Concept of Memory

We have already mentioned that digital computer works on stored programmed concept introduced by Von Neumann. We use memory to store the information, which includes both program and data.

Due to several reasons, we have different kind of memories. We use different kind of memory at different level.

The memory of computer is broadly categories into two categories:

- Internal and
- external

Internal memory is used by CPU to perform task and external memory is used to store bulk information, which includes large software and data.

Memory is used to store the information in digital form. The memory hierarchy is given by:

- Register
- Cache Memory
- Main Memory
- Magnetic Disk
- Removable media (Magnetic tape)

- **Register:**
  This is a part of Central Processor Unit, so they reside inside the CPU. The information from main memory is brought to CPU and keep the information in register. Due to space and cost constraints, we have got a limited number of registers in a CPU. These are basically faster devices.

- **Cache Memory:**
  Cache memory is a storage device placed in between CPU and main memory. These are semiconductor memories. These are basically fast memory device, faster than main memory.
  We cannot have a big volume of cache memory due to its higher cost and some constraints of the CPU. Due to higher cost we cannot replace the whole main memory by faster memory. Generally, the most recently used information is kept in the cache memory. It is brought from the main memory and placed in the cache memory. Now a days, we get CPU with internal cache.

- **Main Memory:**
  Like cache memory, main memory is also semiconductor memory. But the main memory is relatively slower memory. We have to first bring the information (whether it is data or program), to main memory. CPU can work with the information available in main memory only.

- **Magnetic Disk:**
  This is bulk storage device. We have to deal with huge amount of data in many application. But we don't have so much semiconductor memory to keep these information in our computer. On the other hand, semiconductor memories are volatile in nature. It loses its content once we switch off the computer. For permanent storage, we use magnetic disk. The storage capacity of magnetic disk is very high.

- **Removable media:**
  For different application, we use different data. It may not be possible to keep all the information in magnetic disk. So, which ever data we are not using currently, can be kept in removable media. Magnetic tape is one kind of removable medium. CD is also a removable media, which is an optical device.

Register, cache memory and main memory are internal memory. Magnetic Disk, removable media are external memory. Internal memories are semiconductor memory. Semiconductor memories are categorized as volatile memory and non-volatile memory.

RAM: Random Access Memories are **volatile** in nature. As soon as the computer is switched off, the contents of memory are also lost.

ROM: Read only memories are **non volatile** in nature. The storage is permanent, but it is read only memory. We cannot store new information in ROM.

Several types of ROM are available:

- **PROM:** Programmable Read Only Memory; it can be programmed once as per user requirements.

- **EPROM:** Erasable Programmable Read Only Memory; the contents of the memory can be erased and store new data into the memory. In this case, we have to erase whole information.

- **EEPROM:** Electrically Erasable Programmable Read Only Memory; in this type of memory the contents of a particular location can be changed without effecting the contents of other location.

Main Memory

The main memory of a computer is semiconductor memory. The main memory unit of computer is basically consists of two kinds of memory:

RAM: Random access memory; which is volatile in nature.
ROM: Read only memory; which is non-volatile.

The permanent information are kept in ROM and the user space is basically in RAM.

The smallest unit of information is known as bit (binary digit), and in one memory cell we can store one bit of information. 8 bit together is termed as a byte.

The maximum size of main memory that can be used in any computer is determined by the addressing scheme.

A computer that generates 16-bit address is capable of addressing upto 2^{16} which is equal to 64K memory location. Similarly, for 32 bit addresses, the total capacity will be 2^{32} which is equal to 4G memory location.

In some computer, the smallest addressable unit of information is a memory word and the machine is called word-addressable.

In some computer, individual address is assigned for each byte of information, and it is called **byte-addressable computer**. In this computer, one memory word contains one or more memory bytes which can be addressed individually.

A byte addressable 32-bit computer, each memory word contains 4 bytes. A possible way of address assignment is shown in figure3.1. The address of a word is always integer multiple of 4.

The main memory is usually designed to store and retrieve data in word length quantities. The word length of a computer is generally defined by the number of bits actually stored or retrieved in one main memory access.

Consider a machine with 32 bit address bus. If the word size is 32 bit, then the high order 30 bit will specify the address of a word. If we want to access any byte of the word, then it can be specified by the lower two bit of the address bus.

