CACHE MEMORY

Maninder Kaur

professormaninder@gmail.com

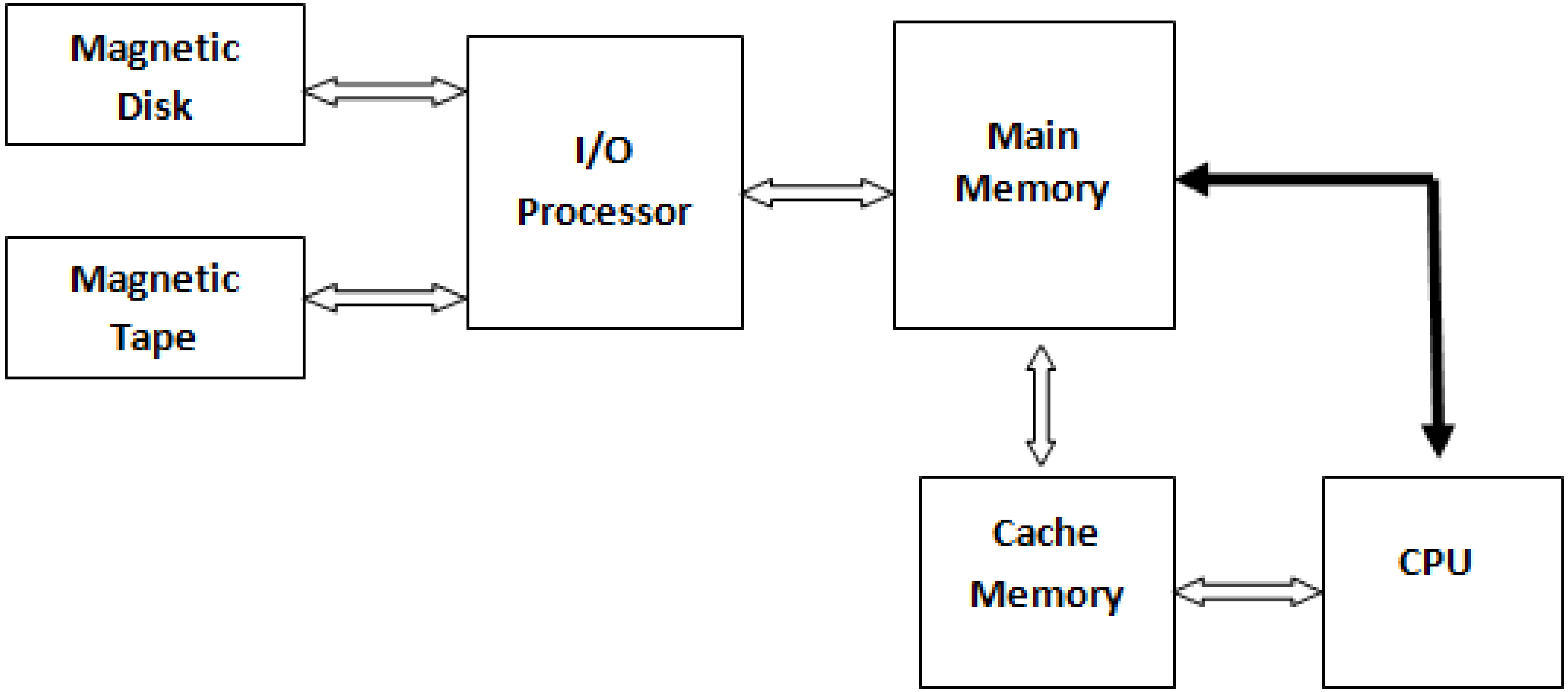
What is Cache Memory?

- Cache memory is a small, high-speed RAM buffer located between the CPU and main memory.

- Cache memory holds a copy of the instructions (instruction cache) or data (operand or data cache) currently being used by the CPU.

- The main purpose of a cache is to accelerate your computer while keeping the price of the computer low.

