Computer Organization
(Memory Organization)

KR Chowdhary
Professor & Head
Email: kr.chowdhary@gmail.com
webpage: krchowdhary.com

Department of Computer Science and Engineering
MBM Engineering College, Jodhpur

November 14, 2013

Memory

- Memory access may be synchronized by bus, or may work in asynchronous mode.
- *Memory access time:* time from req. to supply of data, *memory cycle time:* delay time between two consecutive read requests.

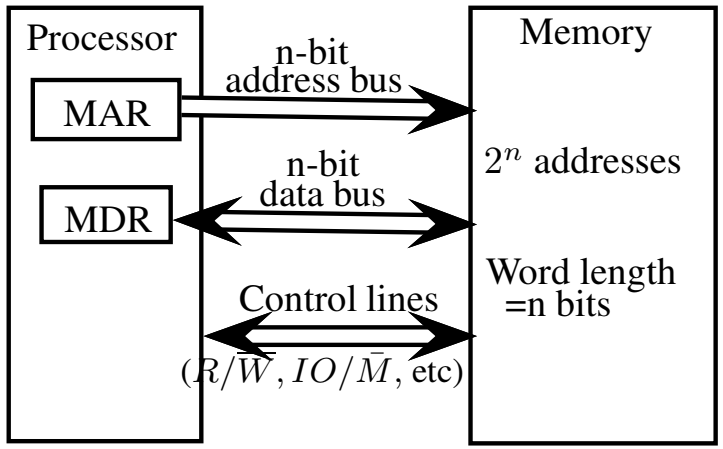


Figure 1: Connection of memory to processor

- *Virtual memory:*?

Memory hierarchy

- Internal Processor Memory, Main Memory, Secondary
- Computer pioneers predicted that programmers would need unlimited amount of fast memory.
- Economical Solution to that is memory hierarchy
- This takes advantage of (1) locality of reference and (2)cost performance of memory technology.

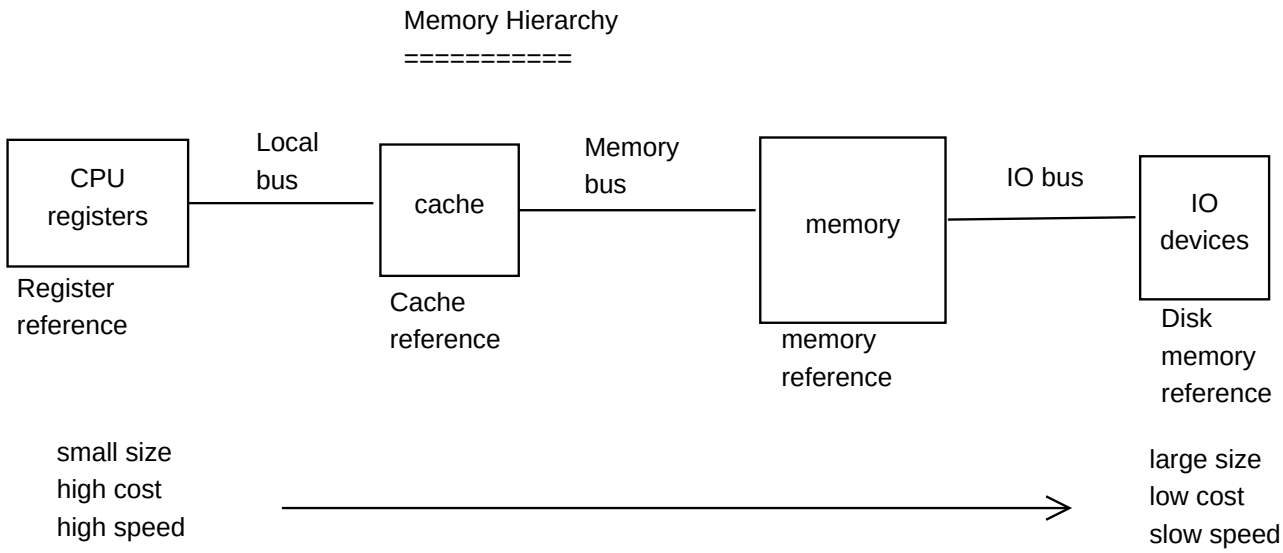


Figure 2: Memory hierarchy in Computers

- Earliest computers used ferromegnetic core as memory

Memory v/s CPU performance

- The rate of growth of cpu's speed has remained much higher than memory
- This necessitated the memory hierarchy

