### Unit # 4 Memory Organization

**Dr. Rajesh Tiwari Professor ( CSE – AIML) CMREC, Hyderabad, Telangana**

### Memory Hierarchy

* The memory unit that communicates directly with the CPU is called the main Memory.
* Devices that provide backup storage are called auxiliary memory.
* The most common auxiliary memory devices used in computer

systems are magnetic disks and tapes.

# Memory Hierarchy

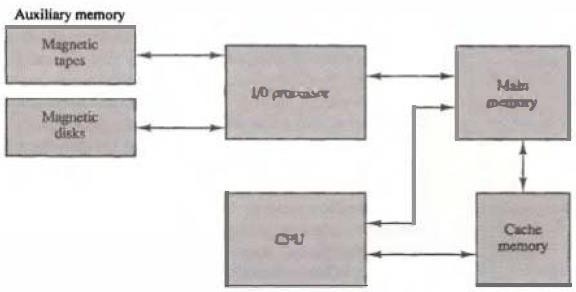


Figure 4·1: Memory hierarchy in Computer system.

# Memory Hierarchy

**Memory Hierarchy**

* A special very-high speed memory called a Cache is sometimes used to increase the speed of processing by making current programs and data available to the CPU at a rapid rate.
* The cache memory is employed in computer systems to compensate for the speed differential between main memory access time and processor logic.
* CPU logic is usually faster than main memory access time, with the result that processing speed is limited primarily by the speed of main memory.
* A technique used to compensate for the mismatch in operating speeds is to employ an extremely fast, small cache between the CPU and main memory whose access time is close to processor logic clock cycle time.
* The cache is used for storing segments of programs currently being executed in the CPU and temporary data frequently needed in the present calculations.
* **Auxiliary and cache** memories are used for different purposes.
* The **cache holds** those parts of the program and data that are most heavily used, while the **auxiliary memory** holds those parts that are not presently used by the CPU.
* The **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 typical **access time** ratio between cache and main memory is about 1 to 7.
  + For example,
    - A typical cache memory may have an access time of 100 ns, while main memory access time may be 700 ns.
    - Auxiliary memory average access time is usually 1000 times that of main memory.
    - Block size in auxiliary memory typically ranges from 256 to 2048 words, while cache block size is typically from 1 to 16 words.

### Main Memory

* + 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.
  + The principal technology used for the main memory is based on semiconductor integrated circuits.
  + Integrated circuit RAM chips are of two operating modes,

**static and dynamic.**

* + The **static RAM** consists essentially of internal flip-flops that store the binary information. 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.

**Main Memory**

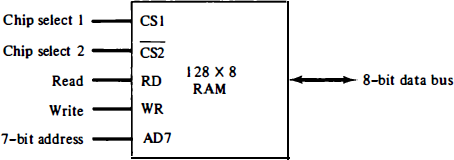
* The ROM portion of main memory is needed for storing an initial program called a bootstrap loader.
* The bootstrap loader is a program whose function is to start the

computer software operating when power is turned on.

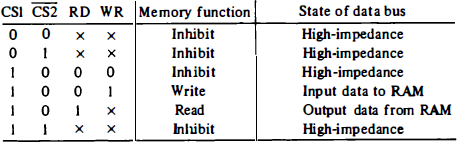
* RAM is volatile, its contents are destroyed when power is turned off.
* The contents of ROM remain unchanged after power is turned off and on again.
* When power is turned on, the hardware of the computer sets the program counter to the first address of the bootstrap loader.
* The bootstrap program loads a portion of the operating system from disk to main memory and control is then transferred to the operating system, which prepares the computer for general use.

## Main Memory

* + RAM and ROM chips are available in a variety of sizes.
  + If the memory needed for the computer is larger than the capacity of one chip, it is necessary to combine a number of chips to form the required memory size.
  + To demonstrate the chip interconnection, we will show an example of a 1024 x 8 memory constructed with 128 x 8 RAM chips and 512 x 8 ROM chips.



* + 1. Block diagram



* + 1. Function table

Figure 4-2: Typical RAM chip.