Placement of Cache in computer



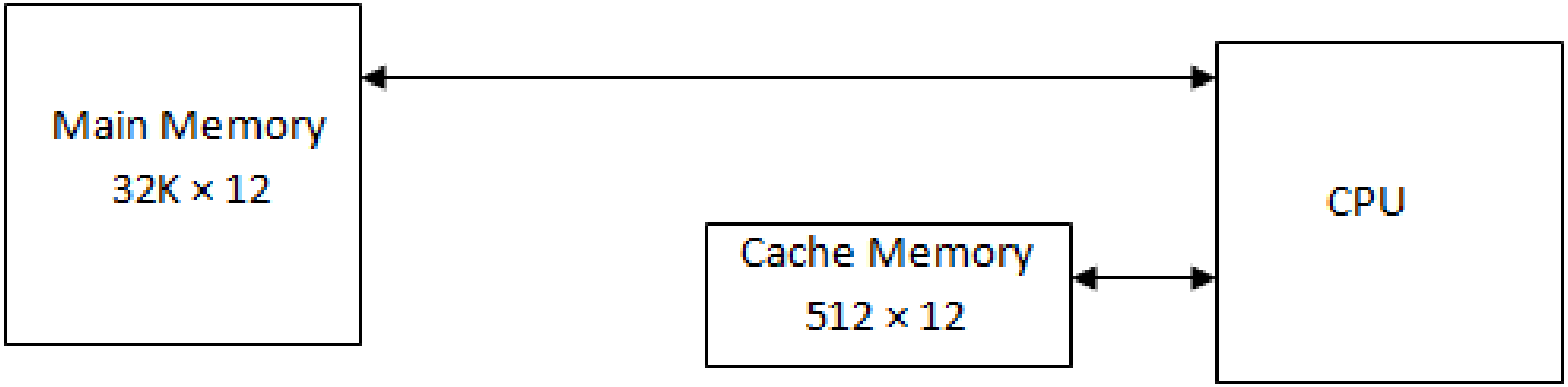
Hit Ratio

- The ratio of the total number of hits divided by the total CPU accesses to memory (i.e. hits plus misses) is called *Hit Ratio.*

- **Hit Ratio = Total Number of Hits / (Total Number of Hits + Total Number of Miss)**

