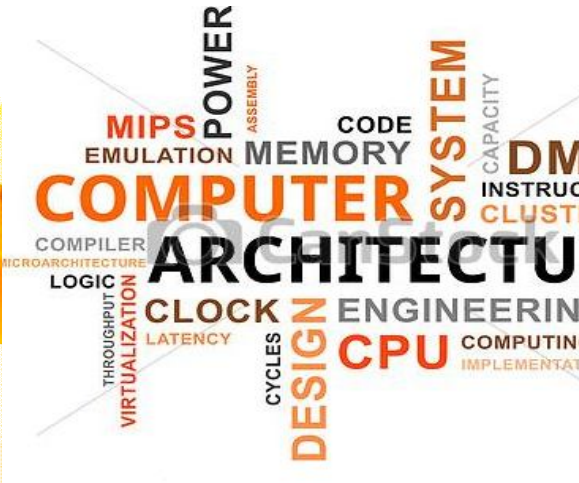
UNIT IV
MEMORY SYSTEM

Basic concepts of Semiconductor RAMs - ROMs – Speed, Size and Cost – Cache memories – Performance consideration – Virtual memory – Memory Management requirements – Secondary storage - Case Study: Memory Organization in Multiprocessors



Recap the previous Class

Introduction

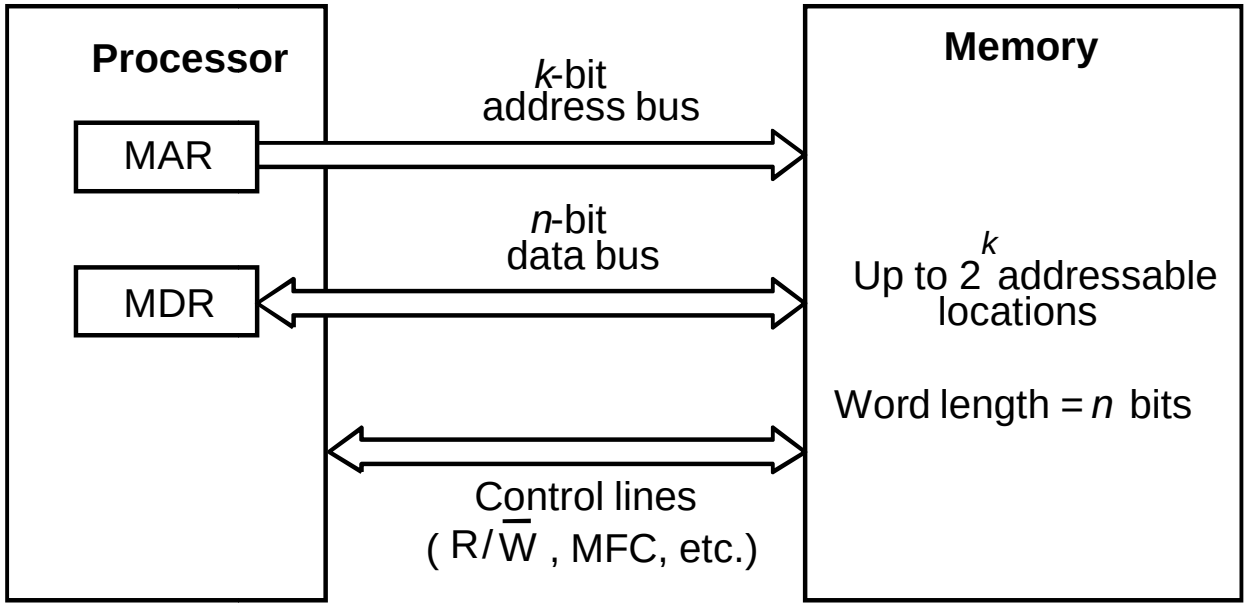
- Memory is one of the most important functional units of a computer.
  - Used to store both instructions and data.
  - Stores as bits (0's and 1's), usually organized in terms of bytes.
    - How are the data stored in memory accessed?
      - Every memory location has a unique address.
  - A memory is said to be byte addressable if every byte of data has a unique address.
  - Some memory systems are word addressable also (every addressed locations consists of multiple bytes, say, 32 bits or 4 bytes).