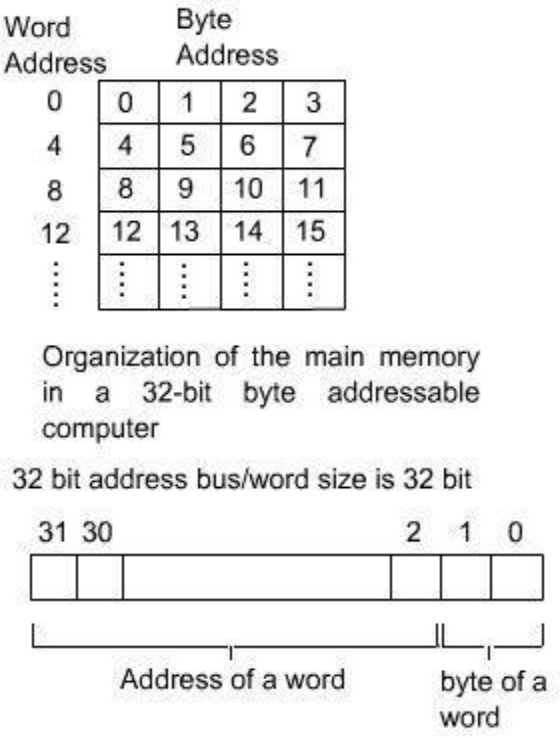


Figure 3.1: Address assignment to a 4-byte word

The data transfer between main memory and the CPU takes place through two CPU registers.

- **MAR:** Memory Address Register
- **MDR:** Memory Data Register.

If the MAR is *k*-bit long, then the total addressable memory location will be 2^k.

If the MDR is *n*-bit long, then the *n* bit of data is transferred in one memory cycle.

The transfer of data takes place through memory bus, which consist of address bus and data bus. In the above example, size of data bus is *n*-bit and size of address bus is *k* bit.

It also includes control lines like Read, Write and Memory Function Complete (MFC) for coordinating data transfer. In the case of byte addressable computer, another control line to be added to indicate the byte transfer instead of the whole word.

For memory operation, the CPU initiates a memory operation by loading the appropriate data i.e., address to MAR.

If it is a memory read operation, then it sets the read memory control line to 1. Then the contents of the memory location is brought to MDR and the memory control circuitry indicates this to the CPU by setting MFC to 1.

If the operation is a memory write operation, then the CPU places the data into MDR and sets the write memory control line to 1. Once the contents of MDR are stored in specified memory location, then the memory control circuitry indicates the end of operation by setting MFC to 1.

A useful measure of the speed of memory unit is the time that elapses between the initiation of an operation and the completion of the operation (for example, the time between Read and MFC). This is referred to as **Memory Access Time**. Another measure is memory cycle time. This is the minimum time delay between the initiation two independent memory operations (for example, two successive memory read operation). Memory cycle time is slightly larger than memory access time.

Binary Storage Cell:

The binary storage cell is the basic building block of a memory unit.

The binary storage cell that stores one bit of information can be modelled by an SR latch with associated gates. This model of binary storage cell is shown in the figure 3.2.

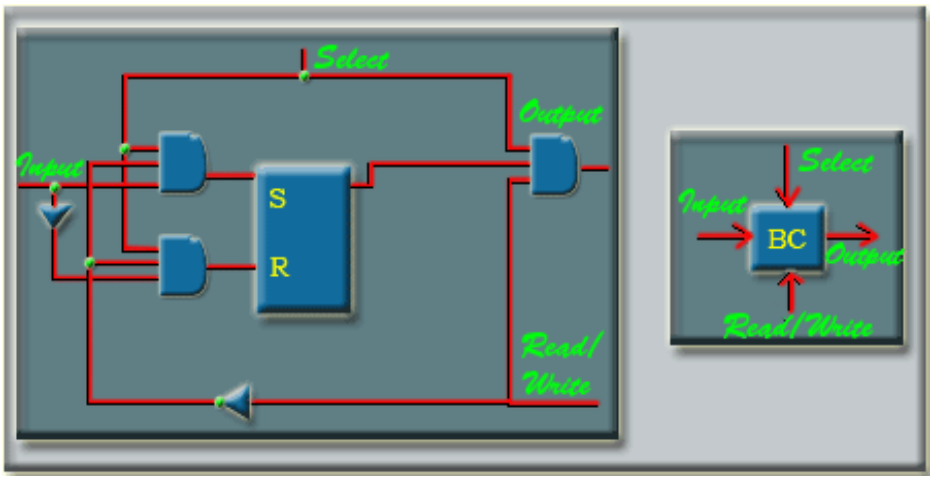


Figure 3.2: Binary Storage cell made up of SR-Latch

1 bit Binary Cell (BC)

The binary cell stores one bit of information in its internal latch.

