Cache Memory

Cache Memory

Locality of reference

The references to memory at any given interval of time tend to be contained within a few localized areas in 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 "Cache Memory".

The performance of the cache memory is measured in terms of a quality called "Hit Ratio".

When the CPU refers to memory and finds the word in cache, it produces a hit. If the word is not found in cache, it counts it as a miss.

The ratio of the number of hits divided by the total CPU references to memory (hits + misses) is the hit ratio.

Cache Memory

The average memory access time of a computer system can be improved considerably by use of cache.

The cache is placed between the CPU and main memory. It is the faster component in the hierarchy and approaches the speed of CPU components.

When the CPU needs to access memory, the cache is examined. If it is found in the cache, it is read very quickly.

If it is not found in the cache, the main memory is accessed.

A block of words containing the one just accessed is then transferred from main memory to cache memory.

.

What is the high speed memory between the main memory and the CPU called?

a) Register Memory
b) Cache Memory
c) Storage Memory
d) Virtual Memory

Cache Memory

The basic characteristic of cache memory is its fast access time. Therefore, very little or no time must be wasted when searching for words in the cache.

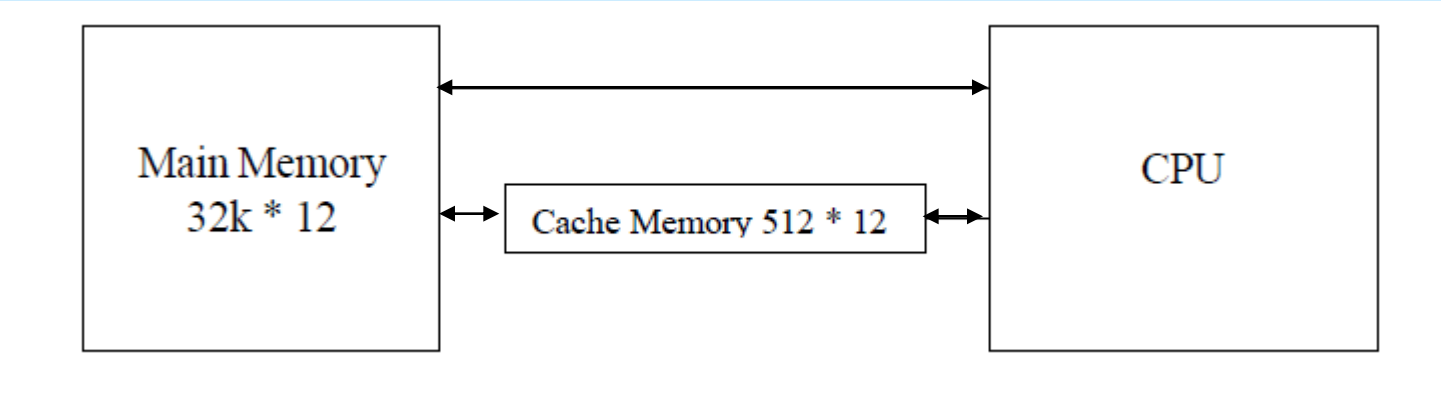
The transformation of data from main memory to cache memory is referred to as a "Mapping Process".

There are three types of mapping procedures are available.

- Associative Mapping
- Direct Mapping
- Set – Associative Mapping.

Cache Memory

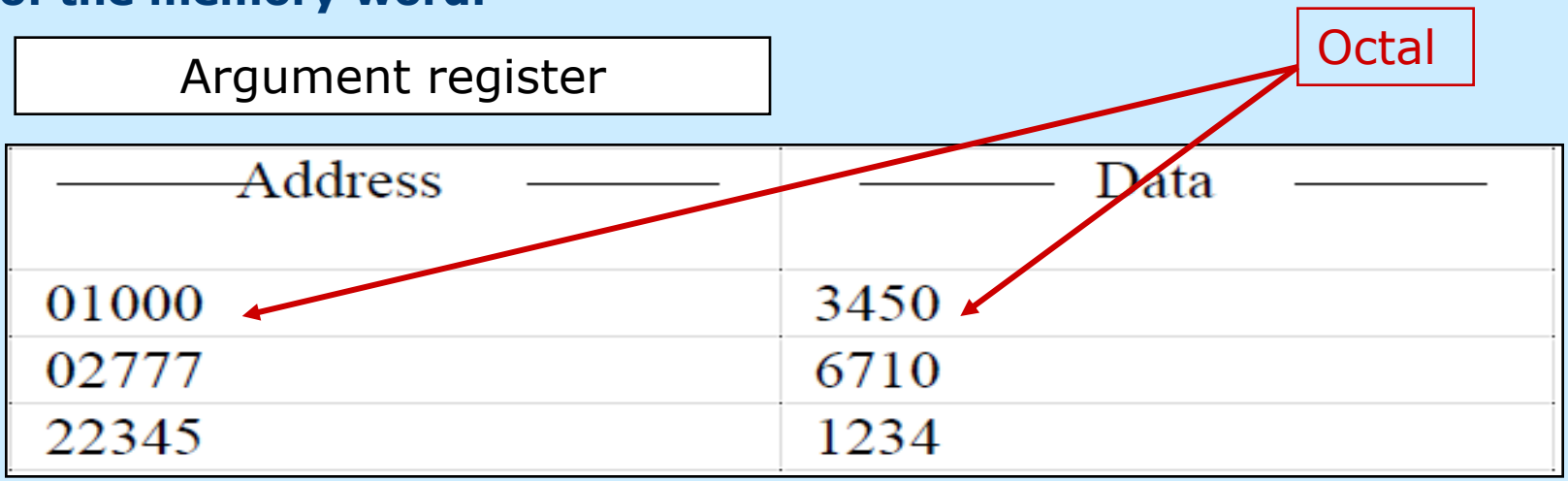
Consider the following memory organization to show mapping procedures of the cache memory.