Connection between Processor and Memory

- Address bus provides the address of the memory location to be accessed.
- Data bus transfers the data read from memory, or data to be written into memory.
- Control bus provides various signals like READ, WRITE, etc.

An Example Memory Module

- *k address lines* :: The maximum number of memory locations that can be accessed is 2^k.
- *n data lines* :: The number of bits stored in every addressable location is n.
- The RD/WR' control line selects the memory for reading or writing (1: read, 0: write).



- Measures for the speed of a memory:
  - memory access time.
  - memory cycle time.
- An important design issue is to provide a computer system with as large and fast a memory as possible, within a given cost target.
- Several techniques to increase the effective size and speed of the memory:
  - Cache memory (to increase the effective speed).
  - Virtual memory (to increase the effective size).

Classification of Memory Systems

a)Volatile versus Non-volatile:

–A *volatile* memory system is one where the stored data is lost when the power is switched off.

- Examples: CMOS static memory, CMOS dynamic memory.
- Dynamic memory in addition requires periodic refreshing.

–A *non-volatile* memory system is one where the stored data is retained even when the power is switched off.

- Examples: Read-only memory, Magnetic disk, CDROM/DVD, Flash memory, Resistive memory.

b)Random-access versus Direct/Sequential access:

–A memory is said to be *random-access* when the read/write time is independent of the memory location being accessed.

•Examples: CMOS memory (RAM and ROM).

–A memory is said to be *sequential access* when the stored data can only be accessed sequentially in a particular order. Examples: Magnetic tape, Punched paper tape.

–A memory is said to be *direct or semi-random access* when part of the access is sequential and part is random. Example: Magnetic disk.

•We can directly go to a track acer which access will be sequential.

c)Read-only versus Random-access:

–*Read-only Memory* (ROM) is one where data once stored in permanent or semi-permanent.

- Data written (programmed) during manufacture or in the laboratory.
- Examples: ROM, PROM, EPROM, EEPROM.

–*Random Access Memory* (RAM) is one where data access time is the same independent of the location (address).

- Used in main / cache memory systems.
- Example: Static RAM (SRAM) data once written are retained as long as power is on.
- Example: Dynamic RAM (DRAM) requires periodic refreshing even when power is on (data stored as charge on tiny capacitors).

Access Time, Latency and Bandwidth

- Terminologies used to measure speed of the memory system.

  a) Memory Access Time: Time between initiation of an operation (Read or Write) and completion of that operation.

  b) Latency: Initial delay from the initiation of an operation to the time the first data is available.

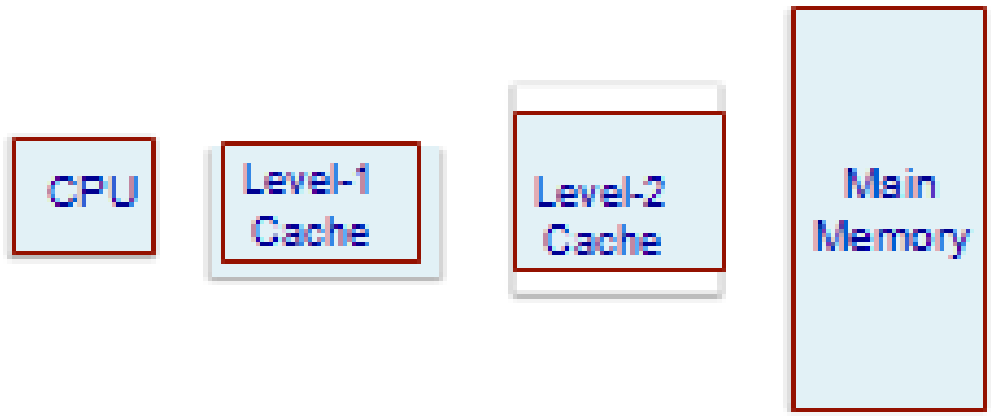
  c) Bandwidth: Maximum speed of data transfer in bytes per second.

