Memory Organization

0.1 Introduction

The ever increasing size of the microprocessor die, coupled with an exponential increase in integration density has led to an extremely high demand for larger and denser memory technologies at all levels of the memory hierarchy. It has often been speculated that, for most of the memory intensive tasks in the microprocessor, the memory bandwidth is the key bottleneck, a hurdle that cannot be solved by simply integrating higher memory capacity. With the advent of mobile devices and a plethora of hand-held devices and an ever shrinking power budget, providing memory solutions that can provide adequate bandwidth within a low-power envelope has become an insurmountable task. This has been further exacerbated by increased device parameter variation that has resulted from technology scaling. The need for faster, lower power, and denser memory arrays is not confined to the microprocessor industry only. Digital Signal Processors (DSPs), Field Programmable Gate Arrays (FGGAs) as well as custom Application Specific Integrated Circuits (ASICs) are all expected to demand better memory solutions over the next decade.

The Memory access may be synchronized by with, or may work in *asynchronous mode*. The *Memory access time* is time from request made to the time supply of data. A *memory cycle time* is delay time between two consecutive read requests.

Virtual memory feature system extends the RAM memory of a computer beyond it physical memory, and can be theoretically as large as the capacity of the secondary storage hard-disk.

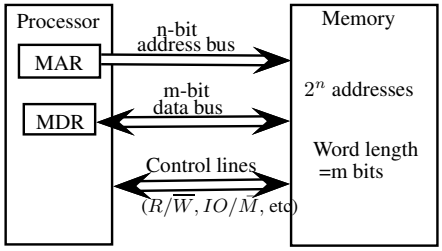

Figure 1: Memory processor Processor Interface.

0.2 Memory Hierarchy

Earliest computers used ferro-magnetic core as main memory, where as the present memories are semiconductor memories. The memory system of a computer is classified as: Internal Processor Memory, Main Memory, Secondary storage.

When compared the progress of memory v/s CPU performance, the rate of growth of CPU's speed has remained much higher than the memory (see figure 2). This necessitated the memory hierarchy. The figure 3 shows the general memory hierarchy of computers.

The Computer pioneers predicted that programmers would need unlimited amount of fast memory, which is an expensive solution. The economical solution to that is *memory hierarchy*. This takes advantage of (1) locality of reference and (2)cost performance of memory technology.

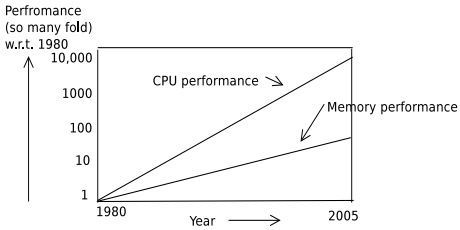

Figure 2: CPU v/s Memory performance improvement.

0.2.1 Memory characteristics

Let cost C is total cost of memory system of size S towards the purpose of *storage* and *access* mechanism, taken together, then

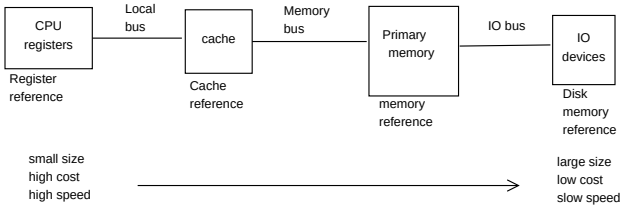

Figure 3: Memory hierarchy in Computers.

$$c = \frac{C}{S}$$

is cost in dollars per bit.

Access time: t_A is time between request made to information delivered.

The access rate in bits/sec, called *band width* is defined as,

$$b_A = w/t_A$$

The w is databus width in bits.

Access mode: is defined differently for *random access* memory (RAM), and *serial access* memory.

RAM is costly because every location has separate access mechanism. For serial access, access mechanism (see fig. 4) is shared among many locations. However, some disk drives are *semi-random*, i.e., (*direct access*), where read/write heads of all recording surfaces access them all simultaneously.

The other classes of memories are:

Non-volatile v/s *volatile*: In volatile memories data is lost when power removed, where a in non-volatile it does not.

Static v/s *dynamic memories*: The dynamic memory requires frequent refreshing, while static does not require it.

Non-destructive readout (NDRO) v/s *Destructive readout* (DRO): In descriptive read out memories, every read operation is followed with write operation, because a read operation destroyes the contents of location which has been read.

The figure 4 shows the access mechanism of random and serial memory system.

