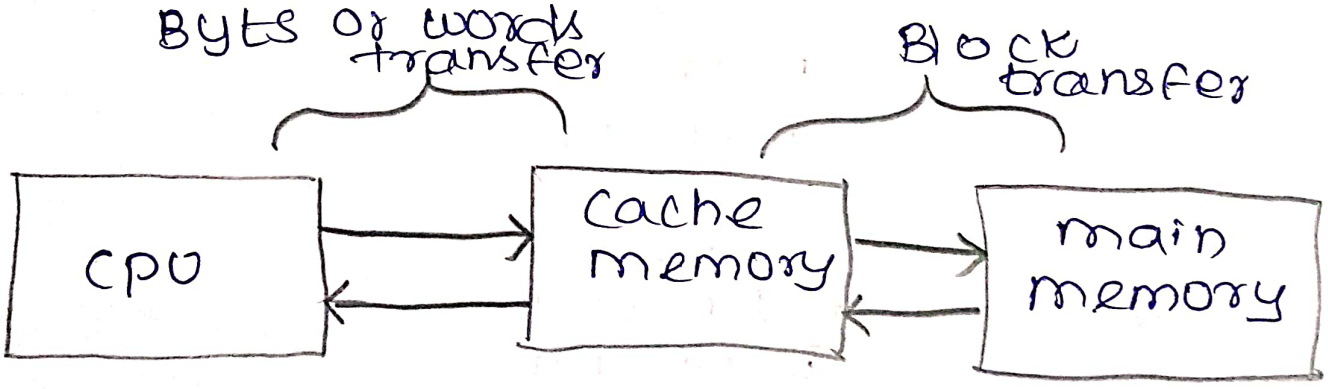
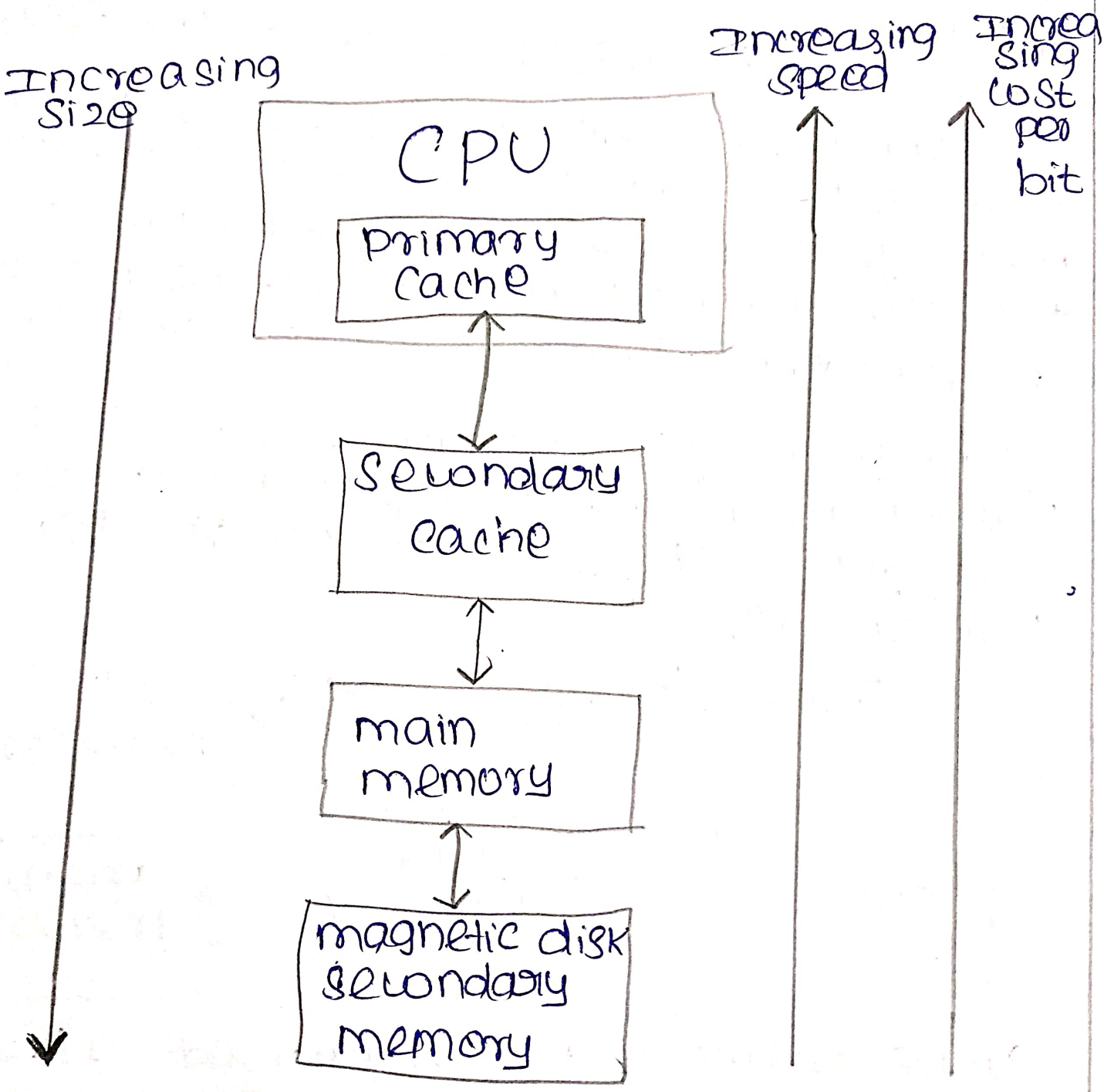
# Cache memory