- In modern memory organizations, every read request reads a block of words into some high-speed registers (LATENCY), from where data are supplied to the processor one by one (ACCESS TIME).

What is Cache Memory?

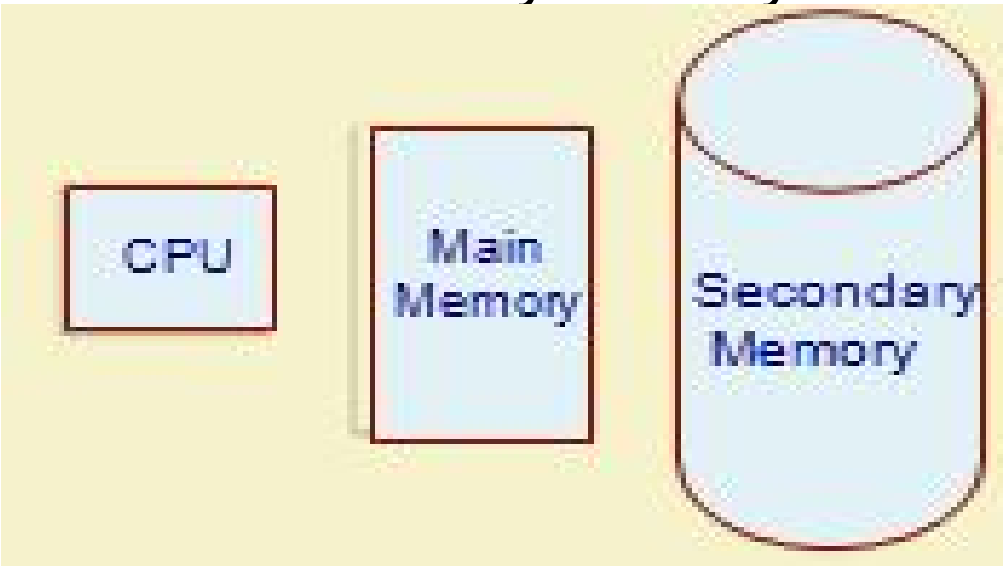
A **fast memory** (possibly organized in several levels) that sits between processor and main memory.

- Faster than main memory and relatively small.
- Frequently accessed data and instructions are stored here.
- Cache memory makes use of the fast SRAM technology.



What is Virtual Memory?