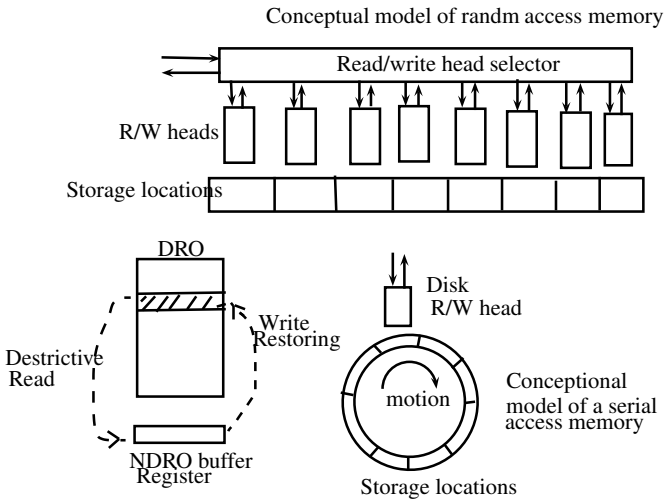

Figure 4: Random and serial access.

0.3 Memory Terminologies

The figure 5 shows a memory cell with read-write operations.

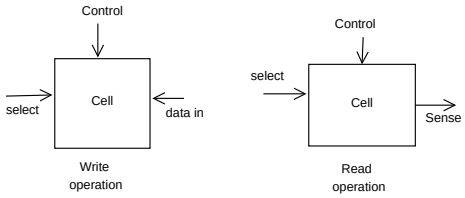

Figure 5: A memory cell with R/W operations.

Following are the various semiconductor memories:

ROM(read only memory): It is random access type, and non-volatile, i.e., on power off its contents remains unchanged.

RAM (random access memory): These are Read/write type, but volatile, i.e, on power loss their contents are lost.

PROM (Programmable ROM): It is just like ROM, in addition it can be

programmed by end user once only.

EPROM (Erasable PROM): They can be erased any number of times by exposing with UV (ultraviolet) rays, and can be reprogrammed.

EEPROM: These are Electrically erasable at byte level, i.e., every by can be individually selected for modification. They are Non-volatile.

Flash memories: These also called *bubble memories*. They allows reading and erasing at block level. The examples are pen-drive, memories in camera, etc.

ROM memories are used as control memories in microprogramming, they store libraries' subroutines for frequently used functions, for Boot ROMs, Function tables, modest size programs and data are kept in ROMs, for system programs. Data are wired in these as part of the fabrication process.

The flash memories are read/write/erase: block-by-block. They do not have byte level structure like other memories, since it uses one transistor per bit, it provides higher storage density. The flash stores information in floating gate transistors. Due to this they do not discharge for many years. The advantage is low power dissipation, non-volatile, compact, and a good alternate to CD (compact disk) and Hard Disk storage. The disadvantage is slow speed compared to disk storage as well lesser reliable. So, they do not appear to be an alternative solution for hard-disks. The figure 6 shows a one-bit storage cell for bubble memory.

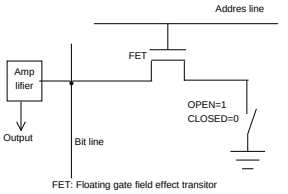

Figure 6: 1-bit ROM cell for bubble memory.

The charge is stored in the floating gate region of the FET, representing binary '1' or '0'. When the switch is open, the charge remains, representing '1'. When the switch is closed, no charge can be held in the transistor gate, hence, 0 remains stored.

0.3.1 Some properties of memories

Cycle-time and data transfer rate: The definition of t_A (access time) is not applicable for dynamic RAMs. The minimum time between two memory requests is longer than t_A due to refreshing cycle. The time between two consecutive memory reads is called the *cycle time of memory* (t_M).

Memory Latency (L): It is delay from request by processor to delivery of word from memory.

Bandwidth (BW): The bandwidth for a memory is $\frac{w}{L}$. If R is number of requests which can be served simultaneously, then $BW = \frac{R*w}{L}$.

An *ideal memory* is infinite capacity, zero latency, infinite bandwidth, which are not achievable practically. Fortunately, memory hierarchy provides decreased latency, hence does require high BW. In addition, the Parallel interleaved memory helps to increase the BW.