Example

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



Types of Cache Mapping

1. Direct Mapping

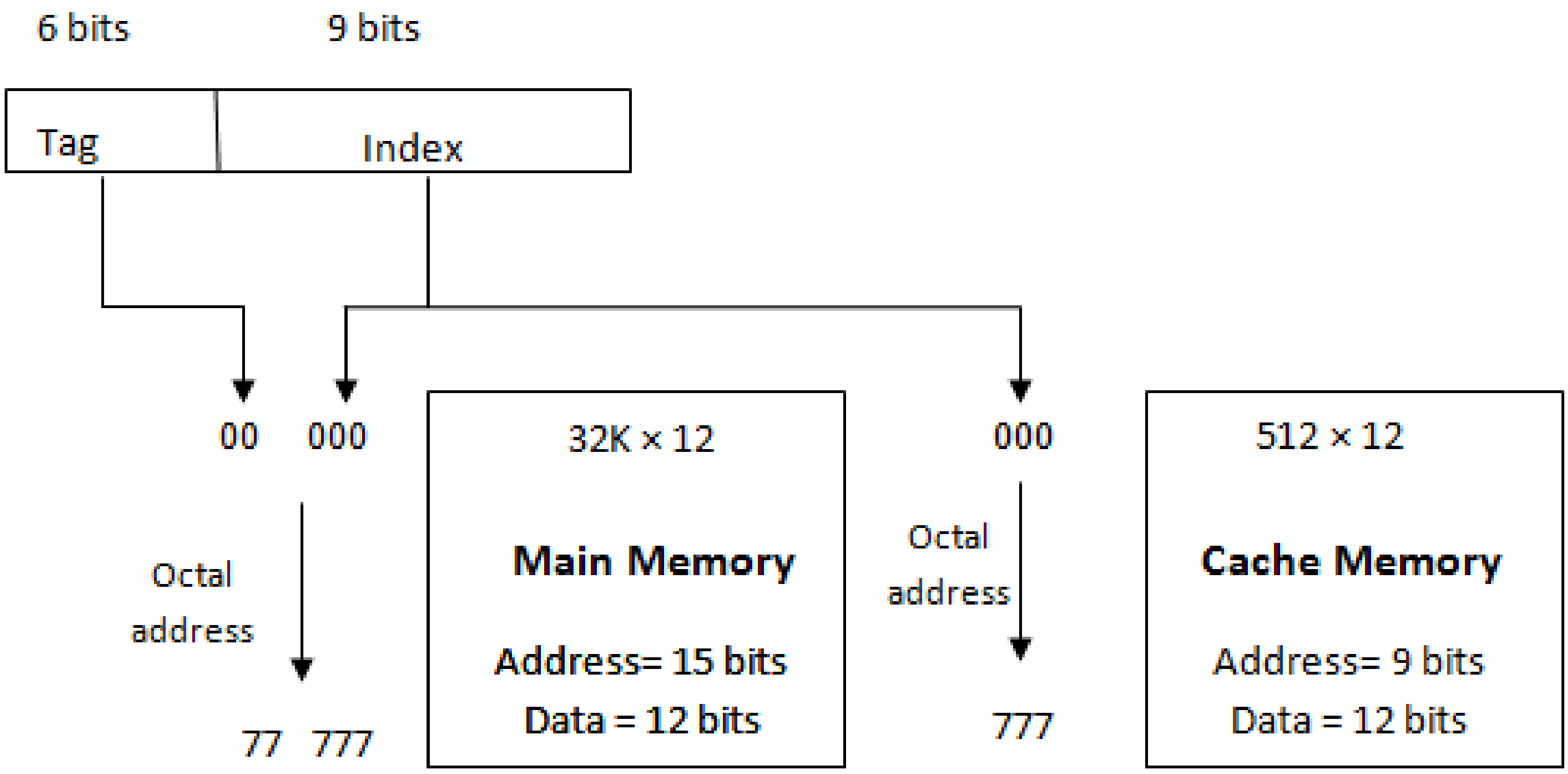
2. Associative Mapping

3. Set Associative Mapping

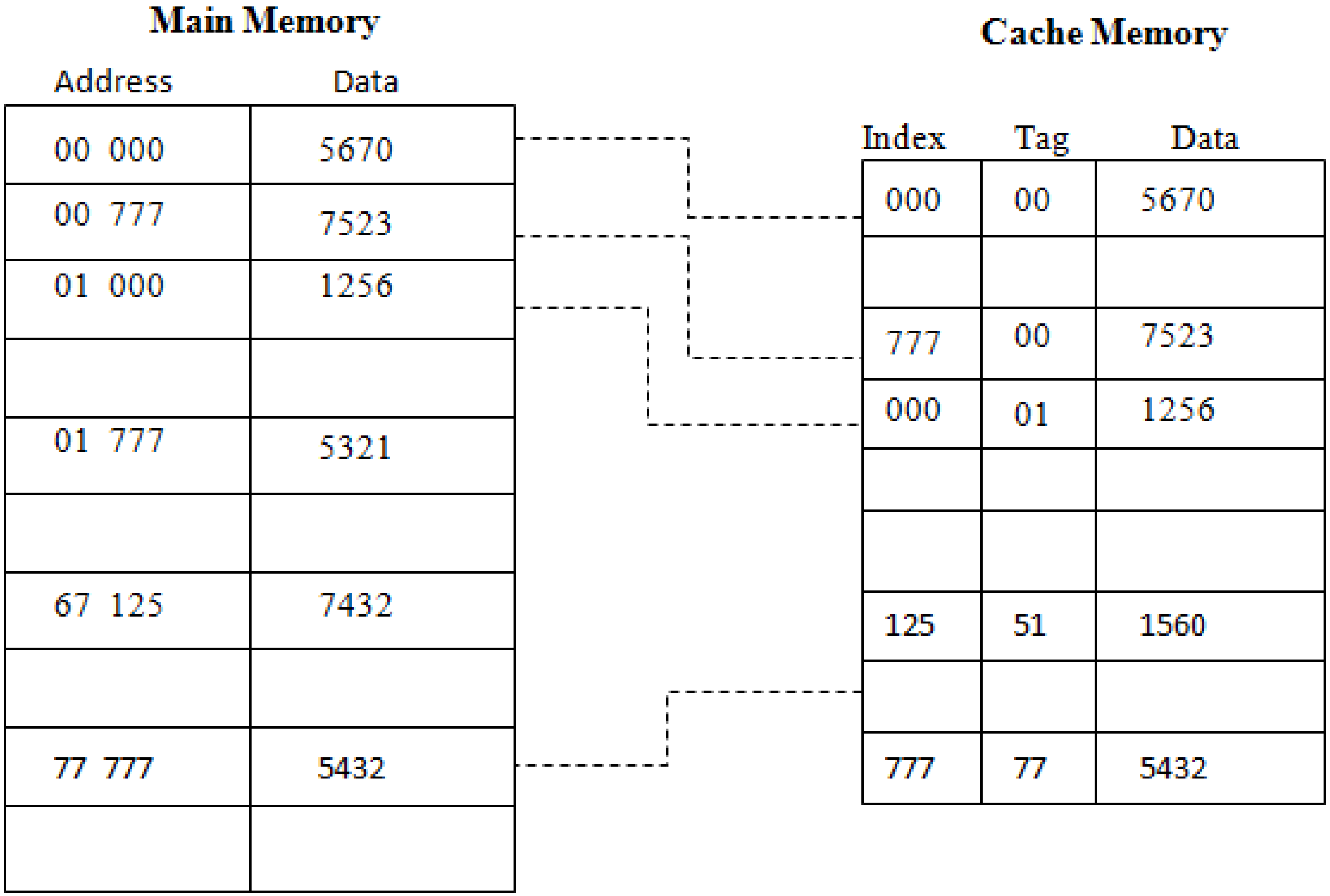
1. Direct Mapping

- The direct mapping technique is simple and inexpensive to implement.
- When the CPU wants to access data from memory, it places a address. The index field of CPU address is used to access address.
- The tag field of CPU address is compared with the associated tag in the word read from the cache.
- If the tag-bits of CPU address is matched with the tag-bits of cache, then there is a *hit* and the required data word is read from cache.
- If there is no match, then there is a *miss* and the required data word is stored in main memory. It is then transferred from main memory to cache memory with the new tag.

1. Direct Mapping



1. Direct Mapping



Main Memory

Address	Data
00 000	5670
00 777	7523
01 000	1256
01 777	5321
67 125	7432
77 777	5432