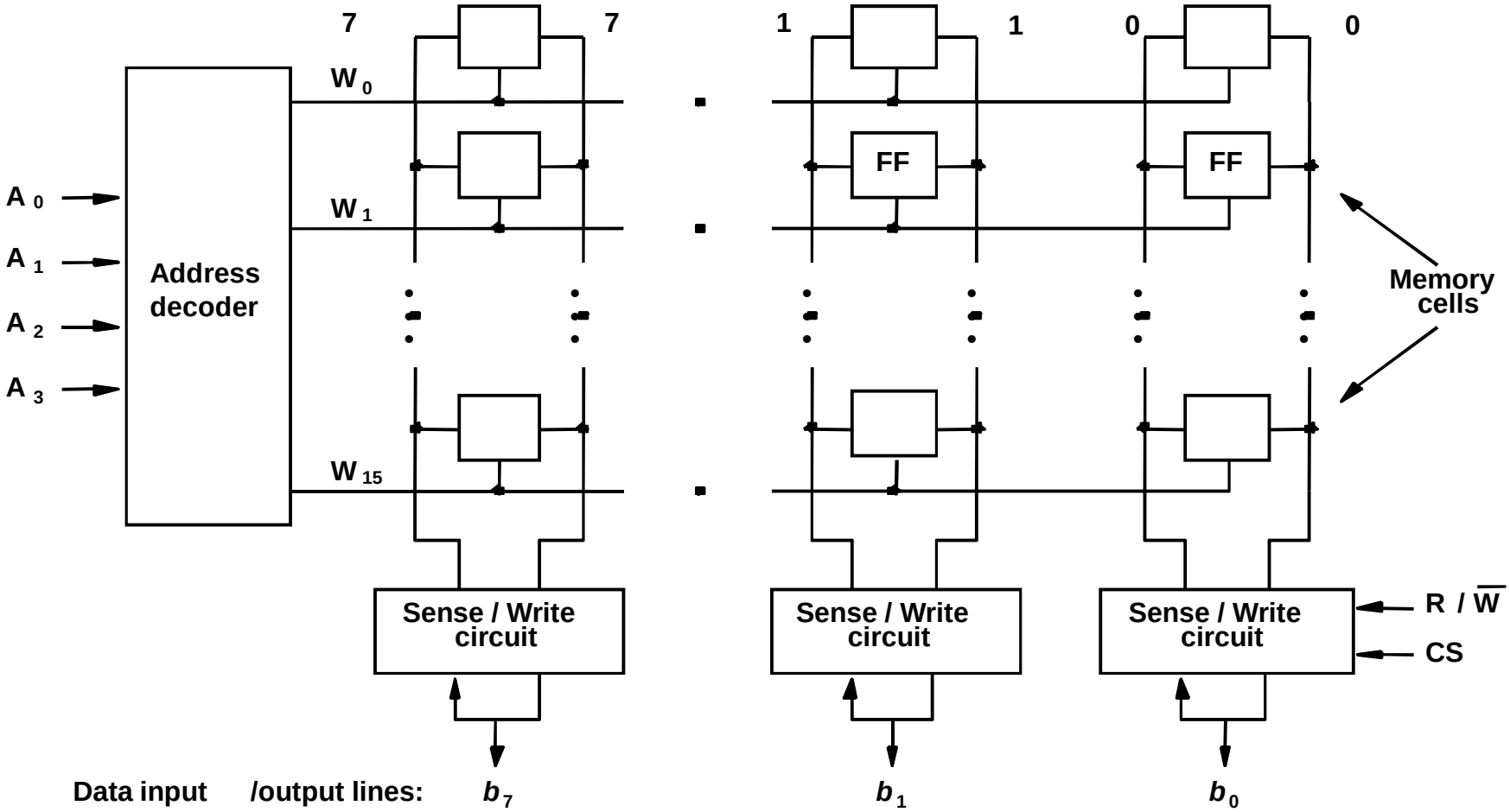
- Technique used by the operating system to provide an illusion of very large memory to the processor.
- Program and data are actually stored on secondary memory that is much larger.
- Transfer parts of program and data from secondary memory to main memory only when needed.



Internal organization of memory chips

- Each memory cell can hold one bit of information.
- Memory cells are organized in the form of an array.
- One row is one memory word.
- All cells of a row are connected to a common line, known as the “word line”.
- Word line is connected to the address decoder.
- Sense/write circuits are connected to the data input/output lines of the memory chip.

Internal organization of memory chips (Contd.,)



Broadly two types of semiconductor memory systems:

a)Static Random Access Memory (SRAM)

b)Dynamic Random Access Memory (DRAM)

i. Asynchronous DRAM

ii. Synchronous DRAM

Vary in terms of speed, density, volatility properties, and cost.

– Present-day main memory systems are built using DRAM.

– Cache memory systems are built using SRAM.

Static Random Access Memory (SRAM)

- SRAM consists of circuits which can store the data as long as power is applied.
- It is a type of semiconductor memory that uses bistable latching circuitry (flip-flop) to store each bit.
- SRAM memory arrays can be arranged in rows and columns of memory cells.
  - Called *word line* and *bit line*.