0.4 Dynamic RAM Memory structures

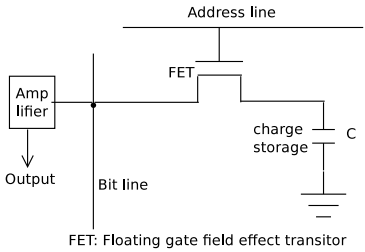

Figure 7: 1-bit DRAM Cell.

Over a time of msecs., the capacitor holding logic 1/0 gets discharged, hence requires *refreshing*.

The address line is to be activated for read/write operation. The transistor acts as switch. Combination of Bit line '1/0' and address line high, causes writing '1/0' at the selected cell. For memory Read, address line high will select the cell, stored value is sensed, amplified and sent at output of sense amplifier. Read operation is destructive, hence capacitor charge needs to be rebuilt periodically (refreshing).

0.4.1 Static RAM (SRAM) Cell

The figure 8 represents a single cell of static RAM.

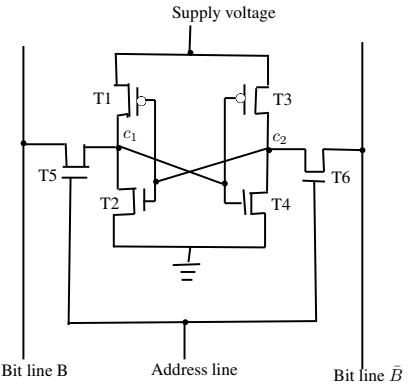

Figure 8: 1-bit SRAM cell.

Write 1: Keep Bit line B as 1, address line high (i.e. selected). This transfers logic 1 to point C_1 and base of T_4. Thus T_4 conducts, causes to lower C_2 voltage, which in turn lowers base of T_2. Hence, T_2 switches off, and T_4 is switched ON. T_1, T_3 remains ON, OFF, respectively (note the -ve true input).

When address line and Bit line are deselected, following states continue to remain: T_2, T_3 off, T_4, T_1 ON. *Write 0* is done in similar way.

READ: Raising the address line high will transfer voltage (stored logic) through T_5, T_6 to bit lines.

0.4.2 SRAM v/s DRAM

Both SRAM and DRAM are volatile memories. DRAMs require refreshing circuit because the stored charge in capacitor decays with time. Also, the read operations is destructive, hence needs re-writing. Refreshing circuit's cost is high for small size DRAMs. Hence, DRAMs are preferred for large capacities.SRAMs are faster than DRAMs. Due to cost and speed SRAMs are used in caches, while DRAMs are used for RAM (Main Memory).

The figure 9 shows the general configuration for RAM memory. The Address drivers and R/W drivers are current drivers. The single data register are used for non-parallel operation.

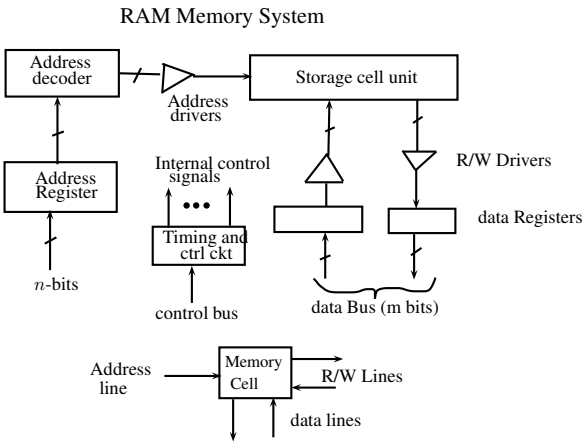

Figure 9: Memory Block diagram.

One dimensional RAM: IF number of address lines are n, then total 2^n words can be stored. The figure 10 shows one dimensional RAM.

In the two-dimensional RAM (see fig. 11), the memory cells are arranged in two-dimensional form. If number of address lines are $x + y$ for horizontals and vertical address decoders then total $2^x \times 2^y$ words can be stored.

Example 0.4.1 *Number of current drivers for memory.*

Considering that there are n^2 number of memory cells in a read/write memory, a single dimensional organization of this will require a total of n^2 number of current drivers. However, when arranged in 2-D organization, there are n drivers for x and n drivers for the y array, making total $2n$ drivers. Thus, for the same memory size of n^2, the ratio of count of drivers for 1-D and 2-D organization of memory cells is

$$\frac{n^2}{2n},$$

which is $n/2$. In other words, for single dimensional organization the drivers in crease quadratic to the number of drivers in 2-D. □

0.4.3 Refreshing Mechanism for DRAM

The figure 12 shows the block diagram for refreshing of DRAM memory system. Refreshing is interleaved with memory R/W operations.