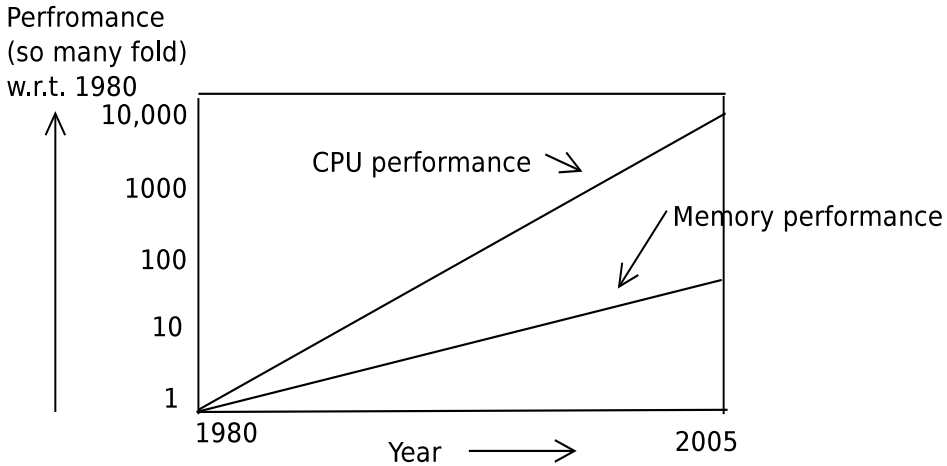


Figure 3: CPU v/s Memory performance improvement

Some characteristics

- Cost(c): If C is total cost of memory system of size S (for storage + access), $c = \frac{C}{S}$ dollars per bit.
- Access time: t_A is time between request made, to information delivered. $b_A = w/t_A$ is access rate in bits/sec(Band Width).
- Access mode: (1) Random access memory (RAM), (2) Serial access memory.
- Random: RAM is costly because every location has separate access mechanism.
- For serial access: access mechanism is shared between many locations. Some disk drives are *semi-random*, i.e., (direct access) - R/W heads of all recording surfaces access them all simultaneously
- Other classes: (1)non-volatile/volatile -data is lost if power removed, (2) static /dynamic memories - requires frequent refreshing.
- Non-destructive/Destructive readout - every read operation is followed with write.

Access mechanisms

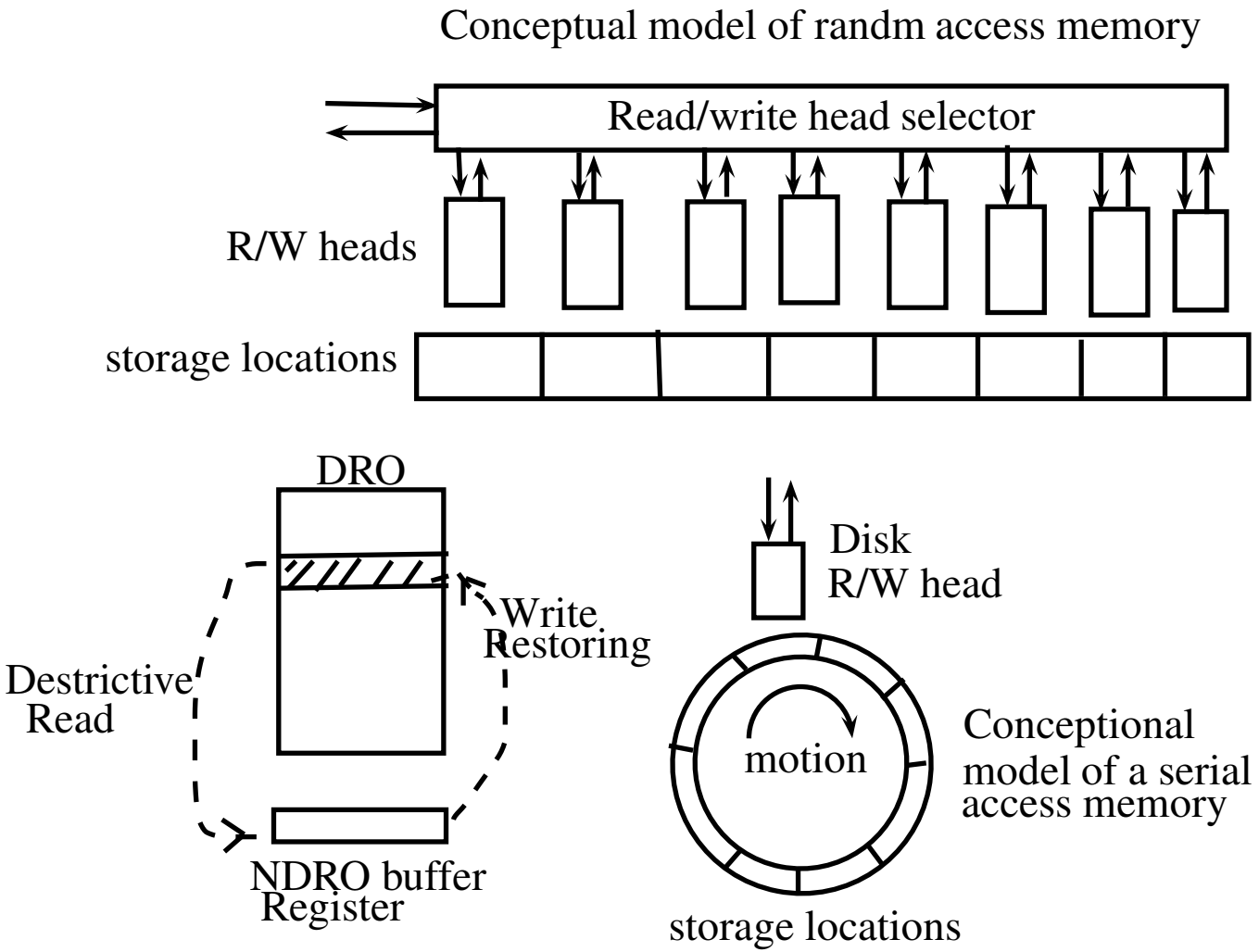


Figure 4: Random and serial access.