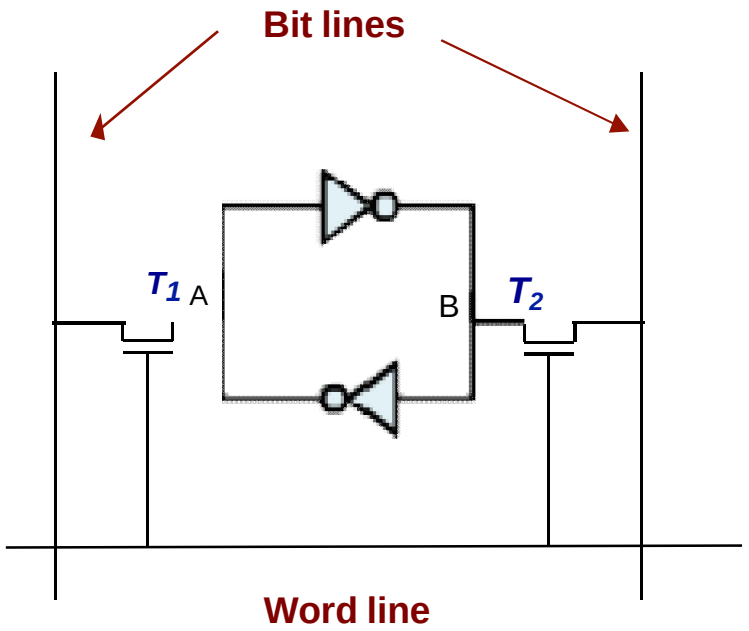
SRAM technology:

- Can be built using 4 or 6 MOS transistors.
- Modern SRAM chips in the market uses 6-transistor implementations for CMOS compatability.
- Widely used in small-scale systems like microcontrollers and embedded systems.
- Also used to implement cache memories in computer systems.

A 1-bit SRAM Cell

Two inverters are cross connected to form a latch.

- The latch is connected to two bit lines with transistors *T1* and *T2*.
- Transistors behave like switches that can be opened (OFF) or closed (ON) under the control of the word line.
- To retain the state of the latch, the word line can be grounded which makes the transistors off.



Features of SRAM

•Moderate / High power consumption.

– Current flows in the cells only when the cell is accessed.

–Because of latch operation, power consumption is higher than DRAM.

•Simplicity – refresh circuitry is not needed.

–Volatile :: continuous power supply is required.

•Fast operation.

–Access time is very fast; fast memories (cache) are built using SRAM.

•High cost.

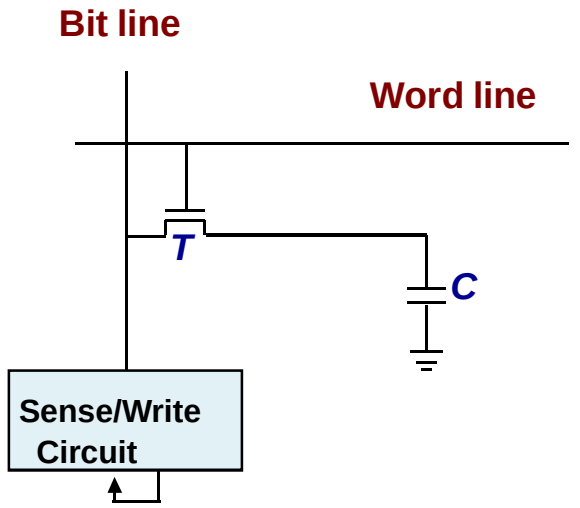
–6 transistors per cell.

•Limited capacity.

–Not economical to manufacture high-capacity SRAM chips.

Dynamic Random Access Memory (DRAM)

- Dynamic RAM do not retain its state even if power supply is on.
  - –Data stored in the form of charge stored on a capacitor.
- Requires periodic refresh.
  - –The charge stored cannot be retained over long time (due to leakage).
- Less expensive that SRAM.
  - –Requires less hardware (one transistor and one capacitor per cell).
- Address lines are multiplexed.



1-transistor DRAM Cell

Types of DRAM

a)Asynchronous DRAM (ADRAM)

–Timing of the memory device is handled asynchronously.

–A special memory controller circuit generates the signals asynchronously.

–DRAM chips produced between the early 1970s to mid-1990s used *asynchronous* DRAM.

b)Synchronous DRAM (SDRAM)

–Memory operations are synchronized by a clock.

