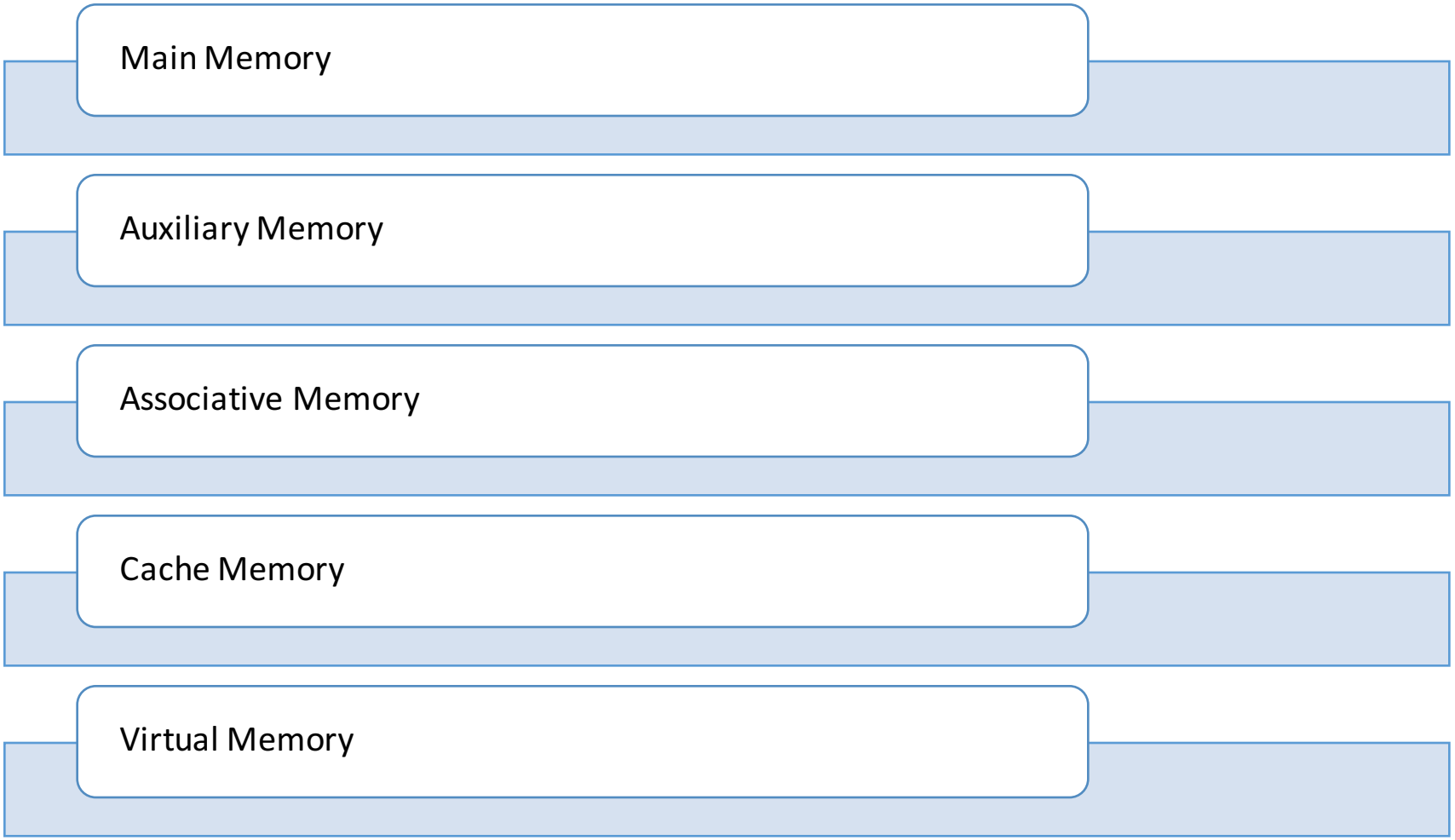
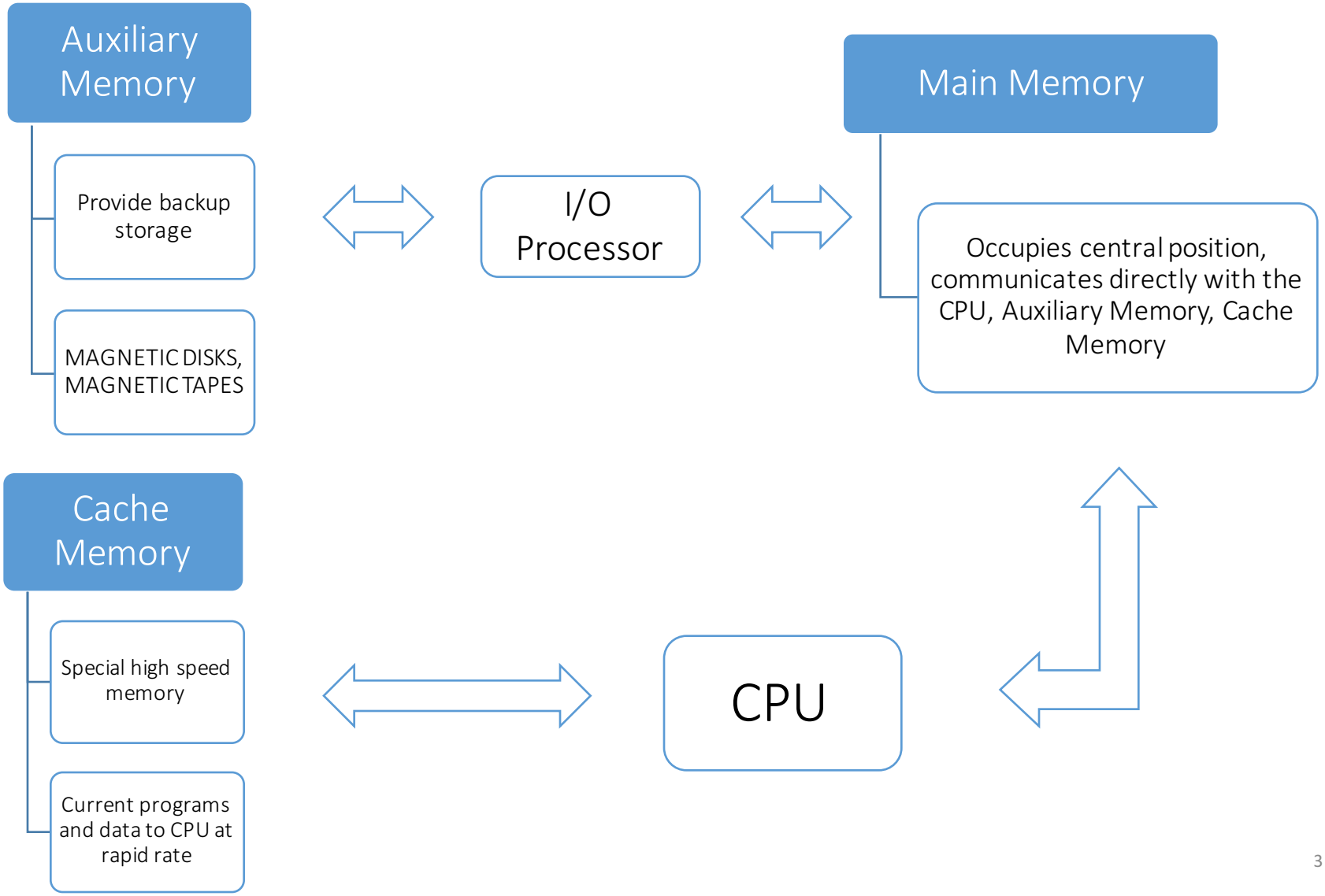
Memory Hierarchy