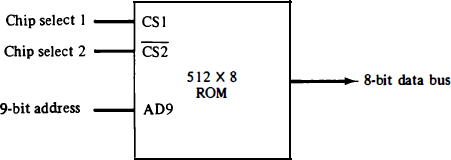


Figure 4-3: Typical ROM chip.

### Memory Address Map

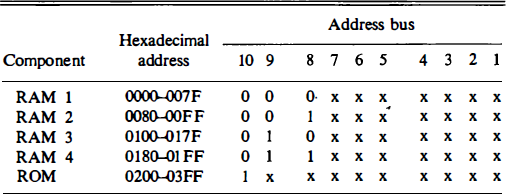
* + The designer of a computer system must calculate the amount of memory required for the particular application and assign it to either RAM or ROM.
  + The interconnection between memory and processor is then established from knowledge of the size of memory needed and the type of RAM and ROM chips available.
  + The addressing of memory can be established by means of a table that specifies the memory address assigned to each chip.
  + The table, called a memory address map, is a pictorial representation of assigned address space for each chip in the system.
  + The memory address map for this configuration is shown in Table 4.1 .
  + The component column specifies whether a RAM or a ROM chip is used.
  + The hexadecimal address column assigns a range of hexadecimal equivalent addresses for each chip.
  + The address bus lines are listed in the third column. Although there are 16 lines in the address bus, the table shows only 10 lines because the other 6 are not used in this example and are assumed to be zero.
  + The small x's under the address bus lines designate those lines that must be connected to the address inputs in each chip.
  + The RAM chips have 128 bytes and need seven address lines.
  + The ROM chip has 512 bytes and needs 9 address lines.
  + The x's are always assigned to the low-order bus lines: lines 1 through 7 for the RAM and lines 1 through 9 for the ROM.
  + It is now necessary to distinguish between four RAM chips by assigning to each a different address.
  + For this particular example we choose bus lines 8 and 9 to

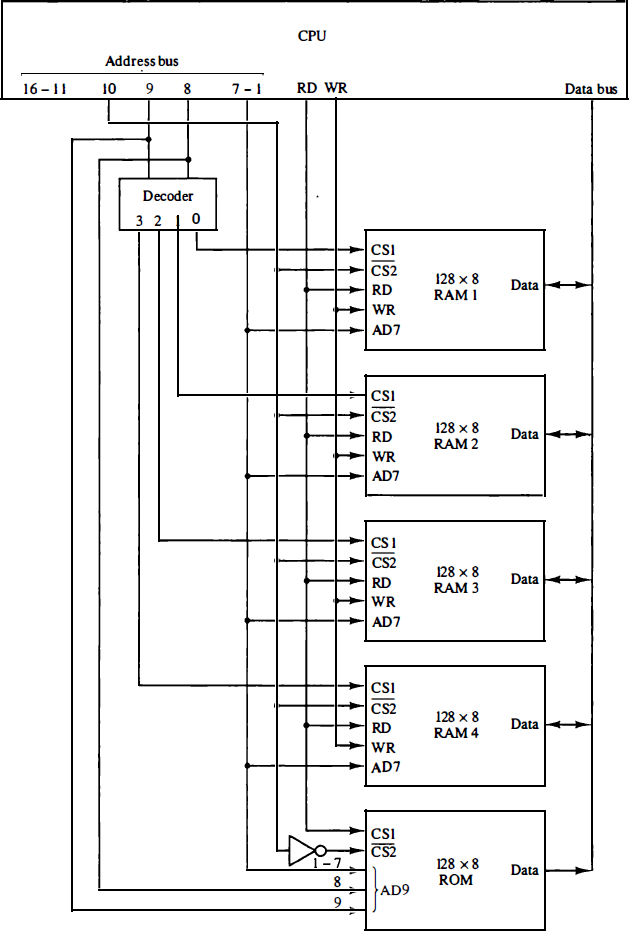
represent four distinct binary combinations.