Memory Cell and Terminology

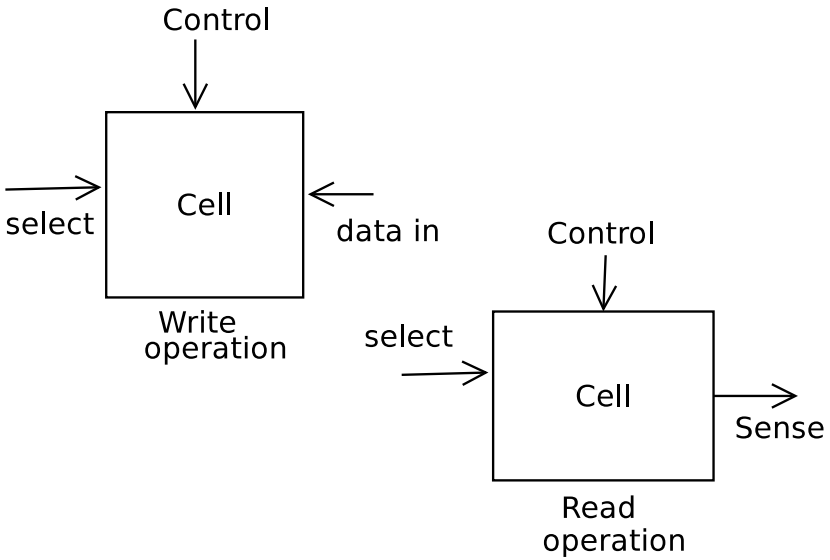


Figure 5: A memory cell with R/W operations

RAMs v/s ROMs

- ROM: random access, non-volatine, read-only
- RAM: Read/write type, volatile
- PROM (Programmable ROM): It can be programmed by end user once only
- EPROM (Erasible PROM): Can be erased any number of times by exposing with UV rays.
- EEPROM Electrically erasible byte level, Non-volatile
- Flash memory (Bubble memory): Reading and erasing block level (pendrive, camera, etc.)

ROM v/s Flash memories

Applications of ROM:

- Microprogramming, stores library subroutines for frequently used functions, Boot ROMs, Function tables, Modest size programs and data are kept in ROMs, System programs, Data are wired as part of the fabrication process

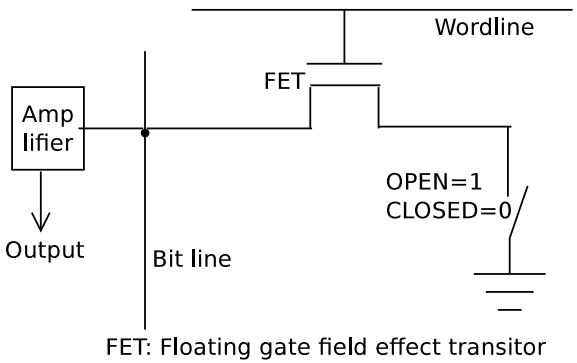


Figure 6: 1-bit ROM Cell.

Flash Memories:

- Read, write, erase: Block by block
- There is no byte level structure like other memories
- One transistor per bit provides higher density.
- Stores information in floating gate transistors. Does not discharge for many years.
- ADV: Low power dissipation, non-volatile, compact, alternate to CD and HDD.
- Disadv: Slow speed.

Some properties of memories

- Cycle-time and data transfer rate: The definition of t_A is not applicable for dynamic RAMs. The minimum time between two memory requests is $> t_A$ due to refreshing cycle. The time between two consecutive memory reads is called the cycle time of memory (t_M).
- Memory Latency (L):= Delay from request by processor to delivery of word from memory. Bandwidth (BW)=$\frac{w}{L}$. If R is no. of requests which can be served simultaneously, then $BW = \frac{R*w}{L}$.
- Ideal memory: Infinite capacity, zero latency, infinite bandwidth, which are not achievable practically.
- But, Memory hierarchy provides decreased latency, hence does require high BW. In addition, the Parallel interleaved memory helps to increase the BW.

Dynamic RAM Memory structures

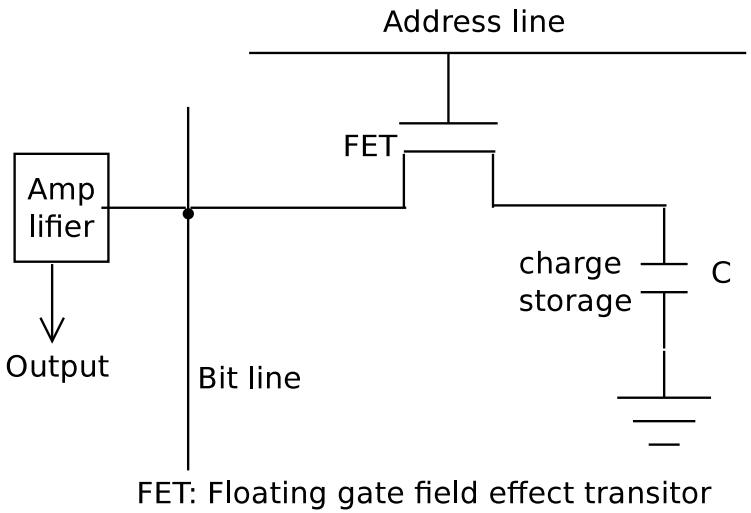
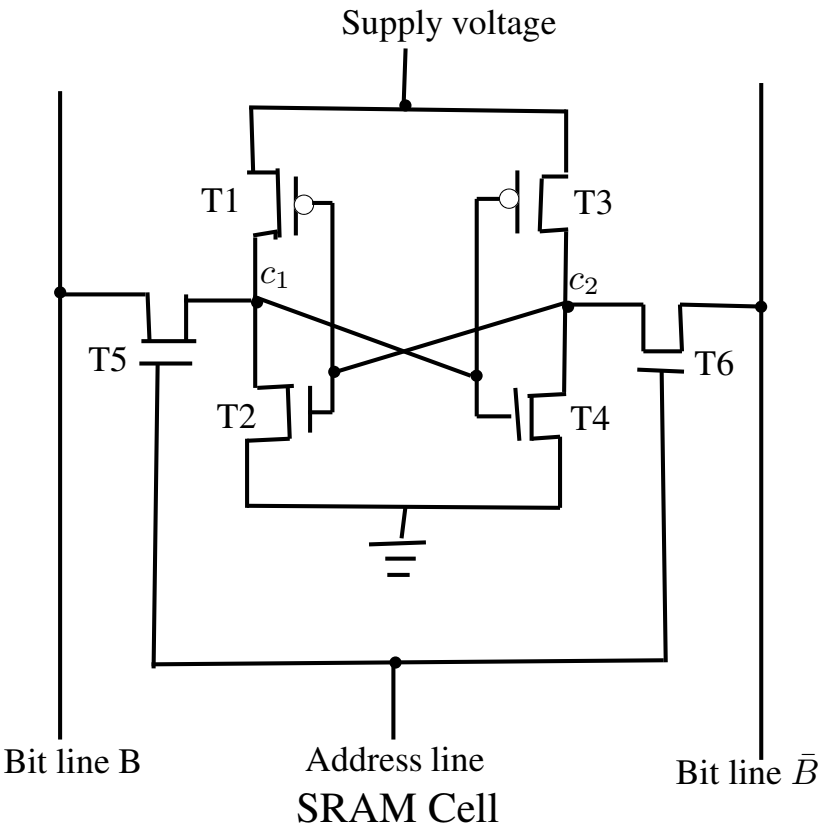