–Concept of SDRAM came in the 1970s.

–Commercially made available only in 1993 by Samsung.

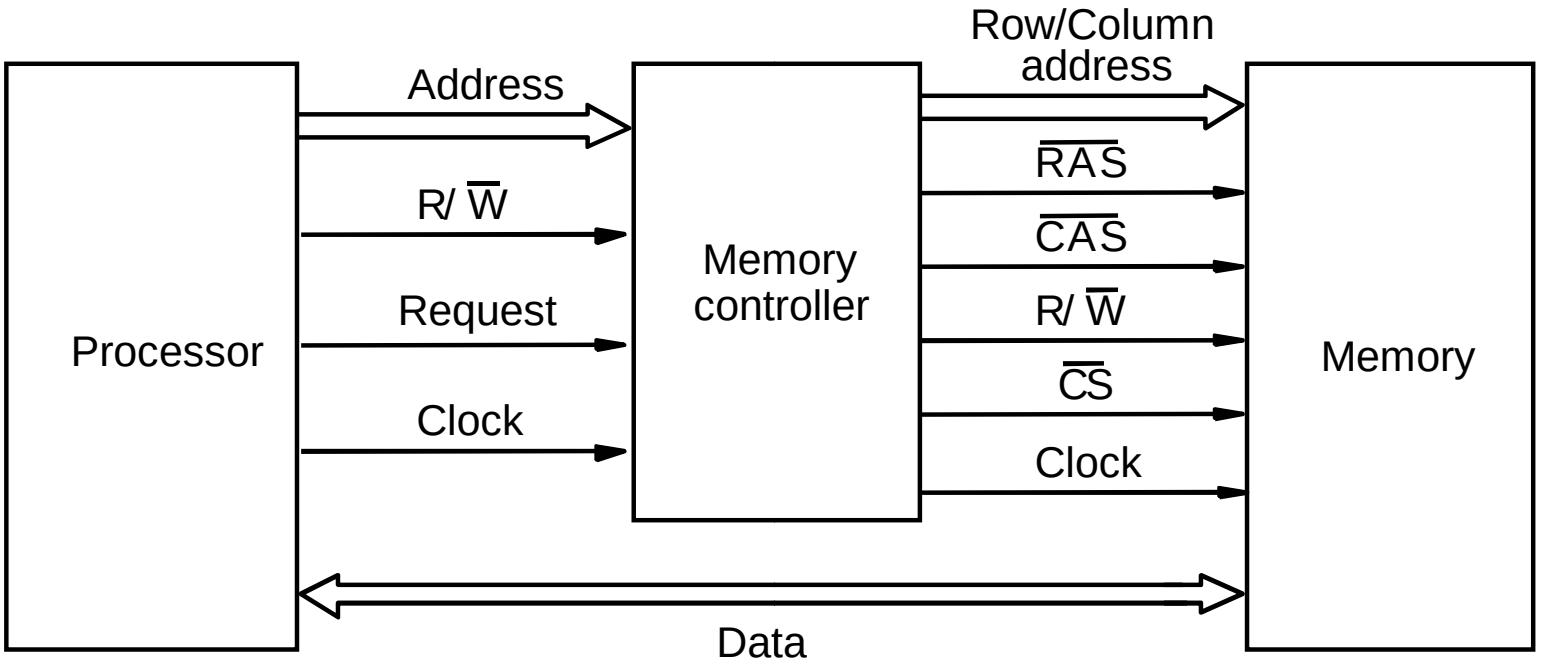
–By 2000 SDRAM replaced almost all types of DRAMs in the market.

–Performance of SDRAM is much higher compared to all other existing DRAM.

Memory controller

- Recall that in a dynamic memory chip, to reduce the number of pins, multiplexed addresses are used.
- Address is divided into two parts:
  - High-order address bits select a row in the array.
  - They are provided first, and latched using RAS signal.
  - Low-order address bits select a column in the row.
  - They are provided later, and latched using CAS signal.
- However, a processor issues all address bits at the same time.
- In order to achieve the multiplexing, memory controller circuit is inserted between the processor and memory.