· The main memory can stores 32k word 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.

Associative Mapping

The associative mapping stores both the address and content (data) of the memory word.



A CPU address of 15 bits is placed in the argument register and associative memory is searched for a matching address.

If the address is found, the corresponding 12 bit data is read and sent to the CPU.

If no match occurs, the main memory is accessed for the word. The address – data pair is then transferred to associative cache memory.

If the cache is full, it must be displayed, using replacement algorithm. FIFO may be used.

Cache Memory is implemented using the DRAM chips.
a) True
b) False

Direct Mapping

The 15-bit CPU address is divided into two fields.

The 9 least significant bits constitute the index field and the remaining 6 bits form the tag fields.

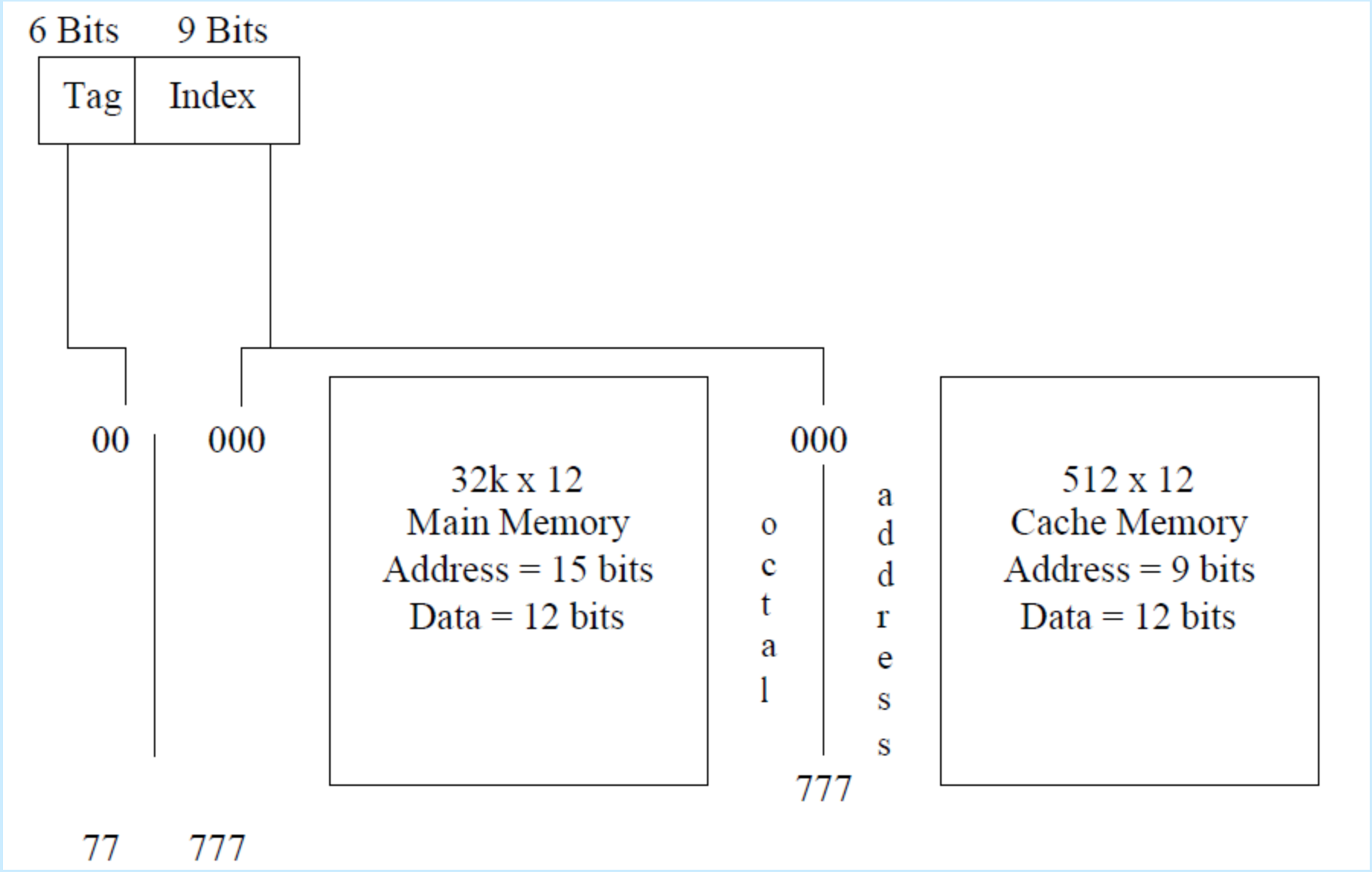
The main memory needs an address but includes both the tag and the index bits.

