Lecture 18: March 23, 2021
Computer Architecture and Organization-I
Biplab K Sikdar

Memory Subsystem

Memory subsystem is one of the major components of computer system (Figure 1).

It includes main memory (MM), secondary memory (SM) and a high speed component of MM (cache).

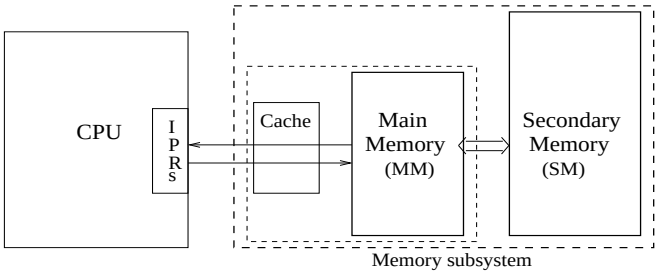


Figure 1: Computer memory subsystem

IPR (internal processor registers) are on-chip highly expensive CPU registers.

This class of memory unit can hold temporary results during computation in progress

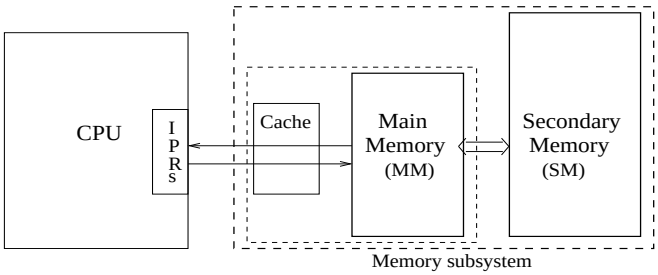
Operates at the CPU speed.

IPRs are are very few in numbers due to its excessive cost.

Cache - A special class of high speed main memory conventionally.

It is a product of Bipolar semiconductor technology.

Normally, cache unit is used to store part of a program currently in execution.



Main/primary memory - Programs (part/whole) that are currently in execution reside in MM.

CPU has direct access to MM.

MM is relatively faster and of moderately large capacity.

The cache is also a part of MM.

Secondary memory - Examples of common SMs are

- magnetic disk, optical disk, magnetic tape etc.

SM acts as the backing storage.

CPU cannot directly communicate with SM.

Conventionally, SM is inexpensive, slow and of large capacity than that of MM.

Objective: design of effective memory subsystem to achieve benefits in terms of

- Access time: It is a measure of efficient access to memory.

  The access time t_A = t_2 - t_1, where at t_1 CPU initiates memory read/write request and read/write operation is completed at t_2.

  We define *read access time* (t_{AR}) and *write access time* (t_{AW}).

  The declared access time of a memory can be

  $$t_A = \frac{t_{AR} + t_{AW}}{2}.$$

  t_{AR} - It is measured as t_4 - t_3 (Figure 2(a)).

  t_{AW} - It is measured as t_6 - t_5 (Figure 2(b)).

  Normally, $t_{AW} > t_{AR}$. However, in some memories,

  $$t_{AR} \simeq t_{AW} = t_A.$$

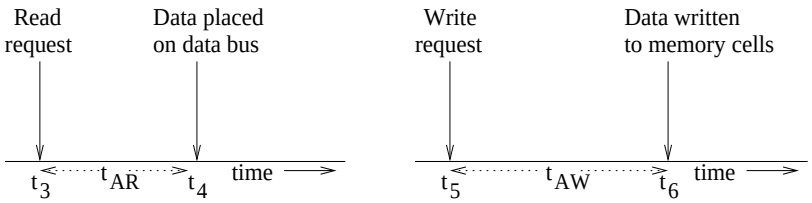


Figure 2: Memory access time

- Access mode: Access time depends also on mode of data access.

  a. *SAM (serial access memory)* - In SAM, information can be accessed only in predetermined order.

  Access time is - function of location being accessed.

  Normally, access time for SAM is proportional to the distance of location of data from a reference position.

  Common example of SAM is *magnetic tape*.

  SAMs are extremely low cost.

  b. *RAM (random access memory)* - The read/write access time for RAM is constant - does't depend on location of data.

  RAMs are classified as RWM and ROM (Figure 3).

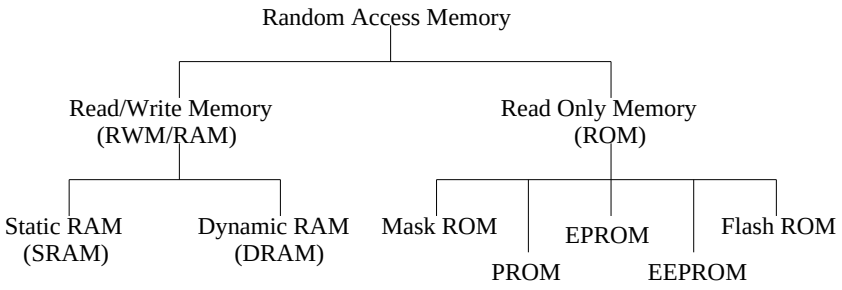


Figure 3: RAM arrays

c. *DAM (direct access memory)* - Access time in DAM is semi-random.

It depends on location/position of data item to be accessed but not exactly proportional to the distance. Magnetic hard disk is a DAM.