Figure 7: 1-bit DRAM Cell.

Over a time (of msecs.) the capacitor holding 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).

Static RAM (SRAM) Cell



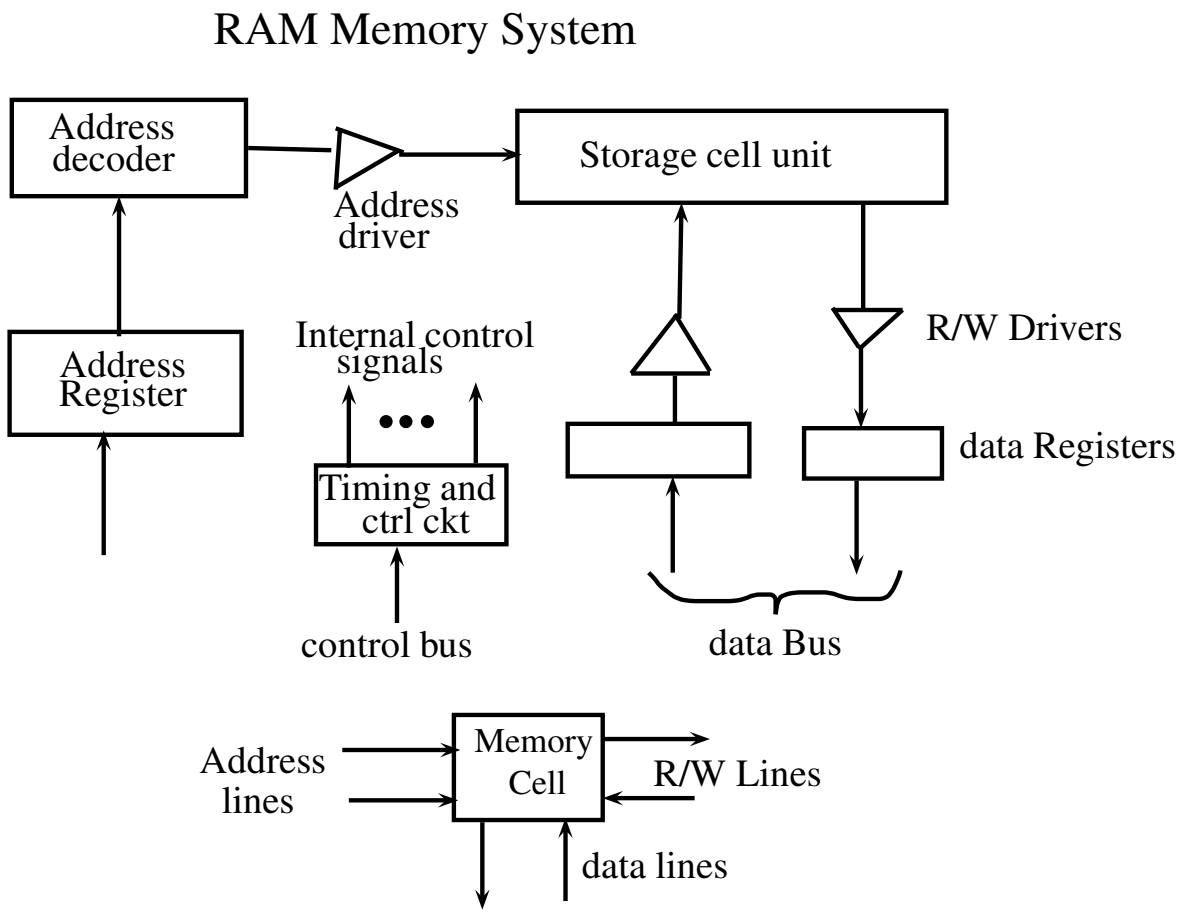
SRAM Cell

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

SRAM v/s DRAM

- Both volatile
- 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, and DRAMs are used for RAM (Main Memory).