The cache memory requires the index bit only i.e., 9 bits.

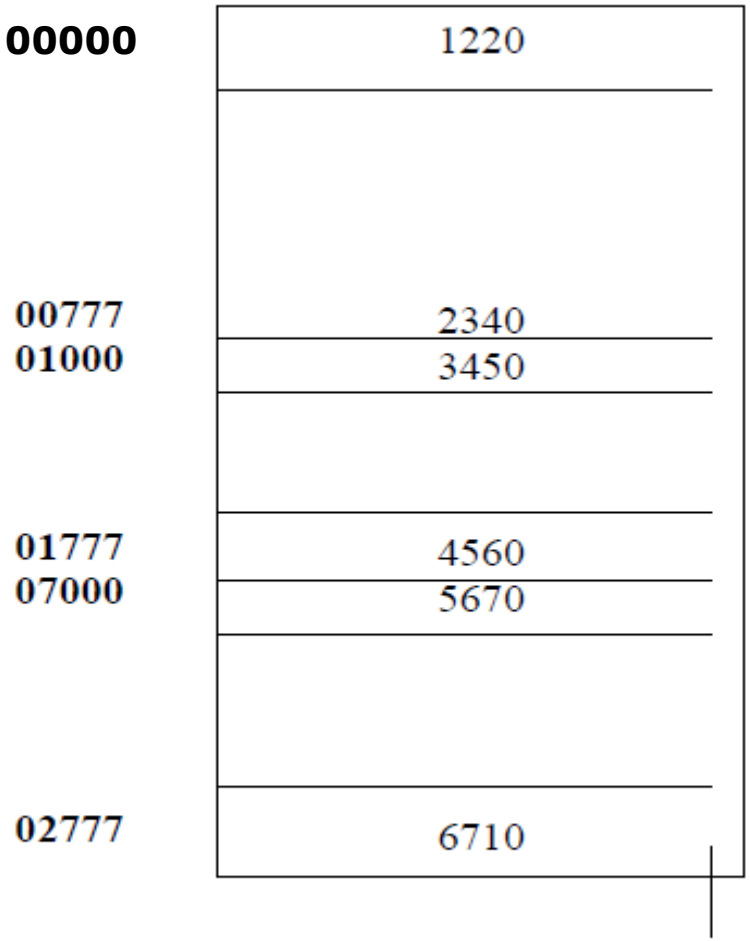
There are 2^k words in the cache memory & 2^n words in the main memory.