* + The table clearly shows that the nine low-order bus lines constitute a memory space for RAM equal to 29 = 512 bytes.
  + The distinction between a RAM and ROM address is done with another bus line.
  + When line 10 is 0, the CPU selects a RAM, and when this line is equal to 1, it selects the ROM.

## Memory Address Map

TABLE 4-1 Memory Address Map for Microprocomputer





**Figure 4-4: Memory connection to the CPU**.

### Auxiliary Memory

* + The average time required to reach a storage location in memory and obtain its contents is called the **access time**.
  + In electromechanical devices with moving parts such as disks and tapes, **the access time consists of a seek time required to position the read-write head to a location and a transfer time required to transfer data to or from the device**.
  + Because the seek time is usually much longer than the transfer

time, auxiliary storage is organized in records or blocks.

* + A **record** is a specified number of characters or words.
  + Reading or writing is always done on entire records.
  + The **transfer rate** is the number of characters or words that the device can transfer per second, after it has been positioned at the beginning of the record.

### Associative Memory

* The time required to find an item stored in memory can be reduced considerably if **stored data can be identified for access by the content of the data** itself rather than by an address.
* A memory unit accessed by content is called an **associative**

**memory** or **content addressable memory** (CAM).

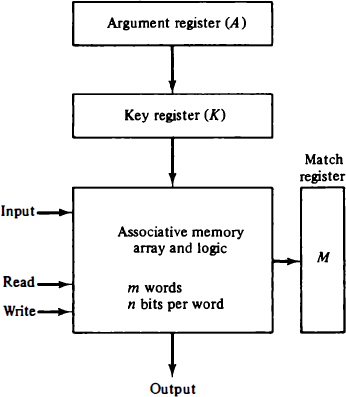
* This type of memory is accessed simultaneously and in parallel on the basis of data content rather than by specific address or location.
* When a word is written in an associative memory, no address is given. The memory is capable of finding an empty unused location to store the word.
* When a word is to be read from an associative memory, the content of the word, or part of the word, is specified.
* The memory locates all words which match the specified content and marks them for reading.

#### Associative Memory

* The block diagram of an associative memory is shown in Fig. 4.6.
* It consists of a memory array and logic for m words with n bits per word.
* The argument register A and key register K each have n bits, one for each bit of a word.
* The match register M has m bits, one for each memory word.
* Each word in memory is compared in parallel with the content of the argument register.
* The words that match the bits of the argument register set a corresponding bit in the match register.
* After the matching process, those bits in the match register that have been set indicate the fact that their corresponding words have been matched.
* Reading is accomplished by a sequential access to memory for those words whose corresponding bits in the match register have

been set.

## Associative Memory



**Figure 4.6: Block diagram of associative memory.**

## Cache Memory

* + If the active portions of the program and data are placed in a fast small memory, the average memory access time can be reduced, thus reducing the total execution time of the program.
  + This fast small memory is referred to as a cache memory.
  + It is placed between the CPU and main memory.
  + The cache memory access time is less than the access time of main memory by a factor of 5 to 10.
  + The cache is the fastest component in the memory hierarchy and approaches the speed of CPU components.

## Cache Memory

* + The performance of cache memory is frequently measured in terms of a quantity called hit ratio .
  + When the CPU refers to memory and finds the word in cache, it is said to produce a hit .
  + If the word is not found in cache, it is in main memory and it counts as a miss .
  + The ratio of the number of hits divided by the total CPU references to memory (hits plus misses) is the hit ratio.
  + The hit ratio is best measured experimentally by running representative programs in the computer and measuring the number of hits and misses during a given interval of time.
  + Hit ratios of 0.9 and higher have been reported.
  + This high ratio verifies the validity of the locality of reference property.

# Cache Memory

* + The average memory access time of a computer system can be

**improved** considerably by use of a cache.

* + If the hit ratio is high enough so that most of the time the CPU accesses the cache instead of main memory, the average access time is closer to the access time of the fast cache memory.
  + For example :
* A computer with cache access time of 100 ns, a main memory access time of 1000 ns, and a hit ratio of 0.9 produces an average access time of 200 ns.
* This is a considerable improvement over a similar computer

without a cache memory, whose access time is 1000 ns.