Components of RAM



- Address drivers and R/W drivers are current drivers
- Single data register for non-parallel operation.

On/Two dimensional RAM

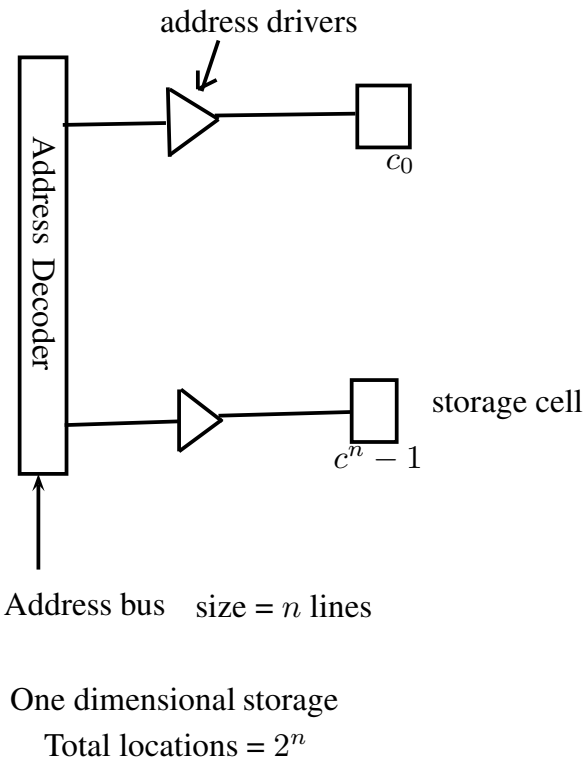


Figure 8: One Dimensional RAM

IF no. of address lines are n, then total 2^n words can be stored.

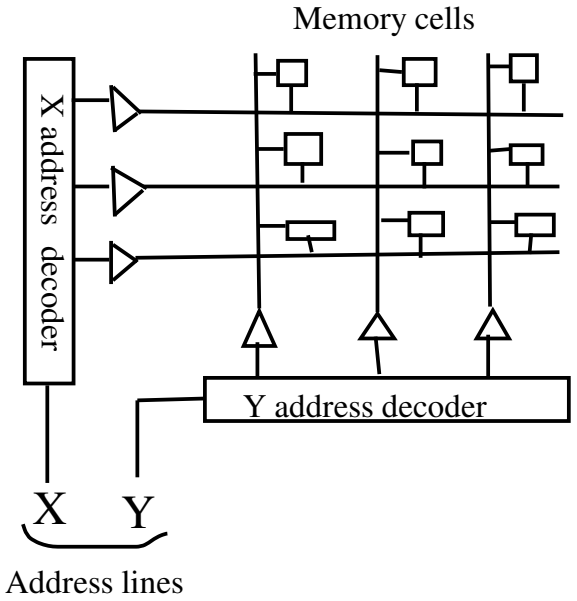


Figure 9: Two Dimensional RAM

IF no. of address lines are $x + y$, for horizontals and vertical address decoders then total $2^x \times 2^y$ words can be stored.