Memory controller (contd..)



Read-Only Memories (ROMs)

- SRAM and SDRAM chips are volatile:
  - Lose the contents when the power is turned off.
- Many applications need memory devices to retain contents after the power is turned off.
  - For example, computer is turned on, the operating system must be loaded from the disk into the memory.
  - Store instructions which would load the OS from the disk.
  - Need to store these instructions so that they will not be lost after the power is turned off.
  - We need to store the instructions into a non-volatile memory.
- Non-volatile memory is read in the same manner as volatile memory.
  - Separate writing process is needed to place information in this memory.
  - Normal operation involves only reading of data, this type of memory is called Read-Only memory (ROM).

Read-Only Memories (Contd.,)

- Read-Only Memory:
  - Data are written into a ROM when it is manufactured.
- Programmable Read-Only Memory (PROM):
  - Allow the data to be loaded by a user.
  - Process of inserting the data is irreversible.
  - Storing information specific to a user in a ROM is expensive.
- Erasable Programmable Read-Only Memory (EPROM):
  - Stored data to be erased and new data to be loaded.
  - Flexibility, useful during the development phase of digital systems.
  - Erasable, reprogrammable ROM.
  - Erasure requires exposing the ROM to UV light.

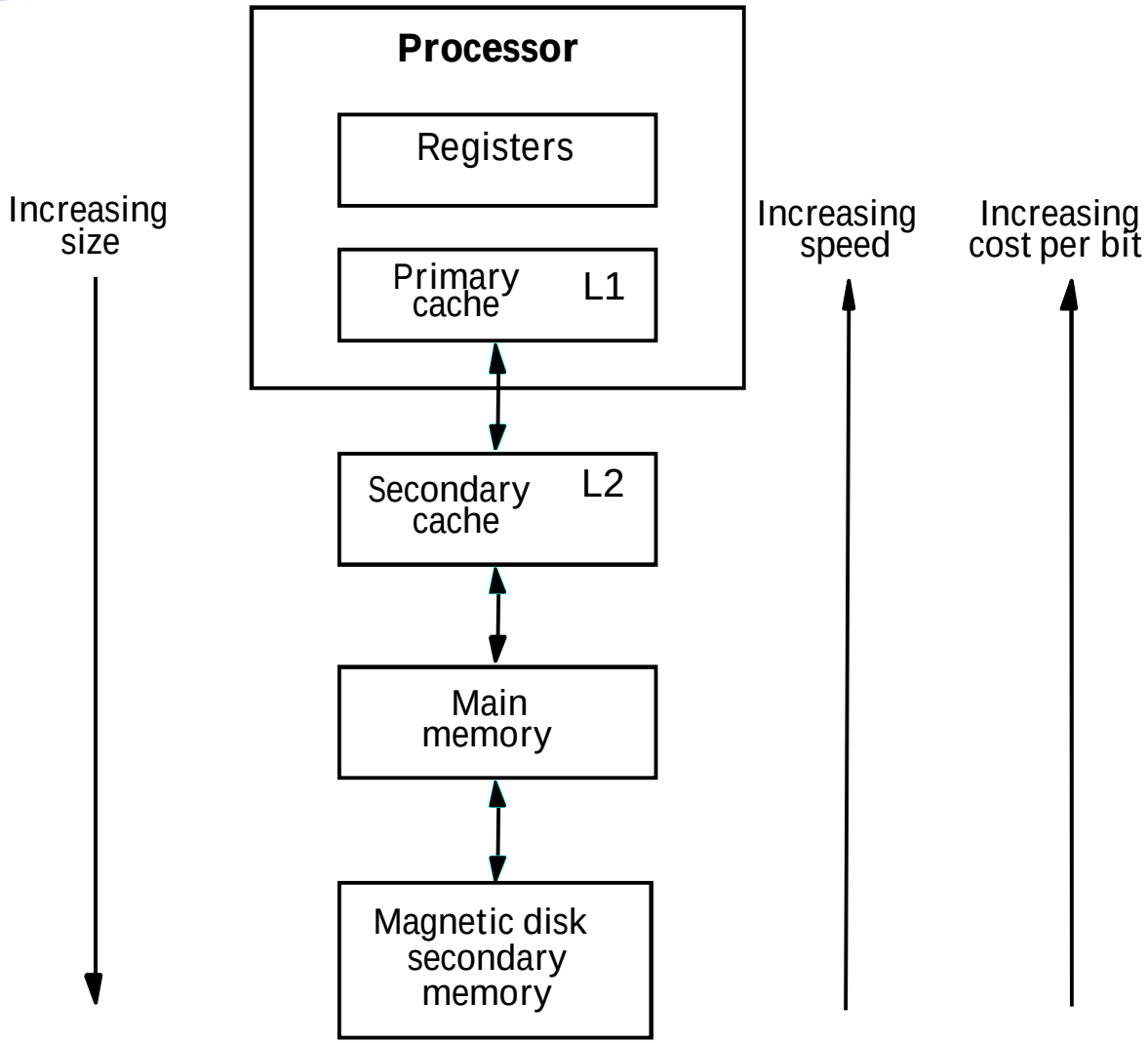
Read-Only Memories (Contd.,)