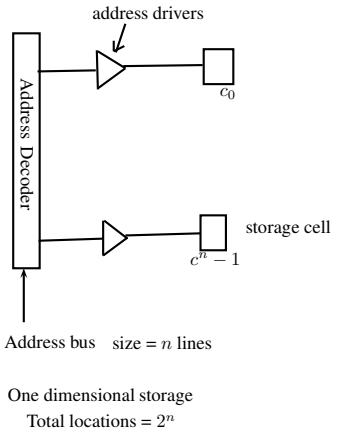

Figure 10: Single Dimension RAM.

For carrying out the refreshing, all cells are required to be refreshed once every 2 msec. So that they do not get discharge to the level where contents are not recoverable. All 128 row are read and recharged to original logic 1/ 0. Refreshing is automatically triggered by an interval timer. All 128 rows are refreshed in 128 x 500 nsec. = 64 μ sec. Since refresh is *interleaved* with R/W operation, max. delay in any R/W can be only 500 nsec. The delay is taken care of by RDY line. Since all cells are refreshed in 2 msec., the fraction of time for refresh is only $\frac{64\mu \text{ sec}}{2msec} \simeq 3.2$ percent. An arbiter selects the row address 0-127 for the purpose of refreshing operation.

If memory access time is t_A, and read starts at time t_0, then data is made available at $t_0 + t_A$ time if no refresh cycle exists with current read cycle, otherwise it is made available at time $t_0 + t_A + 500$(nsec).

0.5 Serial Access Memories

The serial access memories finds their applications in bulk storage, with per bit cost low. Information is stored in tracks, where each cell stores one bit of information. A specified item is accessed by moving R/W head or medium, or both. Conceptually these are shift registers with limited access points.

t_s: *Seek time:* - time to move R/W head from one track to another.

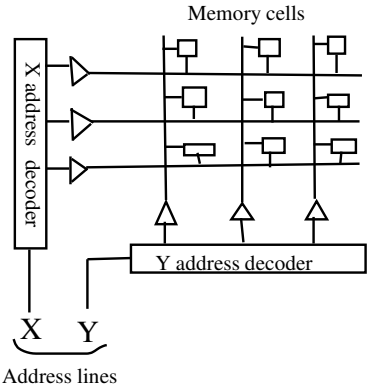

Figure 11: Two Dimensional RAM.

t_l: *Latency time:* - time to reach to data location, once the head has reached to the desired track.

A word of w-bits may be stored serially on a track or it can be stored on $|w|$ number of tracks, reading/writing all in parallel.

Data transfer rate: Bit rate at which information can be transferred to/from the track.

For each track of N-words, rotation speed (of disk) of r rotations per sec., n is number of words per block, the data rate of memory is rN words per sec. Once the head is positioned, a block can be transferred in $\frac{n}{rN)}$ secs. The approximate time to transfer a block can be given by,

$$t_B = t_s + t_l + \frac{n}{rN} \tag{1}$$

0.5.1 Magnetic Surface Recording

Each cell in a track has two stable states, that represent 0 and 1. The direction of current in the coil decide the state of the cell. The readout process is nondestructive. The magnetic surface is nonvolatile. On magnetic disk, the tracks form concentric circles, and on magnetic tape tracks form parallel lines on the surface of long narrow plastic tape. Figure 13 shows the working of magnetic tape memory system.

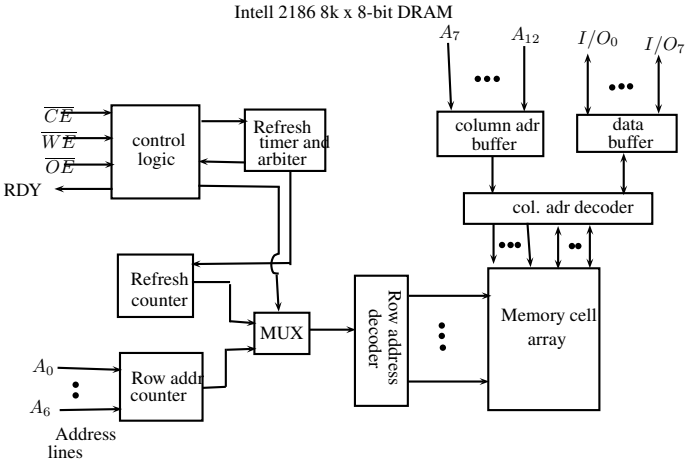

Figure 12: Refreshing mechanism for Dynamic RAM.