DRAM with refreshing

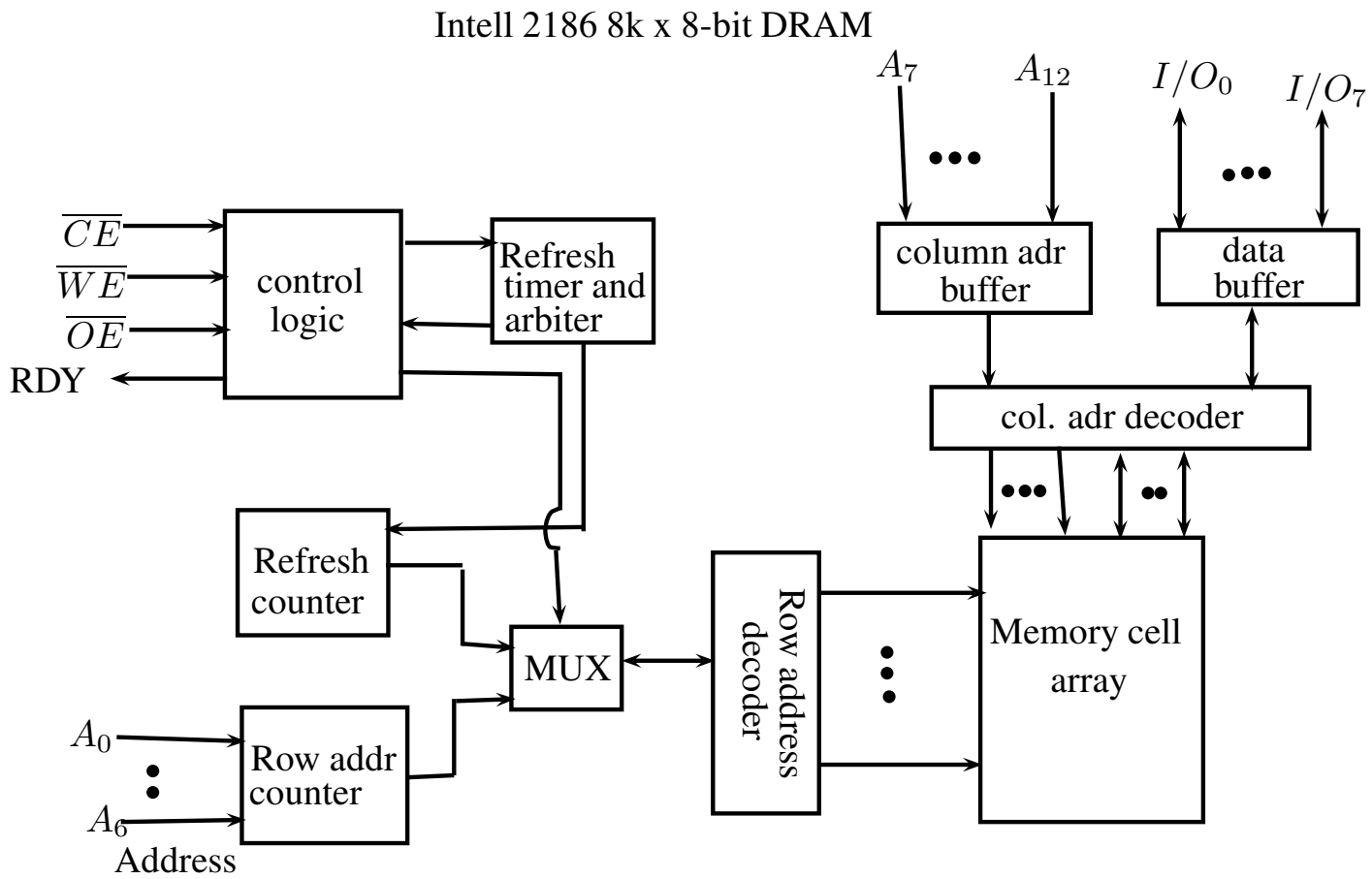


Figure 10: Two Dimensional RAM

Refreshing is interleaved with memory R/W operations.

DRAM with refreshing

- All cells are refreshed once every 2 msec. 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. $= 64\ \mu$ sec. Since refresh is interleaved with R/W, max. delay in any R/W is only 500 nsec. The delay is taken care of by RDY line.

- Since all cells are refreshed in 2 msec., the fraction time is only $\frac{64\mu \text{ sec}}{2msec} \simeq 3.2$ percent.
- An arbiter selects the row address 0-127.
- 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).

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.
- 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 $|w|$ tracks, reading 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 if rN words per sec. Once the head is positioned, a block can be transferred in $n(rN)^{-1}$ secs. The appox time to transfer a block: $t_B = t_s + t_l + n(rN)^{-1}$.

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.

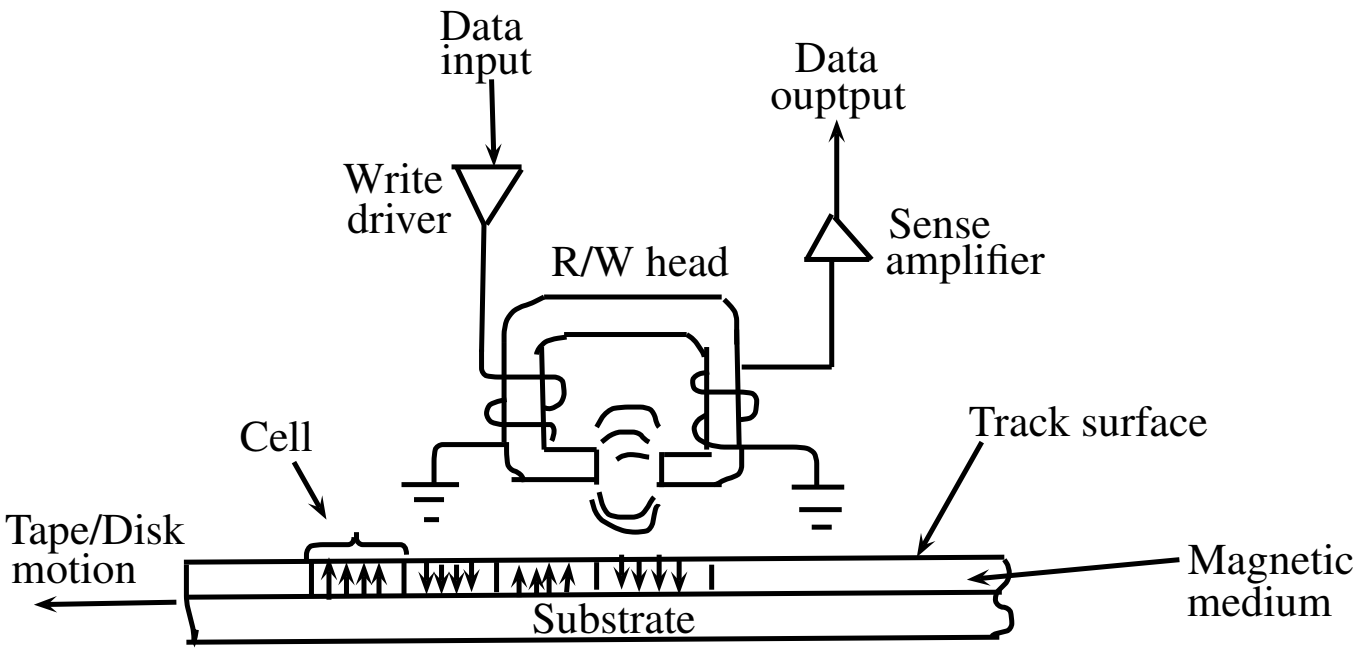


Figure 11: Magnetic surface recording.

Magnetic disk read-write

- Several 100s tracks arranged as concentric circles.
- Several disks attached to a common spindle, rotated at constant speeds (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).