Access to find an item from disk demands placement of the read/write head on proper track of disc and a rotation of disk to place part of track (sector), containing data item, just under disk head.

Access time = time taken to move disk head + rotation time of disk.

d. *CAM (content addressable memory)* - It realizes a special form of random access to data items.

An item/word of memory is retrieved based on the part of word's content.

- Alterability: It defines whether content of memory can be modified by CPU during program execution.

  The read/write memory (RWM/RAM) content is alterable.

  Content of conventional ROM cannot be altered during program execution.

- Permanance of storage: In a non volatile memory (ROM), information once recorded remains as it is (even if the power is withdrawn) until deliberately changed.

  On the other hand, in a volatile memory (RAM), information decays once power is withdrawn.

  Information stored in a static RAM (SRAM) can be kept as it is through constant supply of power.

  In dynamic RAM (DRAM), information decays even if power is on.

  In destructive read out (DRO) memory, if a data is read, data in memory is lost (Figure 4).

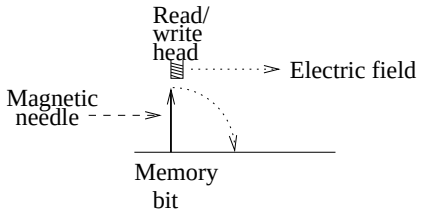


Figure 4: DRO read

The *loss* or *decay* of information adds hardware overhead/delay to CPU instruction execution time (cycle time).

- Cycle time - delay between initiations of two successive memory operations.

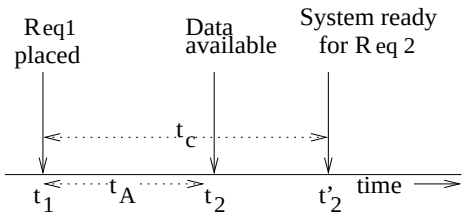


Figure 5: Cycle time

It is - t_A + additional time required before a second access can commence.

Additional time - non-destructive write in DRO, refresh time in DRAM etc.

The cycle time is, therefore,

t_c = t'_2 - t_1 (Figure 5).

In DRAM, if refreshing is required in T_R interval and refresh time is T_r, then

$$t'_2 = t_2 + \frac{T_r}{T_R} \times t_A.$$

In DRO memory, every read operation is followed by a non-destructive write (Figure 6). That is, for DRO memory read,

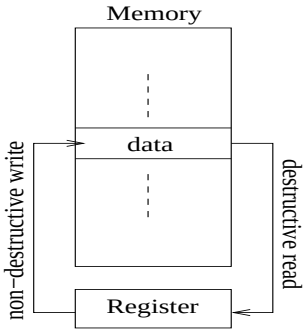


Figure 6: Destructive readout memory

t'_2 = t_2 + restoring time required for non-destructive write.

- Cost/bit: Target of memory subsystem design is to reduce cost per bit of memory interfaced.

  The cost is due to cost of memory cells, peripheral support/access circuitry and memory refresh/restoring logic.

  Cost per bit of B-bit memory unit is - total cost of memory divided by B.

  An approximate relation between t_A and cost/bit of memories is in Figure 7.

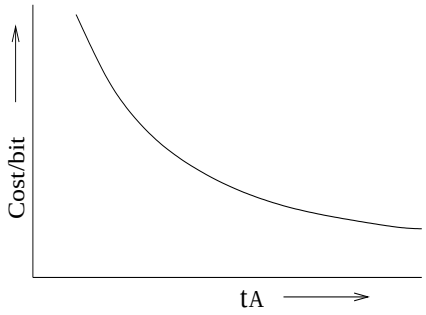


Figure 7: Memory cost and access time

Other than cost of memory, CPU-memory interfacing should be cost effective.

0.1 Memory Interfacing

Basic organization of a memory device is shown in Figure 8.

In addition to power lines, a memory chip consists of -

(i) m address lines to select one out of 2^m memory locations.

(ii) d bidirectional data lines for data transfer with CPU.

(iii) Read/write signal line(s): read (from memory) or write (to memory cells).

In Figure 8, memory module is having only one signal line $RD/\overline{WR}$

$RD/\overline{WR}$ = 1 $\Rightarrow$ read from memory

$RD/\overline{WR}$ = 0 $\Rightarrow$ write to memory

(iv) At least one chip-select line to enable a chip to be ready for read/write opeartion.

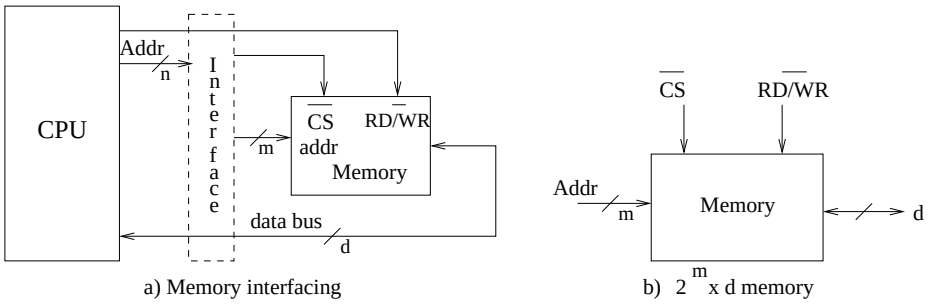


Figure 8: Memory interfacing