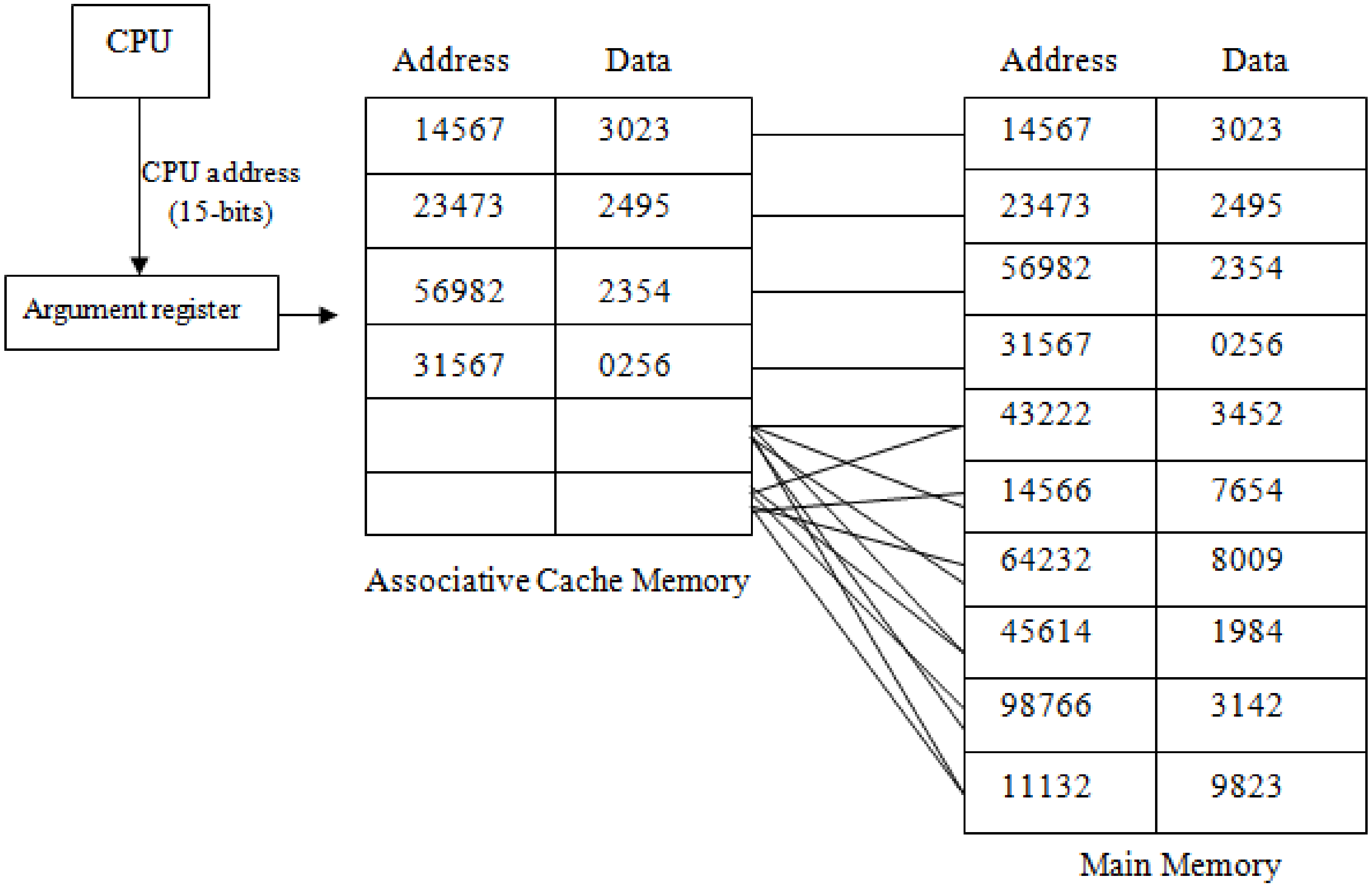
Cache Memory

Index	Tag	Data
000	00	5670
777	00	7523
000	01	1256
125	51	1560
777	77	5432

2. Associative Mapping

- An associative mapping uses an associative memory.
- This memory is being accessed using its contents.
- Each line of cache memory will accommodate the address (main memory) and the contents of that address from the main memory.
- That is why this memory is also called Content Addressable Memory (CAM). It allows each block of main memory to be stored in the cache.

2. Associative Mapping



Address	Data
14567	3023
23473	2495
56982	2354
31567	0256

Associative Cache Memory

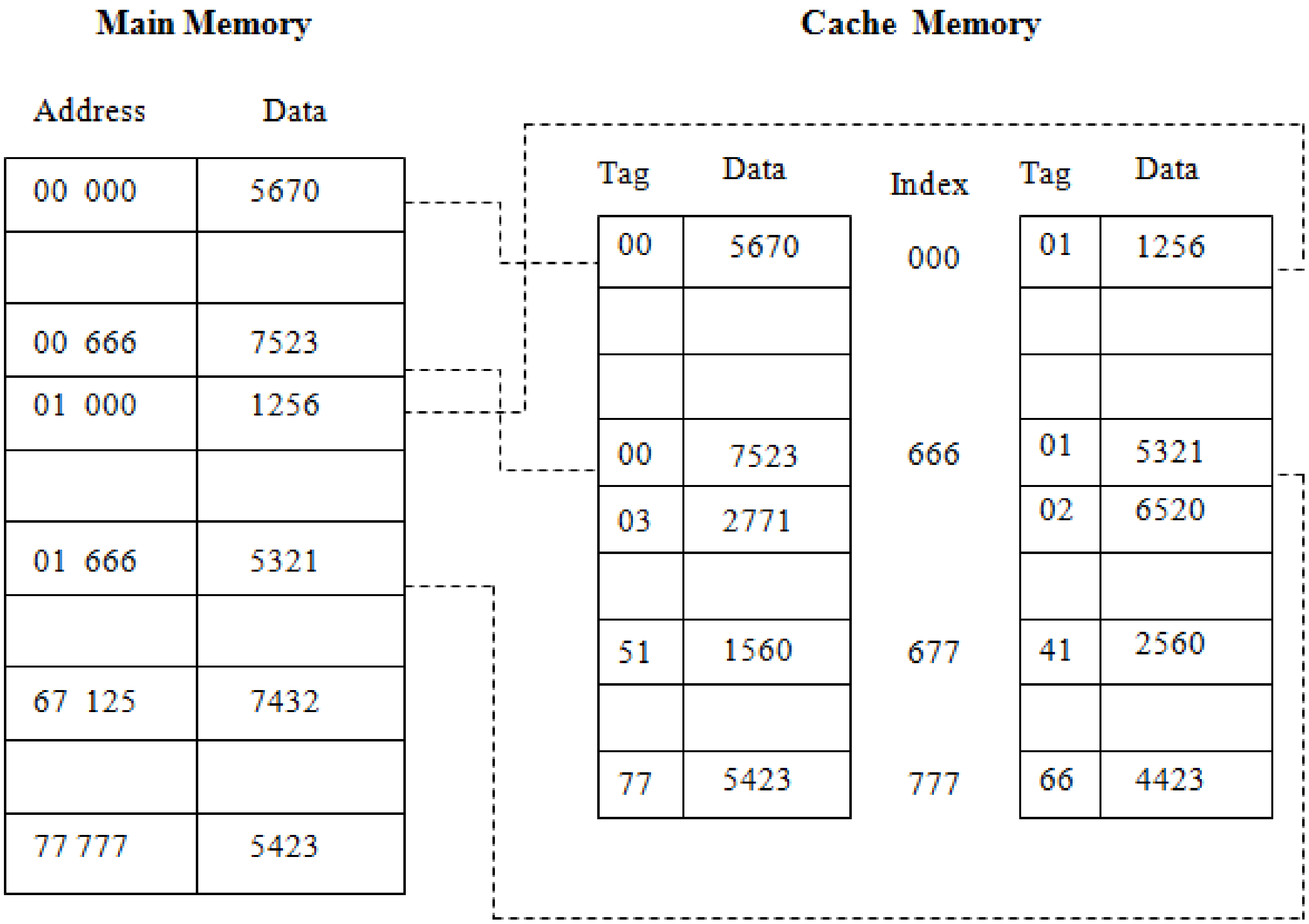
Address	Data
14567	3023
23473	2495
56982	2354
31567	0256
43222	3452
14566	7654
64232	8009
45614	1984
98766	3142
11132	9823