Control input to binary cell

Select	Read/Write	Memory Operation
0	X	None
1	0	Write
1	1	Read

The storage part is modeled here with SR-latch, but in reality it is an electronics circuit made up of transistors.

The memory constructed with the help of transistors is known as semiconductor memory. The semiconductor memories are termed as Random Access Memory (RAM), because it is possible to access any memory location in random.
Depending on the technology used to construct a RAM, there are two types of RAM -

SRAM: Static Random Access Memory.
DRAM: Dynamic Random Access Memory

Dynamic Ram (DRAM):

A DRAM is made with cells that store data as charge on capacitors. The presence or absence of charge in a capacitor is interpreted as binary 1 or 0.

Because capacitors have a natural tendency to discharge due to leakage current, dynamic RAM require periodic charge refreshing to maintain data storage. The term dynamic refers to this tendency of the stored charge to leak away, even with power continuously applied.

A typical DRAM structure for an individual cell that stores one bit information is shown in the figure 3.3.

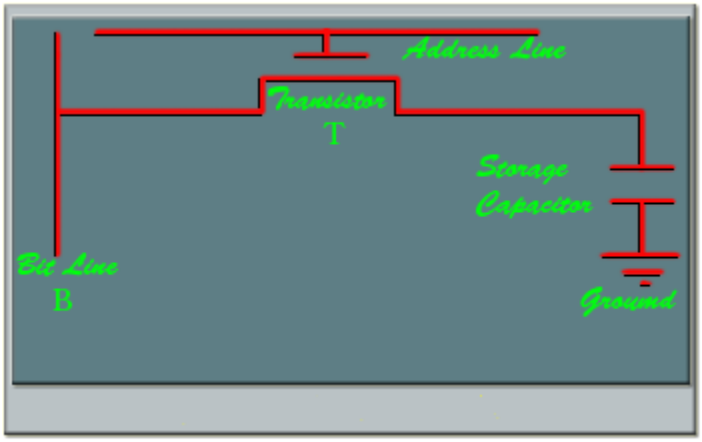


Figure 3.3: Dynamic RAM (DRAM) cell

For the write operation, a voltage signal is applied to the bit line B, a high voltage represents 1 and a low voltage represents 0. A signal is then applied to the address line, which will turn on the transistor T, allowing a charge to be transferred to the capacitor.

For the read operation, when a signal is applied to the address line, the transistor T turns on and the charge stored on the capacitor is fed out onto the bit line B and to a sense amplifier.

The sense amplifier compares the capacitor voltage to a reference value and determines if the cell contains a logic 1 or a logic 0.

The read out from the cell discharges the capacitor, widh must be restored to complete the read operation.

Due to the discharge of the capacitor during read operation, the read operation of DRAM is termed as destructive read out.

Static RAM (SRAM):

In an SRAM, binary values are stored using traditional flip-flop constructed with the help of transistors. A static RAM will hold its data as long as power is supplied to it.

A typical SRAM constructed with transistors is shown in the figure 3.4.

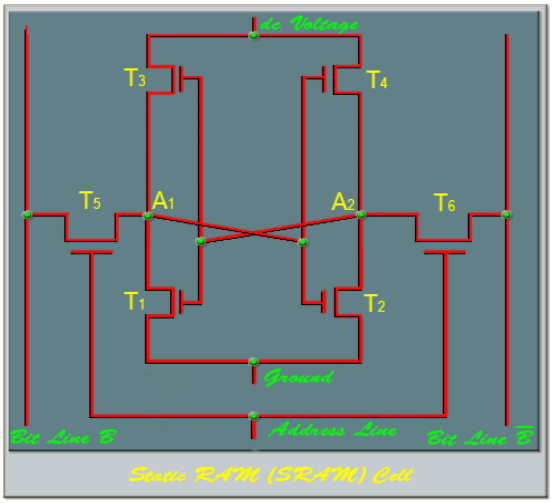


Figure 3.4: Static RAM (SRAM) cell

Four transistors (T_1, T_2, T_3, T_4) are cross connected in an arrangement that produces a stable logic state. In logic state 1, point A_1 is high and point A_2 is low; in this state T_1 and T_4 are off, and T_2 and T_3 are on. In logic state 0, point A_1 is low and point A_2 is high; in this state T_1 and T_4 are on, and T_2 and T_3 are off. Both states are stable as long as the dc supply voltage is applied.