The data or contents of the main memory that are used frequently by CPU are stored in the cache memory so that the processor can easily access that data in a shorter time. Whenever the CPU needs to access memory, it first checks the cache memory. If the data is not found in cache memory, then the CPU moves into the main memory. Cache memory is placed between the CPU and the main memory. The block diagram for a cache memory can be represented

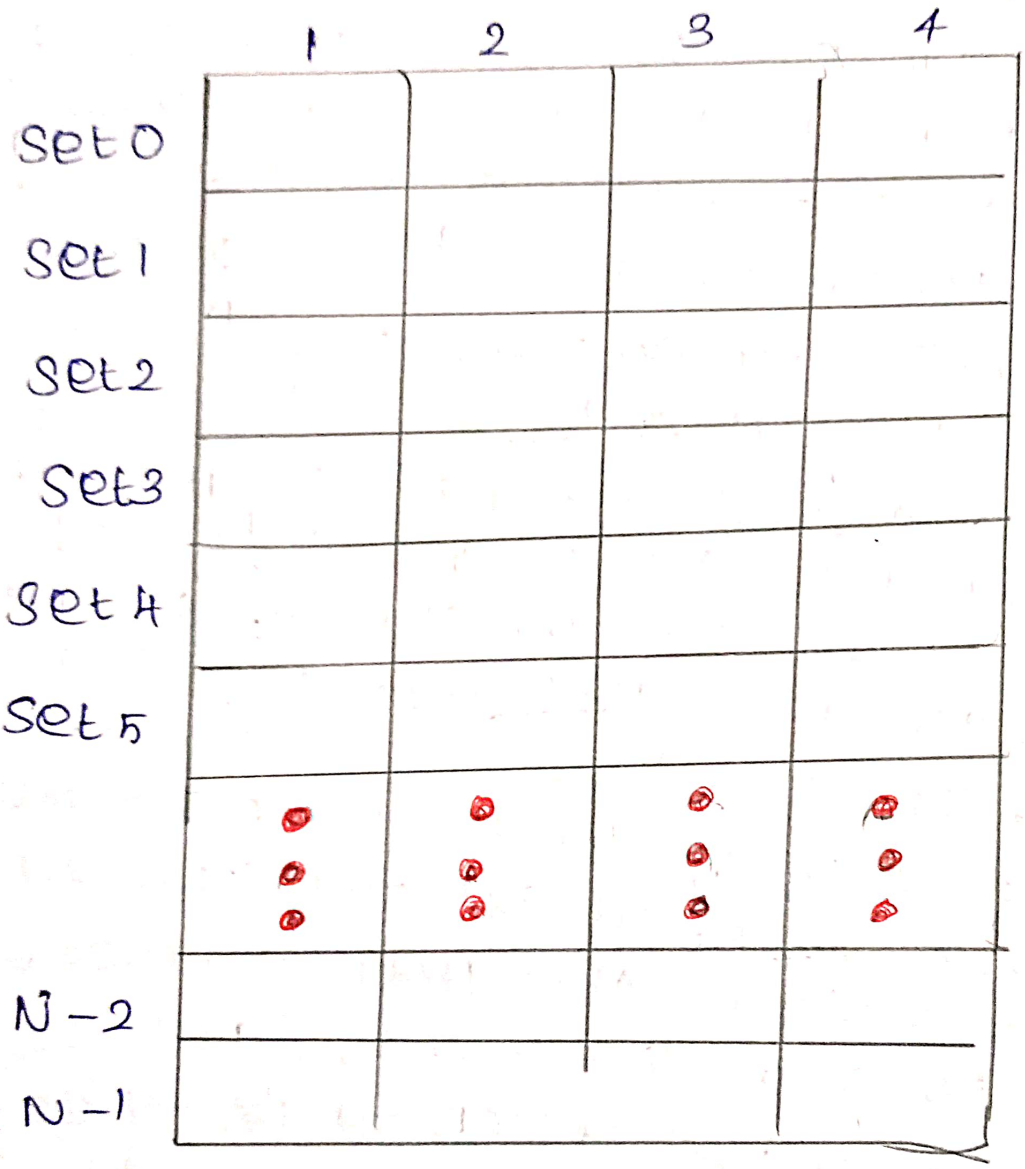
Cache memory: the cache is the fastest component in the memory hierarchy and approaches the speed of CPU components.




Cache memory is organised as distinct set of blocks where each set contains a small fixed number of blocks.


Increasing size
Increasing speed
Increasing cost per bit
CPU
primary cache
Secondary cache
main memory
magnetic disk secondary memory

# Associativity




Logical organization of a 4-way set associate cache

As shown in the above sets are represented by the rows. The example contains N sets and each ~~see~~ set contains four blocks. whenever an access is made to cache, the cache controller does not search the entire cache in order to look for a match. Rather the controller maps maps the address to a particular set of the cache and therefore searches only the set for a match. If a required block is not found in that set, the block is not present in the cache and cache controller does not search it further. This kind of cache organisation is called set associative because the cache is divided into distinct set of blocks. As each set contains four blocks the cache is to be four way set associative. whenthe 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 one just accessed is then transferred from main memory to cache memory. The block size may vary from on word about 16 words adjacent to the one just accessed.

The performance of the 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 the 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 is the hit ratio.

## Levels of memory:

**Level 1:** It is a type of memory in which data is stored and accepted that are immediately stored in CPU. Most commonly used register is accumulator, program counter, address register etc.

**Level 2:** It is the faster memory which has faster access time where data is temporarily stored for faster access.

Level 3: It is memory on which computer works currently. It is small in size and once power is OFF data no longer stays in this memory.

Level 4: It is external memory which is not as fast as main memory but stays permanently in this memory.

Cache mapping: There are three different types of mapping used for the purpose of cache memory which are as follows:

- Direct mapping
- Associative mapping
- Set-Associative mapping

Direct mapping: In direct mapping the cache consists of normal high speed random-access memory. Each location in the cache holds the data, at a specific address in the cache. This address is given by the lower significant bits of the main memory address. This enables the block to be selected directly from the lower significant bit of the memory address. The remaining higher significant bits of the address are stored in the cache with the data to complete the identification of the cached data.