**EXPERIMENT NO. 06**

**DATE OF PERFORMANCE: GRADE:**

**DATE OF ASSESSMENT: SIGNATURE OF LECTURER/TTA:**

**Aim: Introduction of Memory Organization.**

**THEORY:**

**Main Memory in details.**

The main memory is the central storage unit in a computer system. It is a relatively large and fast memory used to store programs and data during the computer operation.

Integrated circuit RAM chips are available in two possible modes: static and dynamic.

The static RAM consists essentially of internal flip-flops. The stored information remains valid as long as power is applied to the unit.

The dynamic RAM stores the binary information in the form of electric charges that are applied to capacitors. The capacitors are provided inside the chip by MOS transistors. The stored charge on the capacitor tends to discharge with time and the capacitors must be periodically recharged by refreshing the dynamic memory.

The dynamic RAM offers reduced power consumption and larger storage capacity in a single memory chip.

Most of the main memory in a general-purpose computer is made up of RAM integrated circuit chips, but a portion of the memory may be constructed with ROM chips.

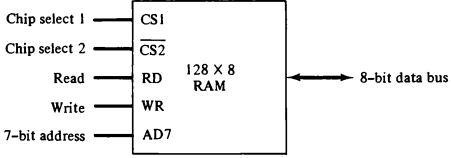
RAM is used for storing the bulk of the programs and data that are subject to change. ROM is used for storing programs that are permanently resident in the computer and for tables of constants that do not change in the value once the production of the computer is completed.

Among other things, the ROM portion of main memory is needed for storing an initial program called as bootstrap loader. The bootstrap loader is a program whose function is to start the computer software operating when power is turned on.

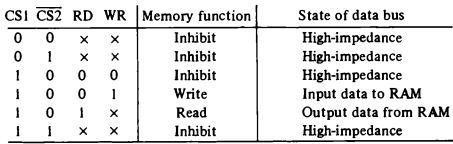
Since RAM is volatile, its contents are destroyed when power is turned off. The contents of ROM remain unchanged after power is turned off and on again.

When power is turned on, the hardware of the computer sets the program counter to the first address of the bootstrap loader which loads a portion of the OS from disk to main memory and control is then transferred to the OS, which prepares the computer for general use.

**A RAM chip** is better suited to communicate with CPU if it has one or more control inputs that select the chip only when needed. The block diagram of a RAM chip is shown below:

 Fig:Typical RAM chip (128 words of eight bits each)

* Requires 7-bit address and an 8-bit bidirectional data bus
* Chip select (CS) control inputs are for enabling the chip only when it is selected by CPU.



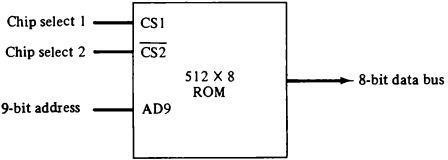
* The unit is in operation only

when CS1=1 and (CS2)’=0.

* High impedance state indicates open circuit i.e. output does not carry a signal and has no logic significance.

Fig: Function table for RAM chip

**ROM Chips**

Since a ROM chip can only read, data bus is unidirectional (output mode only).

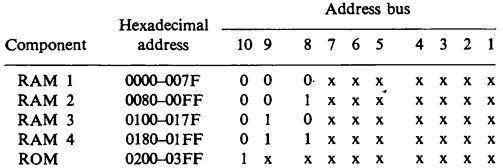
* 9 address lines to address 512 bytes
* Two chip select (CS) inputs CS1=1 and (CS2)’=0 for the unit to operate, otherwise the data bus is in high-impedance state.
* No need for read or write control since the unit can only read.

Fig: Typical ROM chip (512 byte ROM)

**Memory Address Map**

The addressing of memory can be established by means of a table that specifies the memory address assigned to each RAM or ROM chip. This table is called memory address map and is a pictorial representation of assigned address space for particular chip.

Example: Suppose computer system needs 512 bytes of RAM and 512 bytes of ROM.



* Component column specifies RAM or ROM chip. We use four 128 words RAM to make 512 byte size.
* Hexadecimal address column assigns a range of addresses for each chip.
* 10 lines in address bus column: lines 1 through 7 for RAM and 1 through 9 for ROM. Distinction between RAM and ROM chip is made by line 10. When line 10 is 1, it selects ROM and when it is 0, CPU selects RAM.
* X’s represents a binary number ranging from all-0’s to all-1’s.

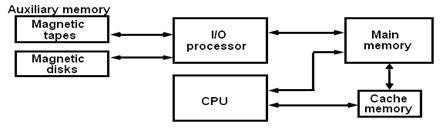
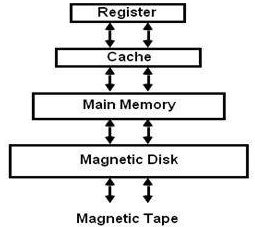
**Memory Hierarchy in details**

Memory unit is an essential component in any digital computer since it is needed for storing programs and data.

The memory unit that communicates directly with the CPU is called the main memory. Devices that provide backup storage are called auxiliary memory.

Only programs and data currently needed+++- by the processor reside in main memory. All other information is stored in auxiliary memory and transferred to main memory when needed.

The memory hierarchy consists of all storage devices employed in a computer system from the slow but high-capacity auxiliary memory to a relatively faster main memory, to an even smaller and faster cache memory accessible to the high-speed processing logic.

