Memory Hierarchy

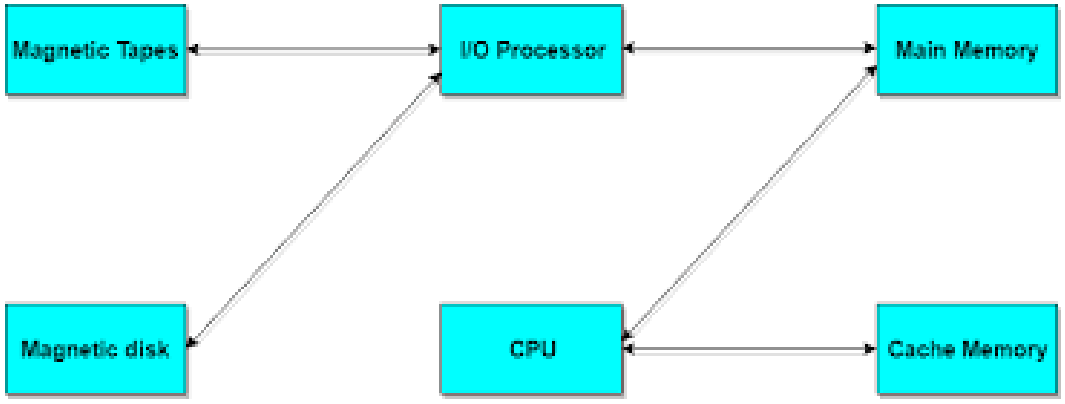
- The memory hierarchy consists of all the storage devices in the computer from the slow but high capacity auxiliary memory to a fast but low capacity cache memory.

- The **magnetic tape** used to store removable files. The most common auxiliary memory devices used in computer systems are magnetic disks and tapes. They are used for storing system programs, large data files, and other backup information. The capacity is larger but speed is very slow.

- The **main memory** interacts with CPU. The data in the auxiliary memory is brought in the main 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.

- 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



- Many operating systems are designed to enable the CPU to process a number of independent programs concurrently. This concept, **called multiprogramming.**

- It refers to the existence of two or more programs indifferent parts of the memory hierarchy at the same time. In this way it is possible to keep all parts of the computer busy by working with several programs in sequence

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 available in two possible 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 unit.

- **The dynamic RAM** stores the binary information in the form of electric charges that are applied to capacitors.

- 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.

- 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.

- 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.

RAM AND ROM CHIPS

- 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 × 8 memory constructed with 128 × 8 RAM chips and 512 × 8 ROM chips.

RAM CHIP

Block Diagram Representing 128 x 8 RAM
(Random Access Memory)

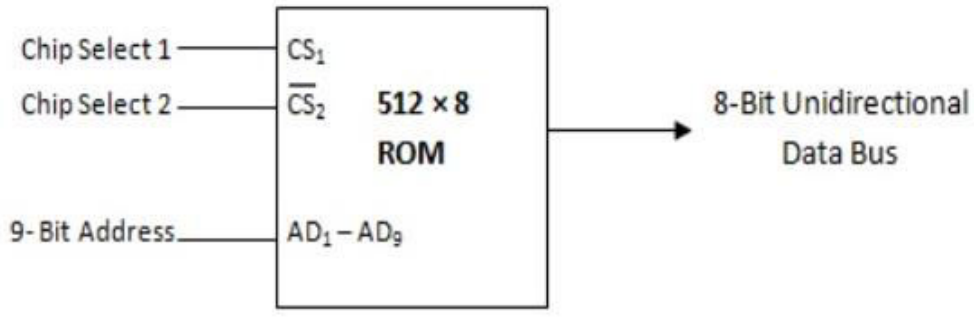
- The memory capacity is of 128 words having word length 8 bits.
- Thus it requires a 7-bit address (as 128=27) and 8 bit bidirectional data bus.
- The RD (read) and WR (write) signal indicates the operations to be performed.

- The two chip select lines CS1 and CS2 used to enable the chip when selected by microprocessor. When CS1=1 and CS2=0, the chip is activated.