The address line is used to open or close a switch which is nothing but another transistor. The address line controls two transistors (T_5 and T_6). When a signal is applied to this line, the two transistors are switched on, allowing a read or write operation.

For a write operation, the desired bit value is applied to line B, and its complement is applied to line $\overline{B}$. This forces the four transistors (T_1, T_2, T_3, T_4) into the proper state.

For a read operation, the bit value is read from the line B. When a signal is applied to the address line, the signal of point A_1 is available in the bit line B.

SRAM Versus DRAM :

- Both static and dynamic RAMs are volatile, that is, it will retain the information as long as power supply is applied.
- A dynamic memory cell is simpler and smaller than a static memory cell. Thus a DRAM is more dense, i.e., packing density is high (more cell per unit area). DRAM is less expensive than corresponding SRAM.

- DRAM requires the supporting refresh circuitry. For larger memories, the fixed cost of the refresh circuitry is more than compensated for by the less cost of DRAM cells
- SRAM cells are generally faster than the DRAM cells. Therefore, to construct faster memory modules (like cache memory) SRAM is used.

Internal Organization of Memory Chips

A memory cell is capable of storing 1-bit of information. A number of memory cells are organized in the form of a matrix to form the memory chip. One such organization is shown in the Figure 3.5.

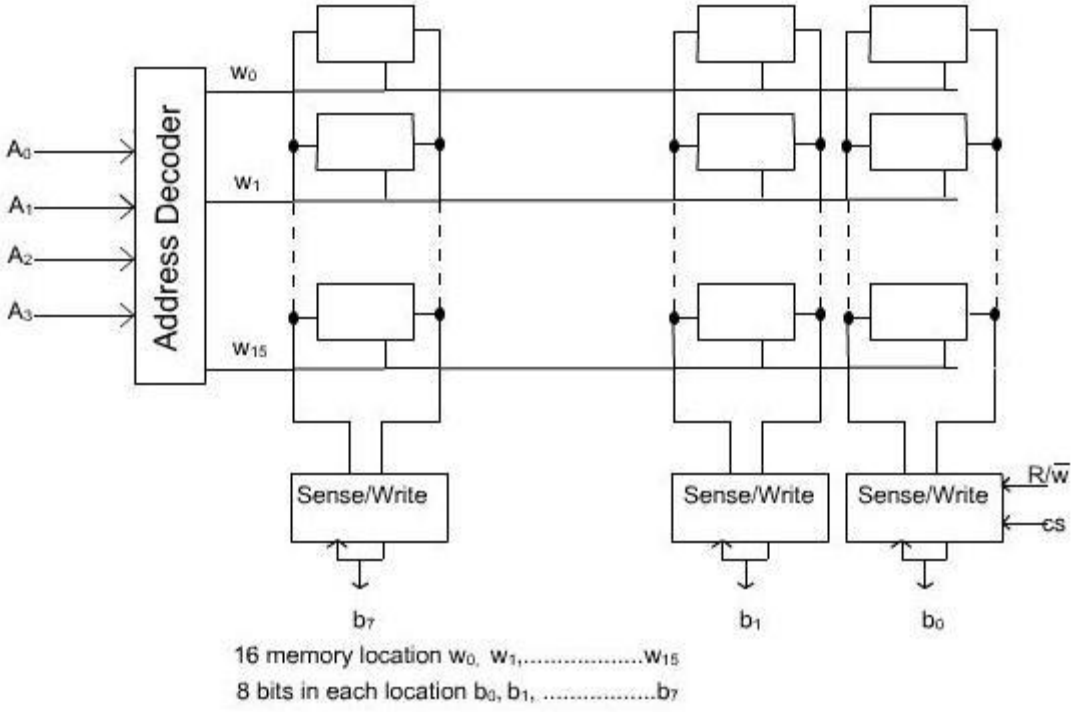


Figure 3.5: 16 X 8 Memory Organization

Each row of cells constitutes a memory word, and all cell of a row are connected to a common line which is referred as word line. An address decoder is used to drive the word line. At a particular instant, one word line is enabled depending on the address present in the address bus. The cells in each column are connected by two lines. These are known as bit lines. These bit lines are connected to data input line and data output line through a Sense/Write circuit. During a Read operation, the Sense/Write circuit sense, or read the information stored in the cells selected by a word line and transmit this information to the output data line. During a write operation, the sense/write circuit receive information and store it in the cells of the selected word.