0.6 Magnetic Disk Read-Write

Several hundred of tracks are arranged as concentric circles. Several disks attached to a common spindle, rotated at constant speeds, which can be any of: 5400, 7200, 10k, 15k rpm. All the heads move in unison. The arm is moved in fixed linear path.

Disk memories have been also designed with one head per track (no need of head movement). Figure 14 shows the working principle of magnetic disk memory.

The figure 15 shows the detailed working of a 4 read/write head magnetic disk storage system.

The Output of address decoder determines: (1) track to be used, and (2) location of desired block of information in the track (i.e., block address). The following are steps of Disk operation:

Track address determines particular R/W head. The head is moved on to the track. Some track position indicator is needed, which is generated when the track passes under the R/W head. The generated address is compared with the block address produced by the address decoder. When they match, the selected head is enabled and data transfer begins. The R/W head is disabled when complete block of information is transferred.

The memory input and output registers are shift registers of parallel/serial I/O data.

Following are the typical data for Magnetic disk memory system.

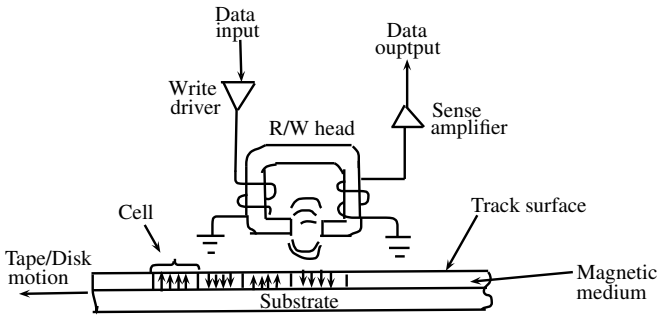

Figure 13: Magnetic surface recording.

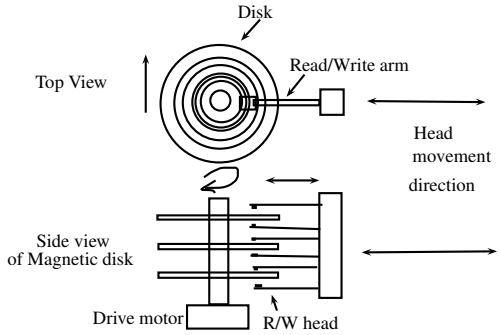

Figure 14: Magnetic Disk storage system.

NEC D2257	
No. of recording surfaces	8
No. of tracks per recording surface	1024
no. of R/W heads per recording surface	1
Track recording density	9420 bits/in.
Storage capacity of track	20,480 bytes
Totals Storage capacity of disk drive	167.7M bytes
Disk rotation speed	3510 rpm
Average seek time	20 ms
Average latency	8.55 ms
Data transfer rate	1.198 bits/s

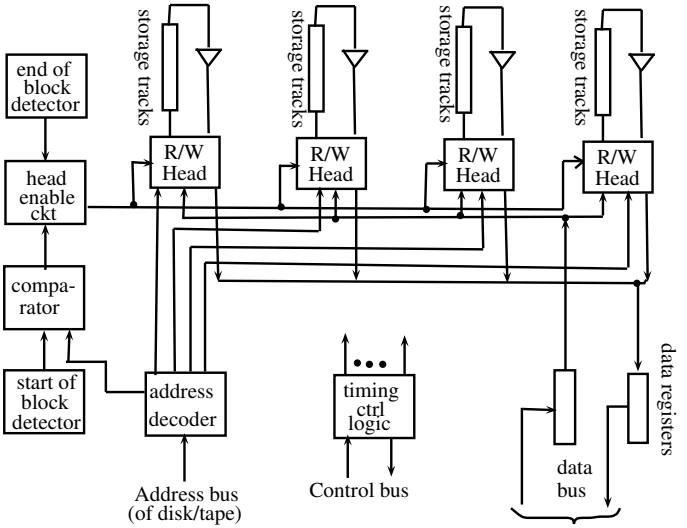

Figure 15: Magnetic Disk memory W/R System with 4-heads

0.7 Magnetic Tape Memories

The figure 16 shows the schematic a magnetic tape storage system.