- **Electrically Erasable Programmable Read-Only Memory (EEPROM):**
  - To erase the contents of EPROMs, they have to be exposed to ultraviolet light.
  - Physically removed from the circuit.
  - EEPROMs the contents can be stored and erased electrically.
- **Flash memory:**
  - Has similar approach to EEPROM.
  - Read the contents of a single cell, but write the contents of an entire block of cells.
  - Flash devices have greater density.
    - Higher capacity and low storage cost per bit.
  - Power consumption of flash memory is very low, making it attractive for use in equipment that is battery-driven.
  - Single flash chips are not sufficiently large, so larger memory modules are implemented using flash cards and flash drives.

Speed, Size, and Cost

- A big challenge in the design of a computer system is to provide a sufficiently large memory, with a reasonable speed at an affordable cost.
- **Static RAM:**
  - Very fast, but expensive, because a basic SRAM cell has a complex circuit making it impossible to pack a large number of cells onto a single chip.
- **Dynamic RAM:**
  - Simpler basic cell circuit, hence are much less expensive, but significantly slower than SRAMs.
- **Magnetic disks:**
  - Storage provided by DRAMs is higher than SRAMs, but is still less than what is necessary.
  - Secondary storage such as magnetic disks provide a large amount of storage, but is much slower than DRAMs.

Memory Hierarchy



- *Fastest access is to the data held in processor registers. Registers are at the top of the memory hierarchy.*
- *Relatively small amount of memory that can be implemented on the processor chip. This is processor cache.*
- *Two levels of cache. Level 1 (L1) cache is on the processor chip. Level 2 (L2) cache is in between main memory and processor.*
- *Next level is main memory, implemented as SIMMs. Much larger, but much slower than cache memory.*
- *Next level is magnetic disks. Huge amount of inexepensive storage.*
- *Speed of memory access is critical, the idea is to bring instructions and data that will be used in the near future as close to the processor as possible.*

TEXT BOOK

Carl Hamacher, Zvonko Vranesic and Safwat Zaky, “Computer Organization”, McGraw-Hill, 6th Edition 2012.

REFERENCES

1. David A. Patterson and John L. Hennessey, “Computer organization and design”, MorganKauffman ,Elsevier, 5th edition, 2014.
2. William Stallings, “Computer Organization and Architecture designing for Performance”, Pearson Education 8th Edition, 2010
3. John P.Hayes, “Computer Architecture and Organization”, McGraw Hill, 3rd Edition, 2002
4. M. Morris R. Mano “Computer System Architecture” 3rd Edition 2007
5. David A. Patterson “Computer Architecture: A Quantitative Approach”, Morgan Kaufmann; 5th edition 2011

THANK YOU