CS_1	$\overline{CS_2}$	RD	WR	Memory function	State of data bus
0	0	X	X	Inhibit	High-impedence
0	1	X	X	Inhibit	High-impedence
1	0	0	0	Inhibit	High-impedence
1	0	0	1	Write	Input data to RAM
1	0	1	X	Read	Output data from RAM
1	1	X	X	Inhibit	High Impedence

ROM CHIP

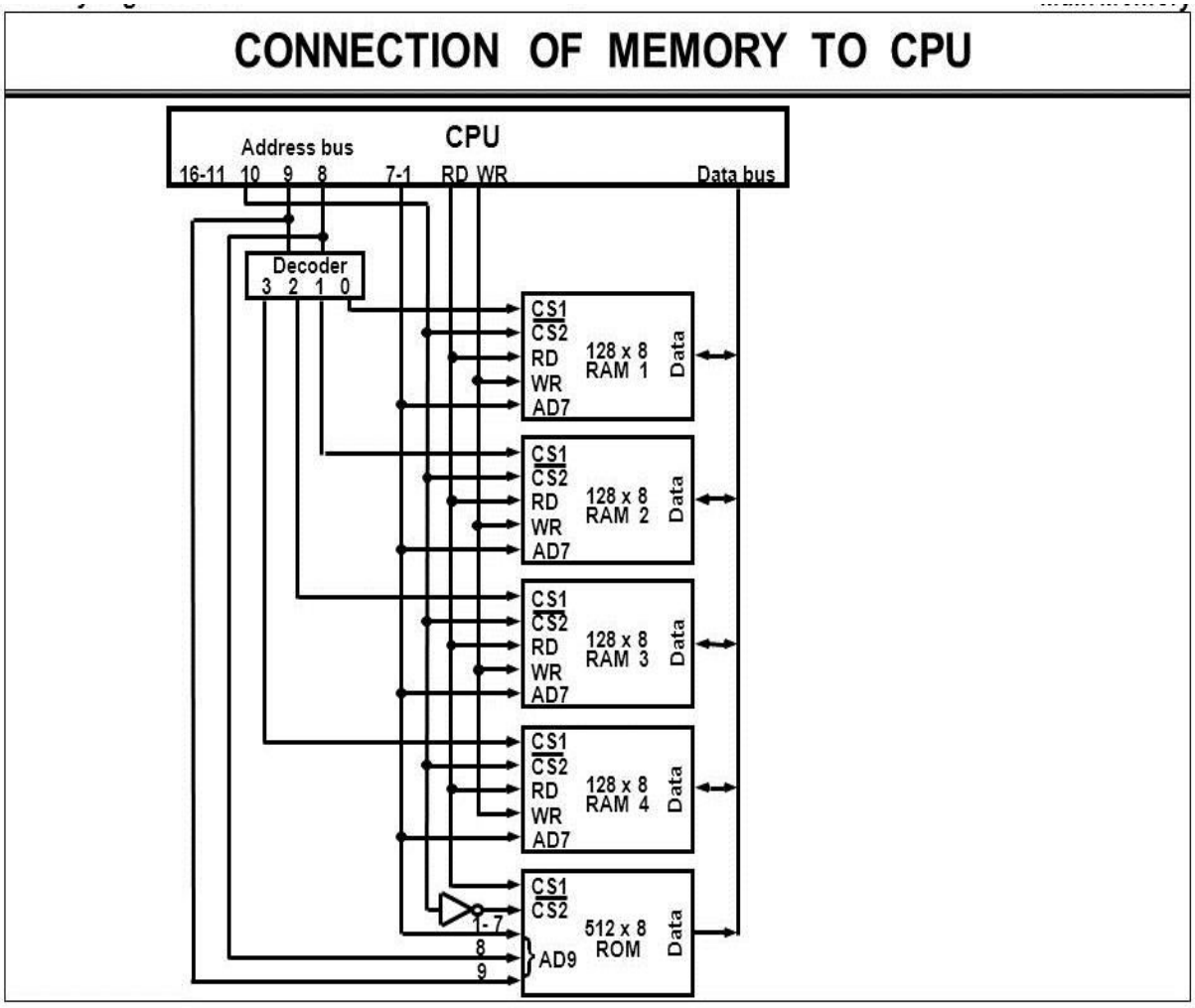
- A ROM chip is organized externally in a similar manner. However, since a ROM can only read, the data bus can only be in an output mode.
- The block diagram of a ROM chip is shown in diagram for the same-size chip, it is possible to have more bits of ROM occupy less space than in RAM. For this reason, the diagram specifies a 512-byte ROM, while the RAM has only 128 bytes.