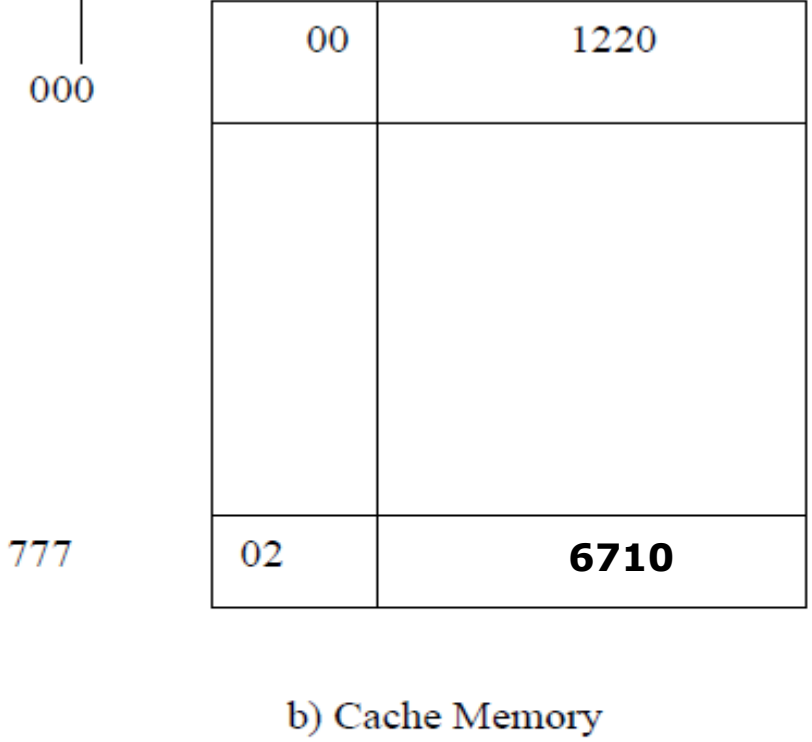
e.g: k = 9, n = 15

Direct Mapping



Direct Mapping





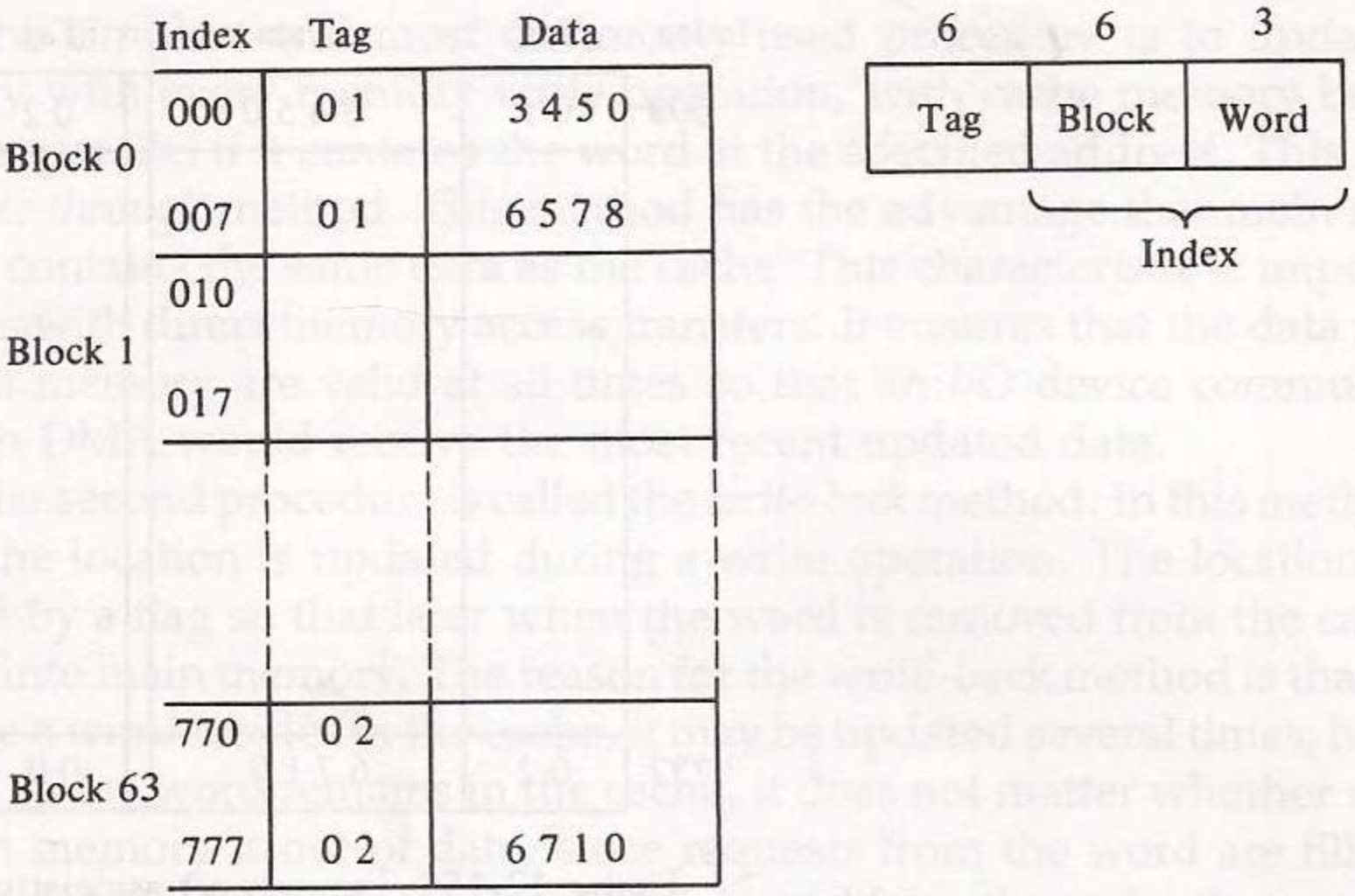


Figure 12-14 Direct mapping cache with block size of 8 words.