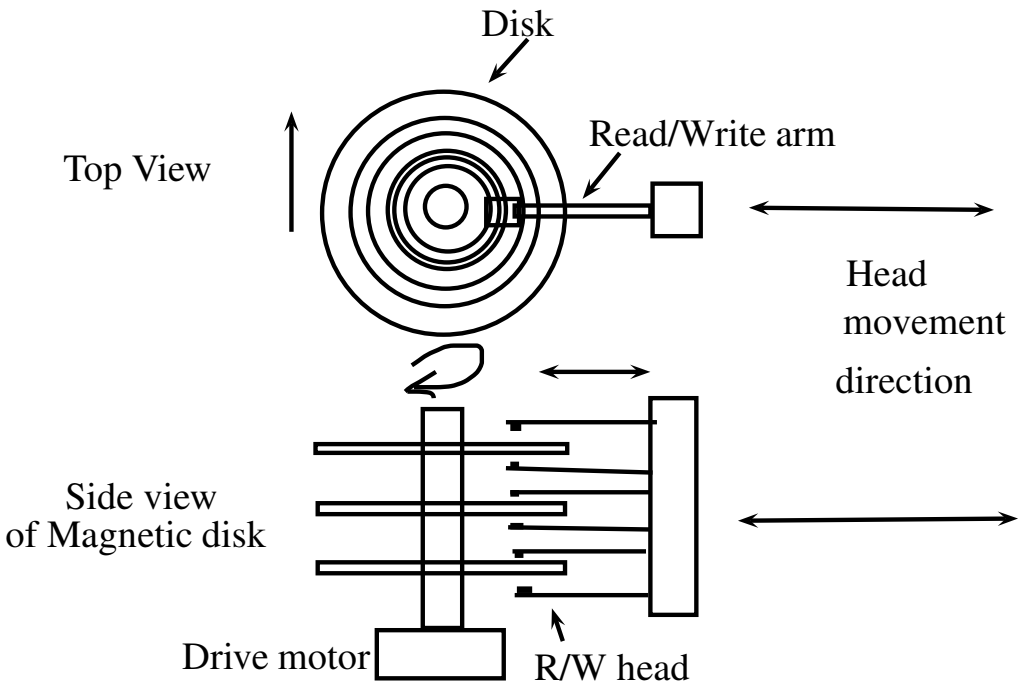


Figure 12: Magnetic Disk storage system.