- The nine address lines in the ROM chip specify any one of the 512 bytes stored in it.

- The two chip select inputs must be CS1 = 1 and CS2 = 0 for the unit to operate. Otherwise, the data bus is in a high-impedance state. There is no need for a read or write control because the unit can only read. Thus when the chip is enabled by the two select inputs, the byte selected by the address lines appears on the data bus.

MEMORY CONNECTION TO CPU



- Each RAM receives the seven low order bits of the address bus to select one of 128 possible bytes. The particular RAM chip selected is determined from lines 8 and 9 in the address bus.

- This is done through a 2 × 4 decoder whose outputs go to the CS1 input in each RAM chip. Thus, when address lines 8 and 9 are equal to 00, the first RAM chip is selected. When 01, the second RAM chip is selected, and

so on. The RD and WR outputs from the microprocessor are applied to the inputs of each RAM chip.

- **The selection between RAM and ROM is achieved through bus line 10. The RAMs are selected when the bit in this line is 0, and the ROM when the bit is 1.**

- The other chip select input in the ROM is connected to the RD control line for the ROM chip to be enabled only during a read operation.

- Address bus lines 1 to 9 are applied to the input address of ROM without going through the decoder. This assigns addresses 0 to 511 to RAM and 512 to 1023 to ROM. The data bus of the ROM has only an output capability, whereas the data bus connected to the RAMs can transfer information in both directions.

Auxiliary Memory

- The most common auxiliary memory devices used in computer systems are magnetic disks and tapes. Other components used, but not as frequently, are magnetic drums, magnetic bubble memory, and optical disks.

Magnetic Disks

- A magnetic disk is a circular plate constructed of metal or plastic coated with magnetized material. Often both sides of the disk are used and several disks may be

stacked on one spindle with read/write heads available on each surface.

- All disks rotate together at high speed and are not stopped or started from access purposes. Bits are stored in the magnetized surface in spots along concentric circles called tracks.

- The tracks are commonly divided into sections called sectors. In most systems, the minimum quantity of information which can be transferred is a sector. The sub division of tone disk surface into tracks and sectors.

Magnetic Tape

- A magnetic tape transport consists of the electrical, mechanical, and electronic components to provide the parts and control mechanism for a magnetic-tape unit.

- The tape itself is a strip of plastic coated with a magnetic recording medium. Bits are recorded as magnetic spots on the tape along several tracks. Usually, seven or nine bits are recorded simultaneously to form a character together with a parity bit.

- Read/write heads are mounted one in each track so that data can be recorded and read as a sequence of characters.

- Magnetic tape units can be stopped, started to move forward or in reverse, or can be rewound. However, they cannot be started or stopped fast enough between individual characters.

- For this reason, information is recorded in blocks referred to as records. Gaps of unrecorded tape are inserted between records where the tape can be stopped.