A memory chip consisting of 16 words of 8 bits each, usually referred to as 16 x 8 organization. The data input and data output line of each Sense/Write circuit are connected to a single bidirectional data line in order to reduce the pin required. For 16 words, we need an address bus of size 4. In addition to address and data lines, two control lines, $R/\overline{W}$ and CS, are provided. The $R/\overline{W}$ line is to used to specify the required operation about read or write. The CS (Chip Select) line is required to select a given chip in a multi chip memory system.

Consider a slightly larger memory unit that has 1K (1024) memory cells...

128 x 8 memory chips:

If it is organized as a 128 x 8 memory chips, then it has got 128 memory words of size 8 bits. So the size of data bus is 8 bits and the size of address bus is 7 bits ($2^7 = 128$). The storage organization of 128 x 8 memory chip is shown in the figure 3.6.

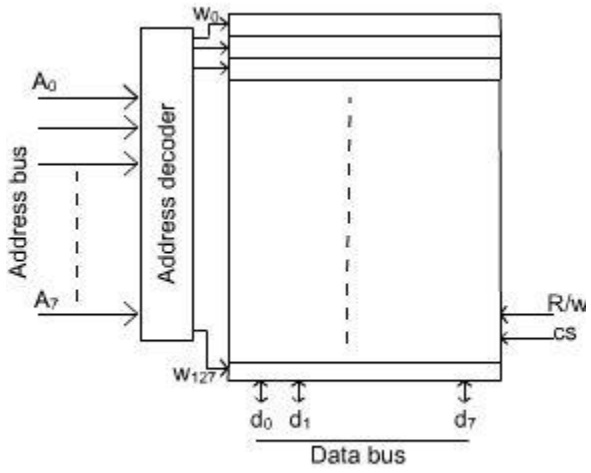


Figure 3.6: 128 x 8 Memory Chip

1024 x 1 memory chips:

If it is organized as a 1024 x 1 memory chips, then it has got 1024 memory words of size 1 bit only.

Therefore, the size of data bus is 1 bit and the size of address bus is 10 bits ($2^{10} = 1024$).

A particular memory location is identified by the contents of memory address bus. A decoder is used to decode the memory address. There are two ways of decoding of a memory address depending upon the organization of the memory module.

In one case, each memory word is organized in a row. In this case whole memory address bus is used together to decode the address of the specified location. The memory organization of 1024 x 1 memory chip is shown in the figure 3.7.

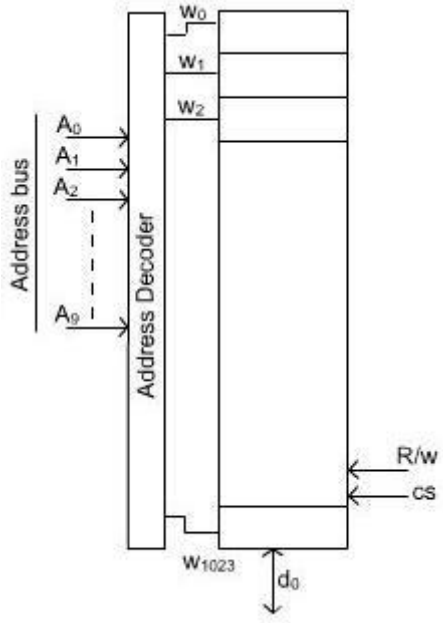


Figure 3.7: 1024 x 1 Memory chip

In second case, several memory words are organized in one row. In this case, address bus is divided into two groups.

One group is used to form the row address and the second group is used to form the column address. Consider the memory organization of 1024 x 1 memory chip. The required 10-bit address is divided into two groups of 5 bits each to form the row and column address of the cell array. A row address selects a row of 32 cells, all of which are accessed in parallel. However, according to the column address, only one of these cells is connected to the external data line via the input output multiplexers. The arrangement for row address and column address decoders is shown in the figure 3.8.