Magnetic disk memories

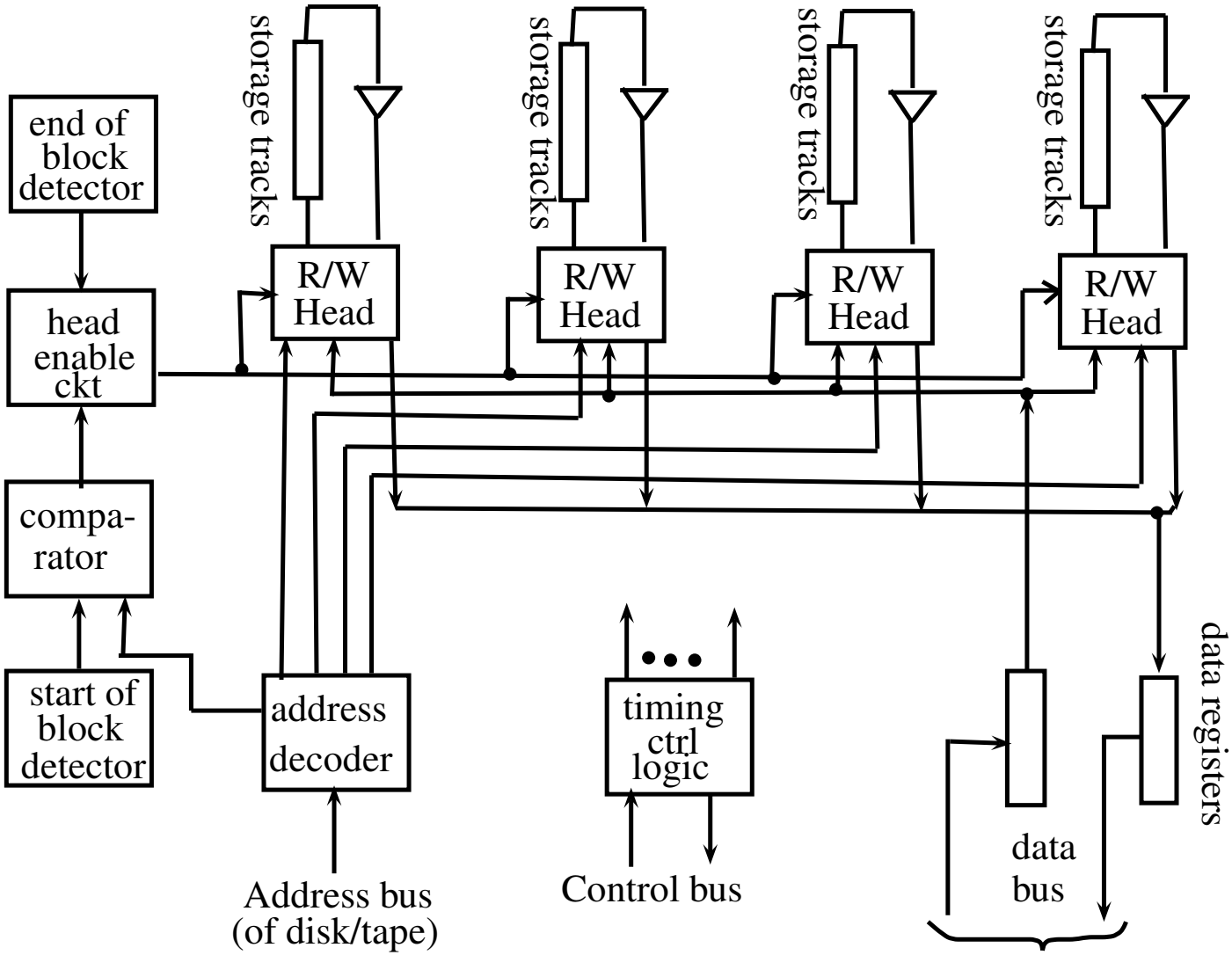


Figure 13: Magnetic Disk memory W/R System with 4-heads

Magnetic disk memory system

- 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 steps of Disk R/W are:
  1. Track address determines particular R/W head
  2. The head is moved on to the track.
  3. Some track position indicator is needed, which is generated when the track passes under the R/W head.
  4. The generated address is compared with the block address produced by the address decoder.
  5. When they match, the selected head is enabled and data transfer begins.
  6. 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.

Magnetic disk memory typical data:

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

Magnetic tape memories

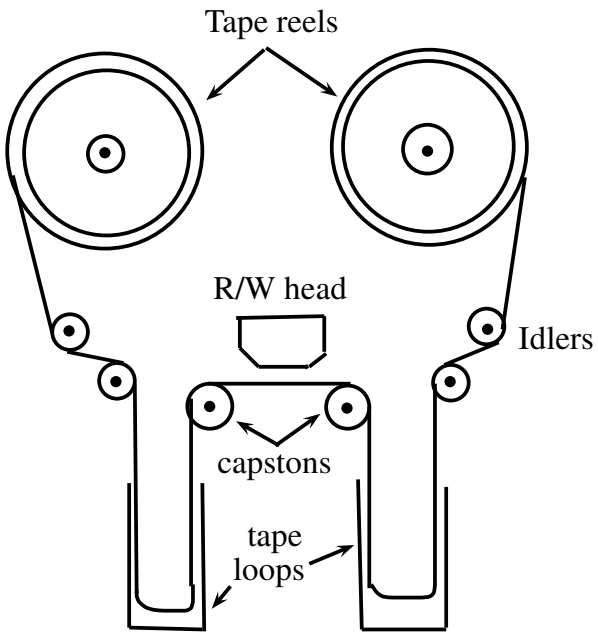


Figure 14: Magnetic tape storage.

- 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 info+ 1 parity bit).
- Packed in cartridges/cassettes; resemble audio tapes. Medium is not in continuous motion. When request recved, tape is moved forward/ backward to desired location.
- Loops 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 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. 1.1 How many 128×8 RAM chips are required to provide a memory capacity of 2048 bytes?
   1.2 How many address lines are required for accessing 2048 bytes? How many chips are common for all the chips?
   1.3 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.
   4.1 What is the time period between successive refresh requests?
   4.2 How long a refresh address counter do we need?

Exercises

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.
   - 6.1 What must the linear tape-recording density be in order to achieve a data-transfer rate of 10^7 bits/s?
   - 6.2 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?

Exercises

7. Figure 15 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.
   - 7.1 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?
   - 7.2 Constructing a 32-bit wide memory system using these chips yields what data transfer? (RAS=row addr select, CAS=col. addr select)

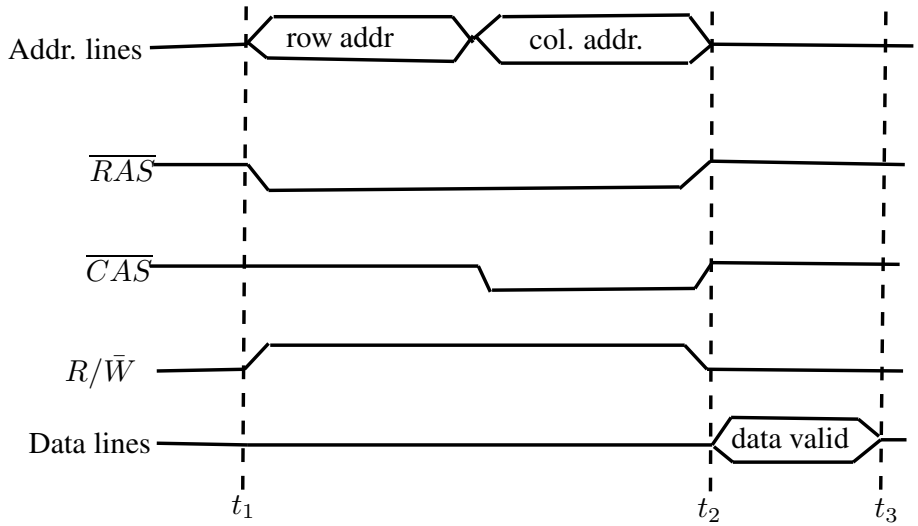


Figure 15: Timing diagram of DRAM.

Exercises

8. The memory of a particular microcomputer is built from 128K $\times$ 1 DRAMs. According to the data sheet, the cell array of DRAM is organized into 512 rows. Each row must be refreshed atleast once every 4 ms. Suppose we refresh memory on a strictly periodic basis:
   - 8.1 What is the time period between successive refresh requests?
   - 8.2 How long a refresh address counter do we need?
9. Design a 16 X 4-bits RAM using 4X2-bits ICs.
10. Design a 16-bit memory of total capacity 8192 bits using SRAM chips of 64 $\times$ 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


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


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