Associative Memory

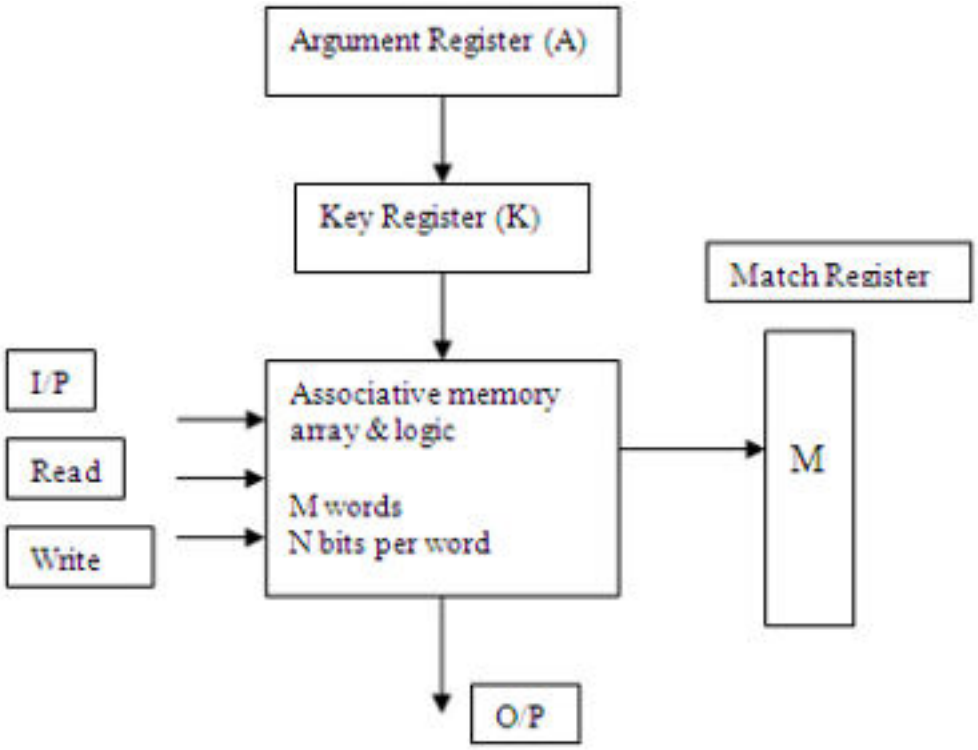
- The time required to find an item stored in memory can be reduced considerably if stored data can be identified for access by the content of the data itself rather than by an address.

- 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.

Hardware Organization

- The block diagram of an associative memory is shown in diagram. It consists of a memory array and logic from words with n bits per word.



- The key register provides a mask for choosing a particular field or key in the argument word. The entire argument is compared with each memory word if the key register contains all 1's.

- Otherwise, only those bits in the argument that have 1's in their corresponding position of the key register are compared.

- Thus the key provides a mask or identifying piece of information which specifies how the reference to memory is made.

- To illustrate with a numerical example, suppose that the argument register A and the key register K have the bit configuration shown below. Only the three leftmost bits of A are compared with memory words because K has 1's in these positions.

A	101 111100
K	111 000000
Word 1	100 111100 no match
Word 2	101 000001 match

- Word 2 matches the unmasked argument field because the three leftmost bits of the argument and the word are equal.

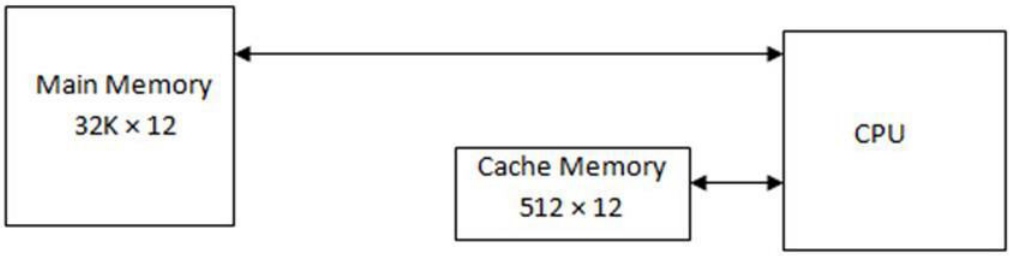
Cache Memory

- 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. Such a fast small memory is referred to as a cache memory.

- Cache memory is an intermediate buffer between CPU main memory. The objective is to reduce CPU waiting time during main memory access.

Example

A system with 512 x 12 cache and 32 K x 12 of main memory.