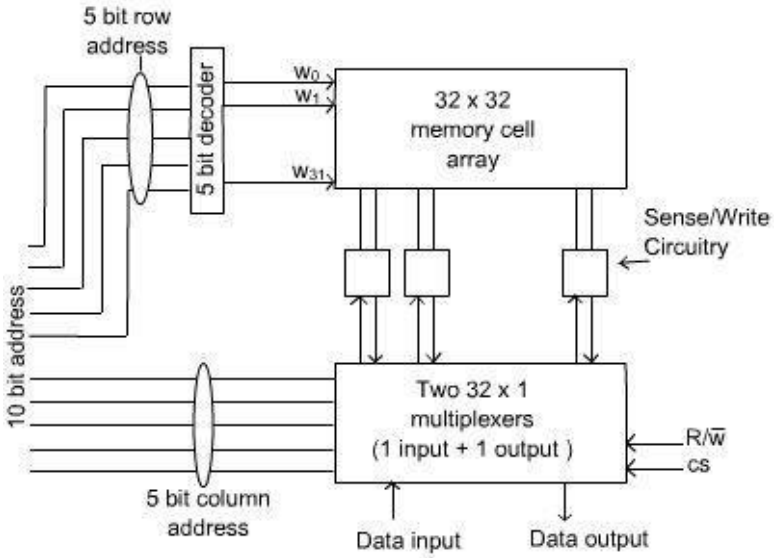


Figure 3.8: Organization of 1k x 1 Memory chip

The commercially available memory chips contain a much larger number of cells. As for example, a memory unit of 1MB (mega byte) size, organized as 1M x 8, contains $2^{20} \times 8$ memory cells. It has got 2^{20} memory location and each memory location contains 8 bits information. The size of address bus is 20 and the size of data bus is 8.

The number of pins of a memory chip depends on the data bus and address bus of the memory module. To reduce the number of pins required for the chip, we use another scheme for address decoding. The cells are organized in the form of a square array. The address bus is divided into two groups, one for column address and other one is for row address. In this case, high- and low-order 10 bits of 20-bit address constitute of row and column address of a given cell, respectively. In order to reduce the number of pin needed for external connections, the row and column addresses are multiplexed on ten pins. During a Read or a Write operation, the row address is applied first. In response to a signal pulse on the **Row Address Strobe (RAS)** input of the chip, this part of the address is loaded into the row address latch. All cell of this particular row is selected. Shortly after the row address is latched, the column address is applied to the address pins. It is loaded into the column address latch with the help of

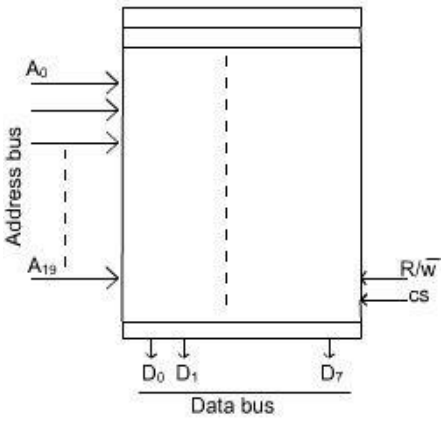
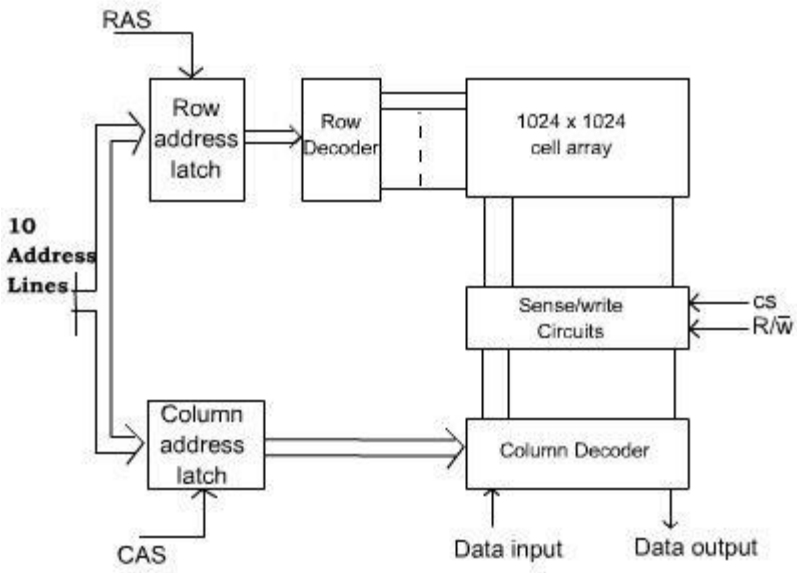


Figure 3.9: 1 MB(Mega Byte) Memory Chip

Column Address Strobe (CAS) signal, similar to RAS. The information in this latch is decoded and the appropriate **Sense/Write** circuit is selected.