The main memory occupies a central position by being able to communicate directly with the CPU and with auxiliary memory devices through an I/O processor,

The auxiliary memory has a large storage capacity, is relatively inexpensive, but has low access speed compared to main memory.

The cache memory is very small, relatively very small, relatively expensive, and has very high access speed.

CPU has direct access to both cache and main memory but not to auxiliary memory. The transfer from auxiliary to main memory is usually done by means of direct memory access of large blocks of data.

The part of the computer system that supervises the flow of information between auxiliary memory and main memory is called **memory management system.**

**Cache and Cache Mapping Techniques**

A cache is a small amount of very fast associative memory.

It sits between normal main memory and CPU.

When CPU needs to access contents of memory location, then cache is examined for this data.

If present, get from cache (fast).

If not present, read required block from main memory to cache, then deliver from cache to CPU.

Cache includes tags to identify which block of main memory is in each cache slot.

The performance of cache memory is frequently measured in terms of a quantity called hit ratio. When CPU refers to memory and finds the word in cache, then it is said to produce hit. If word is not found in cache, it is in main memory and it counts as a miss. The ratio of the number of hits (success in finding the words in cache) to the total CPU references to memory (hits + misses) is known as hit ratio.

The basic characteristic of cache memory is its fast access time.

**Cache Mapping**

The process of transferring the data from main memory to cache is known as mapping process. There are three types of cache mapping techniques:

Associative mapping

Direct mapping

Set-associative mapping

The main memory can store 32K words of 12 bits each. The cache is capable of storing 512 of these words at any given time. For every word stored in cache, there is a duplicate copy in main memory. The CPU communicates with both memories. It first sends a 15-bit address to cache. If there is a hit, the CPU accepts the 12-bit data from cache. If there is a miss, the CPU reads the word from main memory and the word is then transferred to cache.

#### 

#### Direct Mapping

##### The tag field of the CPU address is compared with the tag in the word read from the cache. If the two tags match, there is a hit and the desired data word is in cache. If there is no match, there is a miss and the required word is read from main memory. It is then stored in the cache together with the new tag, replacing the previous value.

The disadvantage of direct mapping is that two words with the same index in their address but with different tag values cannot reside in cache memory at the same time.

**Set-Associative Mapping**

### It is an improvement over the direct mapping in that each word of cache can store two or more words of memory under the same index address.

Each data word is stored together with its tag and the number of tag-data items in one word of cache is said to form a set.

Each index address refers to two data words and their associated tags. Each tag requires six bits and each data word has 12 bits, so the word length is 2(6 + 12) = 36 bits.

An index address of nine bits can accommodate 512 words. Thus, the size of cache memory is 512\*36. It can accommodate 1024 words of main memory since each word of cache contains two data words.

**Fully-Associative Mapping**

It is the combination of Direct Mapping and Set-associative mapping.

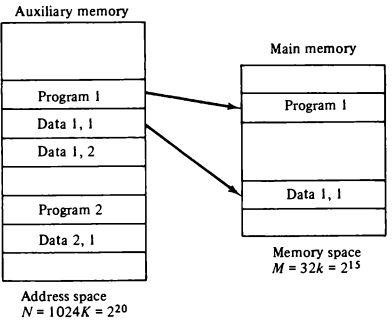
#### Virtual Memory

**Address Space and Memory Space**

An address used by the programmer is a virtual address (virtual memory addresses) and the set of such addresses is the Address Space.

An address in main memory is called a location or physical address. The set of such locations is called the Memory Space. Thus the address space is the set of addresses generated by the programs as they reference instructions and data; the memory space consists of actual main memory locations directly addressable for processing. Generally, the address space is larger than the memory space.

Example: consider main memory: 32K words (K = 1024) = 215 and auxiliary memory 1024K words = 220. Thus we need 15 bits to address physical memory and 20 bits for virtual memory (virtual memory can be as large as we have auxiliary storage).

 Fig: Relation between address and memory space in a virtual memory system

* Here auxiliary memory has the capacity of storing information equivalent to 32 main memories.
* Address space N = 1024K
* Memory space M = 32K
* In multi-program computer system, programs and data are transferred to and from auxiliary memory and main memory based on the demands imposed by the CPU.

In virtual memory system, address field of an instruction code has a sufficient number of bits to specify all virtual addresses. In our example above we have 20-bit address of an instruction (to refer 20-bit virtual address) but physical memory addresses are specified with 15-bits. So a table is needed to map a virtual address of 20-bits to a physical address of 15-bits. Mapping is a dynamic operation, which means that every address is translated immediately as a word is referenced by CPU.

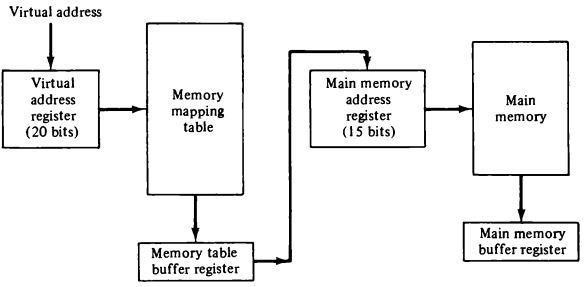


Fig: Memory table for mapping a virtual address