- The cache memory access time is less than the access time of main memory The cache is the fastest component in the memory hierarchy **The fundamental idea of cache organization is that by keeping the most frequently accessed instructions and data in the fast cache memory.**

Basic operation of the cache

- When the CPU needs to access memory, the cache is examined. If the word is found in the cache, it is read from the fast memory. If the word addressed by the CPU is not found in the cache, the main memory is accessed to read the word. A block of words containing the one just accessed is then transferred from main memory to cache memory.

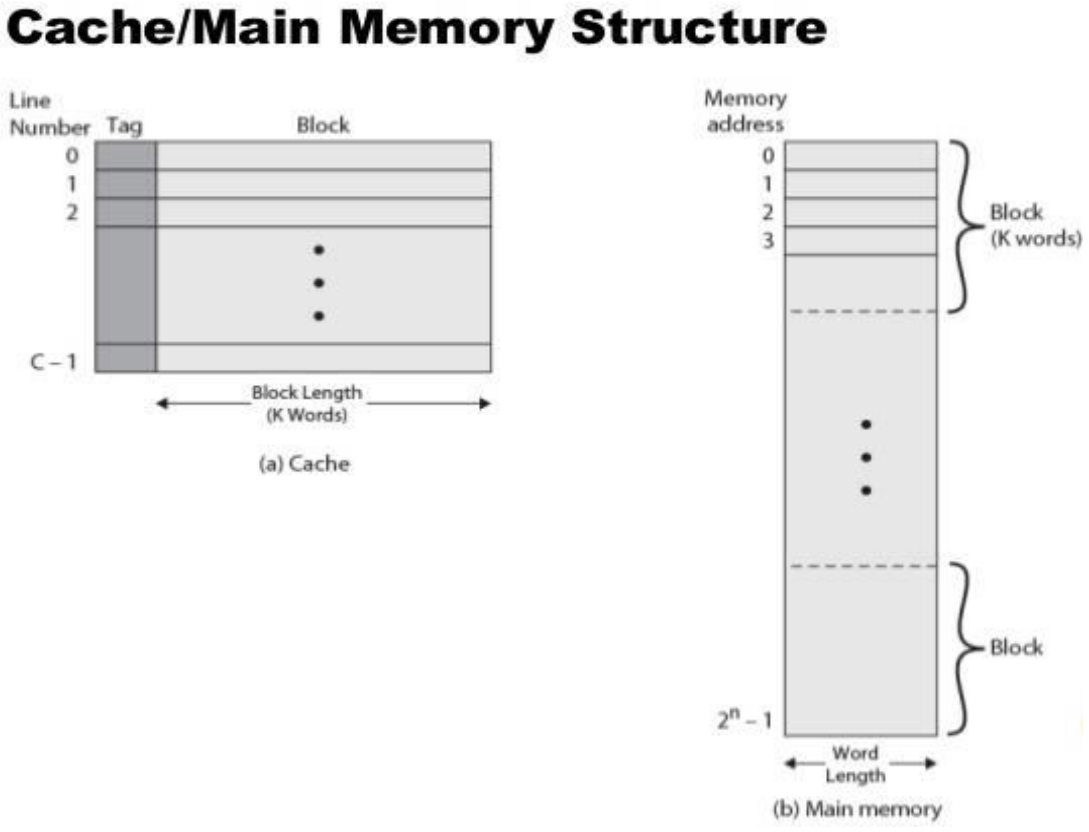
Hit Ratio

- 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**

Structure of cache



- The main memory is conceptually divided into many blocks. Each containing a fixed number of consecutive locations.

- While reading a location from main memory the content of entire block is transferred & stored in cache memory.

- The cache memory is organized as number of lines or blocks & size of each line is same as the capacity of main memory.

Mapping

- The transformation of data from main memory to cache memory is referred to as a mapping process.