For a Write operation, the information at the input lines are transferred to the selected circuits.



The 1MB (Mega byte) memory chip with 20 address lines as shown in the figure 3.9. The same memory chip (1MB) with 10 address lines (where row & column address are multiplexed) is shown in Figure 3.10.

Figure 3.10: Organization of a 1M x 1 Memory chip.

Now we discuss the design of memory subsystem using memory chips. Consider memory chips of capacity 16K x 8. The requirement is to design a memory subsystem of capacity 64K x 16. Each memory chip has got eight lines for data bus, but the data bus size of memory subsystem is 16 bits. The total requirement is for 64K memory location, so four such units are required to get the 64K memory location. For 64K memory location, the size of address bus is 16. On the other hand, for 16K memory location, size of address bus is 14 bits. Each chip has a control input line called Chip Select (CS). A chip can be enabled to accept data input or to place the data on the output bus by setting its Chip Select input to 1. The address bus for the 64K memory is 16 bits wide. The high order two bits of the address are decoded to obtain the four chip select control signals. The remaining 14 address bits are connected to the address lines of all the chips. They are used to access a specific location inside each chip of the selected row. The $R/\overline{W}$ inputs of all chips are tied

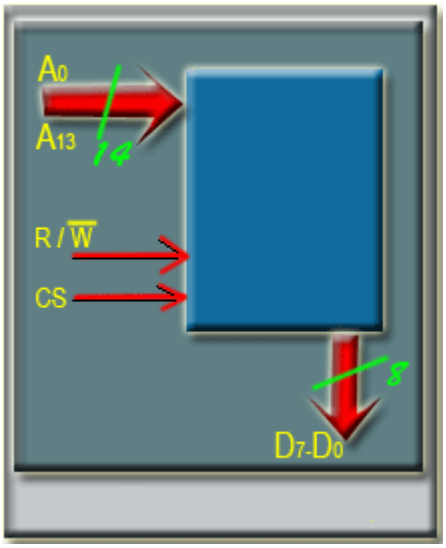


Figure 3.11: 16k x 8 Memory chip

together to provide a common $READ/\overline{WRITE}$ control.

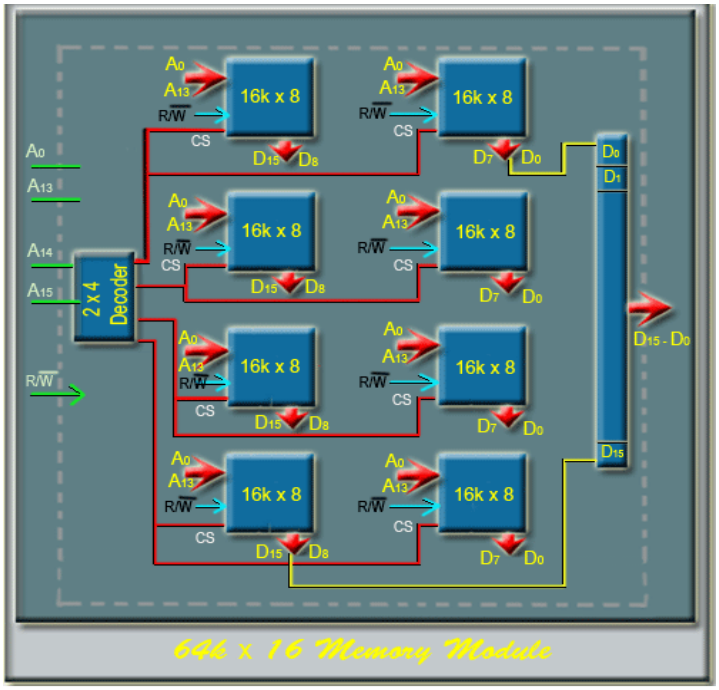


Figure 3.12: 64k x 16 Memory chip

The block diagram of a *16k x 8* memory chip is shown in the figure 3.11. The block diagram of a *64k x 16* memory module constructed with the help of eight *16k x 8*memory chips is shown in the figure 3.12.