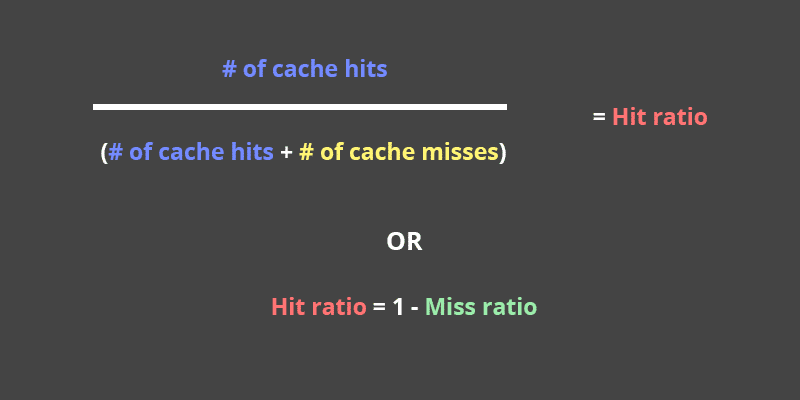
## Cache Memory

* + The basic characteristic of cache memory is its fast access time.
  + Very little or **no time must** be wasted when searching for words in the cache.
  + The transformation of data from main memory to cache memory is referred to as a mapping process.
  + Three types of mapping procedures are of practical interest

when considering the organization of cache memory:

* Associative mapping
* Direct mapping
* Set-associative mapping

## Hit Ratio



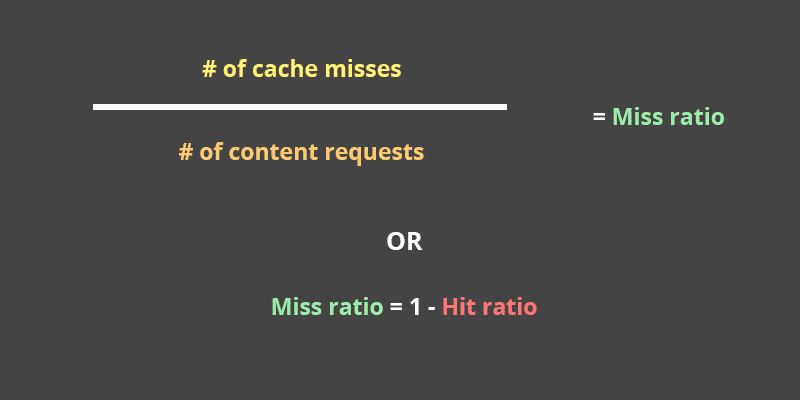
### Hit Ratio

* + To calculate a hit ratio, divide the number of cache hits with the sum of the number of cache hits, and the number of cache misses.
  + For example:
* if you have 51 cache hits and three misses over a period of time, then that would mean you would divide 51 by 54.
* The result would be a hit ratio of 0.944.
  + You can also choose to express this as a percentage by

multiplying the end result by 100.

* + In the example above, the 0.944 result would be multiplied by 100 to get a hit ratio of 94.4%.
  + Alternatively, you can find out the hit ratio if you already know the miss ratio. Then, you can subtract one by the miss ratio.

# Miss Ratio



**Miss Ratio**

* + We can calculate a miss ratio by dividing the number of misses with the total number of content requests.
  + For example,

– if you look over a period of time and find that the misses your cache experienced was11, and the total number of content requests was 48, you would divide 11 by 48 to get a miss ratio of 0.229.

* + Similar to hit ratios, we can calculate a miss ratio if you

already know the hit ratio.

* + In that case, you can calculate it by subtracting one from the hit ratio.

•

* + **What’s the Ideal Hit and Miss Ratios in Caches?**
  + Generally speaking, for most sites, a hit ratio of 95-99%, and a miss ratio of one to five percent is ideal.
  + Keep in mind that every site is different so these aren’t one-size-fits-all numbers.

#### Thank you