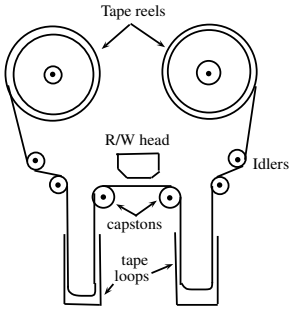

Figure 16: Magnetic tape storage.

It is a Compact inexpensive portable medium for storing large amount information. Information is stored in binary, unlike audio and video tapes, in 9 parallel tracks. Basic memory word is 9-bits: 8-bit information with 1 parity bit. It is Packed in cartridges/cassettes, and resemble audio tapes. Medium is not in continuous motion. When request re-

ceived, tape is moved forward/ backward to desired location. The Loops are for rapid start-stop. Capstans accelerate tape. Later it moves at tape speed, at which the data transfer takes place. Typical storage density: 1600 bytes/in., speed is 18.75 in./s. Thus, data tr. rate = 1600 * 18.75 = 30,000 bytes/s. Info. stored in blocks with block gap. Rewind time $\approx$ 1 min.
Some other characteristics:

- Storage density, per unit area
- Energy consumption (may or may not require cooling)
- Reliability: MTBF (mean time between failures)
- Maximum Number of read/write operations (life-time)

Exercises

1. (a) How many 128×8 RAM chips are required to provide a memory capacity of 2048 bytes?
   (b) How many address lines are required for accessing 2048 bytes? How many are common for all the memories?
   (c) How many lines must be decoded for chip-select? Specify the size of chip select?

2. What is transfer rate of 8-track magnetic tape whose speed is 120 inches per sec. and whose density is 1600 bits per inch?

3. Consider a dynamic RAM that must be given a refresh cycle of 64 times per msec. Each refresh operation requires 100 ns. A memory cycle requires 200 nsec. What percentage of the memory's total operating time must be given to refreshes?

4. The memory of a particular microcomputer is built from 128k X 1 DRAMs. According to the data-sheet, the cell array of the DRAM is organized into 512 rows. Each row must be refreshed at least once every 4 ms. Suppose we refresh memory on a strictly periodic basis.
   (a) What is the time period between successive refresh requests?
   (b) How long a refresh address counter do we need?

5. A certain moving arm disk-storage device has the following specs.

No. of tracks per recording surface	200
Disk rotation speed	2400 rpm
Track storage capacity	62,500 bits

   Estimate the average latency and the data transfer rate of this device.

6. A magnetic-tape system accommodates 2400-ft reels of standard nine-track tape. The tape is moved past the recording head at a rate of 200 in./s.

   (a) What must the linear tape-recording density be in order to achieve a data-transfer rate of 10^7 bits/s?

   (b) Suppose that the data on the tape is organized into blocks each containing 32k bytes. A gap of 0.3 in. separates each block. How many bytes may be stored on the tape?

7. Figure 17 shows a simplified timing diagram of a DRAM read operation over a bus. The access time is considered to last from t_1 to t_2. Then there is a recharge time, lasting from t_2 to t_3, during which the DRAM chips will have to recharge before the processor can access them again.

   (a) Assume that the access time is 60 ns and recharge time is 40 ns. What is the memory cycle time? What is the maximum data rate this DRAM can sustain, assuming a 1-bit output?

   (b) Constructing a 32-bit wide memory system using these chips yields what data transfer? (RAS=row addr select, CAS=col. addr select)

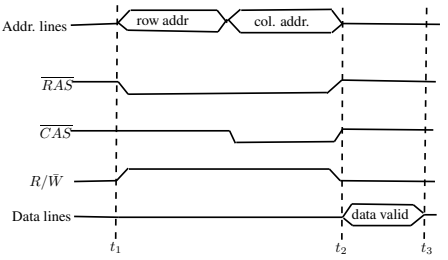

Figure 17: Timing diagram of DRAM.

8. Design a 16 X 4-bits RAM using 4X2-bits ICs.

9. Design a 16-bit memory of total capacity 8192 bits using SRAM chips of 64×1 bits. Give the array configuration of the chips on the memory board showing all required input and output signals for assigning this memory to the lowest address space. The design should allow for both byte and 16-bit word accesses.

Bibliography

[1] John P. Hayes, "Computer Architecture and Organization", 2nd Edition, McGraw-Hill, 1988. (chapter 5)

[2] William Stalling, "Computer Organization and Architecture", 8th Edition, Pearson, 2010. (chapter 5)