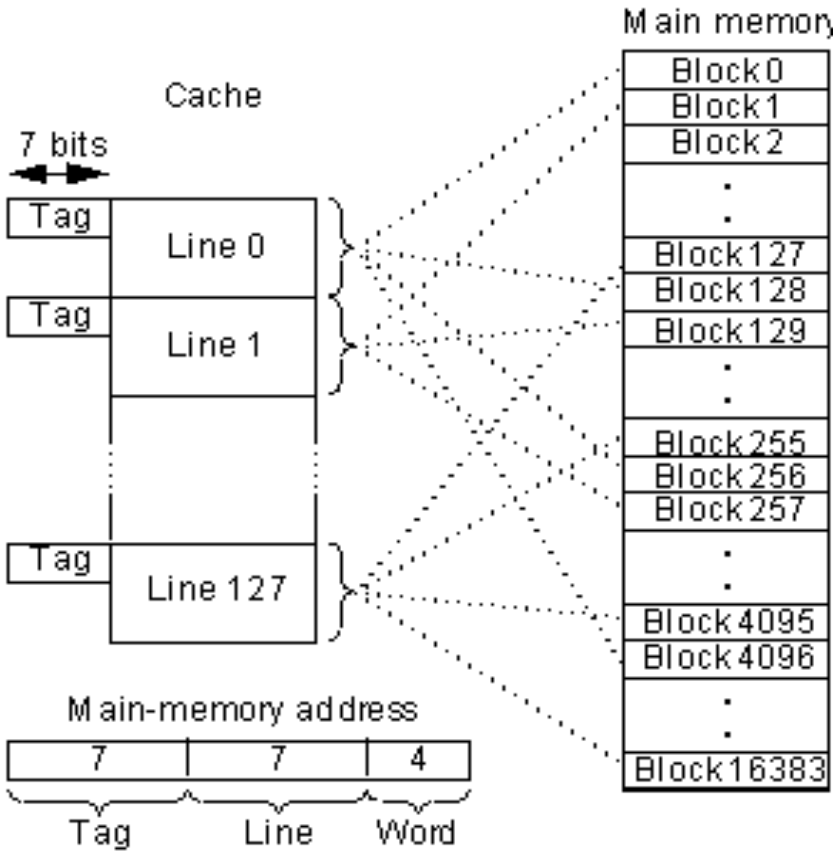
- When a main memory block is stored in cache, the line number where block has been written is determined from **mapping function.**

- Three types of mapping

    1. Direct mapping
    2. Associative mapping
    3. Set-associative mapping

Let's take an example of main memory stores 256k words and the cache stores 2k (2048) words with 16 words per block. Thus cache has 128 blocks and main memory has 16,384 blocks.
Thus CPU generates $256k=2^8*2^{10}=2^{18}$ .so 18 bit address used to looks into cache for data.

1. Direct mapping

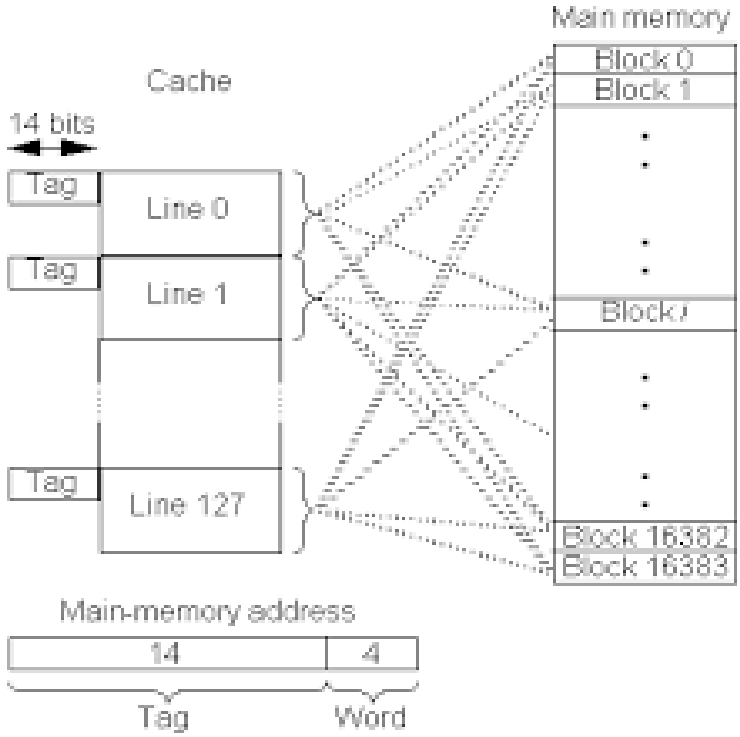


- In this mapping, block i of the main memory maps into the block frame of 128 of the cache.