Main Memory

3. Set Associative Mapping

- That is the easy control of the direct mapping cache and the more flexible mapping of the fully associative cache.
- In set associative mapping, each cache location can have more than one pair of tag + data items.
- That is more than one pair of tag and data are residing at the same location of cache memory. If one cache location is holding two pair of tag + data items, that is called *2-way set associative mapping*.

3. Two-Way Set Associative Mapping



Main Memory

Address	Data
00 000	5670
00 666	7523
01 000	1256
01 666	5321
67 125	7432
77 777	5423

Cache Memory

Tag	Data	Index	Tag	Data
00	5670	000	01	1256
00	7523	666	01	5321
03	2771		02	6520
51	1560	677	41	2560
77	5423	777	66	4423

Replacement Algorithms of Cache Memory

- Replacement algorithms are used when there are no available space in a cache in which to place a data. Four of the most common cache replacement algorithms are described below:
- ***Least Recently Used (LRU):***
  - The LRU algorithm selects for replacement the item that has been least recently used by the CPU.
- ***First-In-First-Out (FIFO):***
  - The FIFO algorithm selects for replacement the item that has been in the cache from the longest time.
- ***Least Frequently Used (LRU):***
  - The LRU algorithm selects for replacement the item that has been least frequently used by the CPU.
- ***Random:***
  - The random algorithm selects for replacement the item randomly.

Writing into Cache

- When memory write operations are performed, CPU first writes into the cache memory. These modifications made by CPU during a write operations, on the data saved in cache, need to be written back to main memory or to auxiliary memory.

- These two popular cache write policies (schemes) are:
  - *Write-Through*
  - *Write-Back*

Write-Through

- In a write through cache, the main memory is updated each time the CPU writes into cache.

- The advantage of the write-through cache is that the main memory always contains the same data as the cache contains.

- This characteristic is desirable in a system which uses direct memory access scheme of data transfer. The I/O devices communicating through DMA receive the most recent data.

Write-Back

- In a write back scheme, only the cache memory is updated during a write operation.

- The updated locations in the cache memory are marked by a flag so that later on, when the word is removed from the cache, it is copied into the main memory.

- The words are removed from the cache time to time to make room for a new block of words.

Thank You

Have a Nice Day