- The memory address consists of 3 fields tag(7 bit), block(7 bit) and word(4 bit).
- Each block frame has its own tag associated with it.
- When a block from main memory exists in a block frame in cache, the **tag** associated with that frame contains the high order 7 bit of the main memory address.
- The 7 bit **block** address is used to address the corresponding block frame.
- The 7 bit **tag field** is compared with the tag in the cache block frame. If they match, the data in the block frame is accessed by 4 bit word.

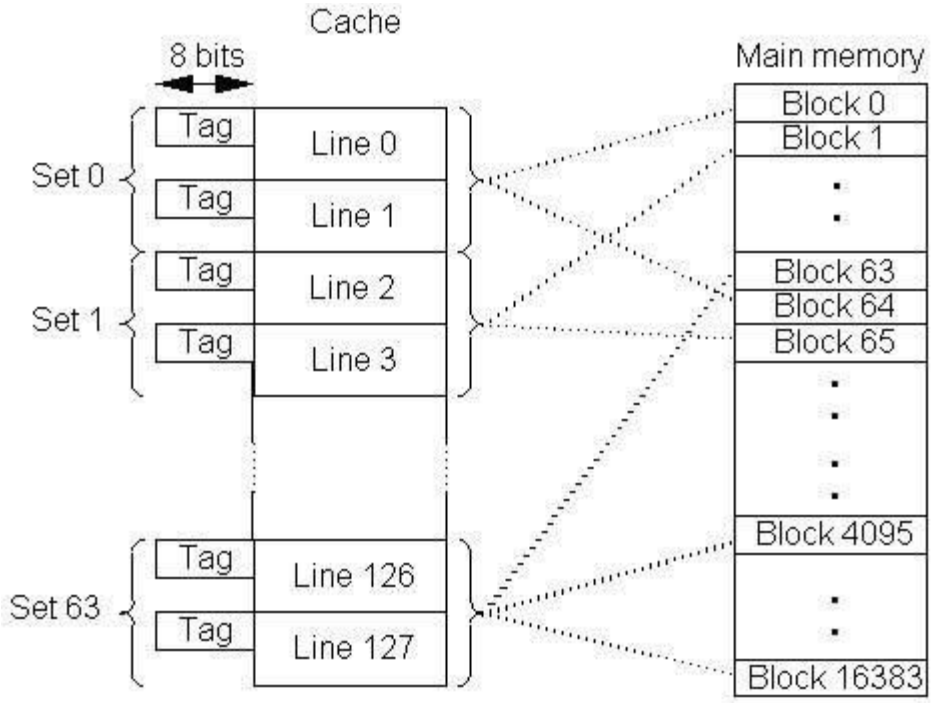
2. Associative Mapping

- In this mapping, any block in memory can be can be mapped in any block frame of cache memory.
- When request comes for a block, all the entries are compared simultaneously to determine if the requested block is present or not.
- The format of memory address has **tag field** (14 bit) and **word** (4 bit).**tag field** is used to identify memory block when it is in cache.



3. Set-Associative Mapping

- It combines the concepts of direct mapping with associative mapping.
- It is an improvement over the direct mapping as cache is divided into S sets. Each set has multiple lines of equal number.
- A main memory block can be mapped into a specific set only within a set,the memory block can be placed in any of K lines.
- Within the set, they can be mapped associatively to one of two block frames.



Virtual Memory

- It is a concept that helps on run a large program in a small physical memory.
- Each address that is referenced by the CPU goes through an address mapping from the so called virtual address to a physical address in main memory.
- Virtual memory is used to give programmers the illusion that they have a very large memory, even though the computer actually has a relatively small main memory.
- This is done dynamically, while programs are being executed in the CPU. The translation or mapping is handled automatically by the hardware by means of a mapping table.