Sabina Batyrkhanovna

Memory Hierarchy

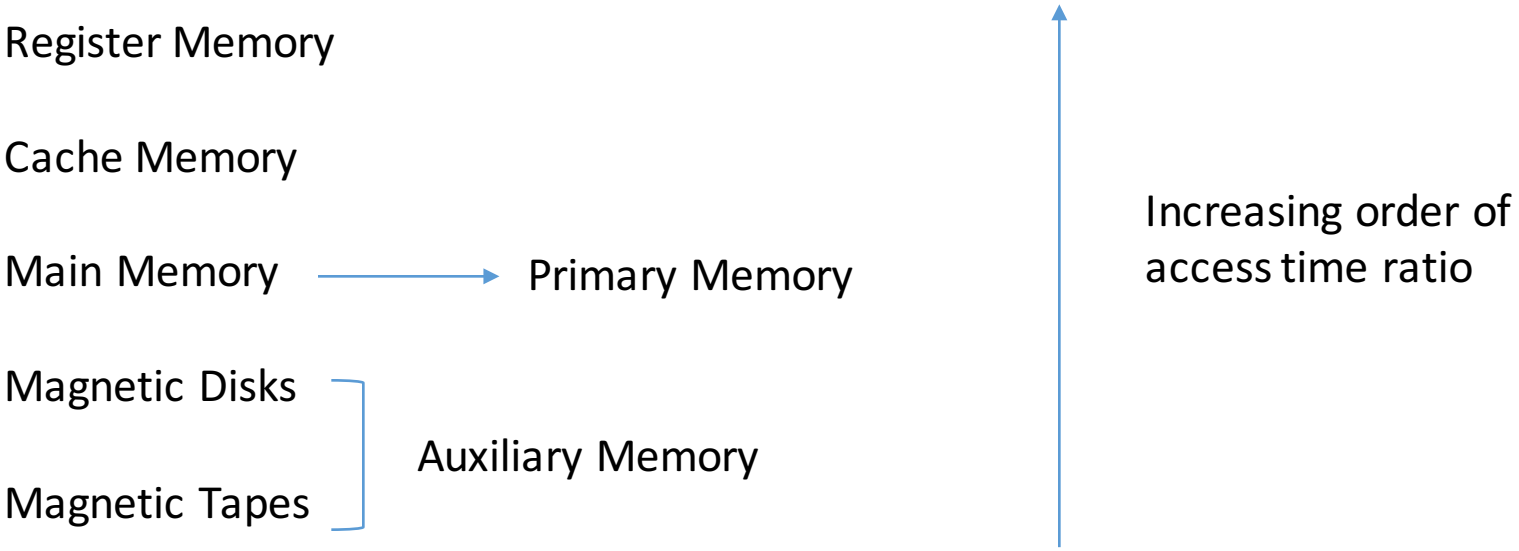


Memory Hierarchy



Memory Hierarchy

- The memory hierarchy system consists of all storage devices employed in a computer system from the slow by high-capacity auxiliary memory to a relatively faster main memory, to an even smaller and faster cache memory.



Access Methods

- Each memory is a collection of various memory location. Accessing the memory means finding and reaching desired location and than reading information from memory location. The information from locations can be accessed as follows:

1. Random access
2. Sequential access
3. Direct access

- Random Access: It is the access mode where each memory location has a unique address. Using these unique addresses each memory location can be addressed independently in any order in equal amount of time. Generally, main memories are random access memories(RAM).

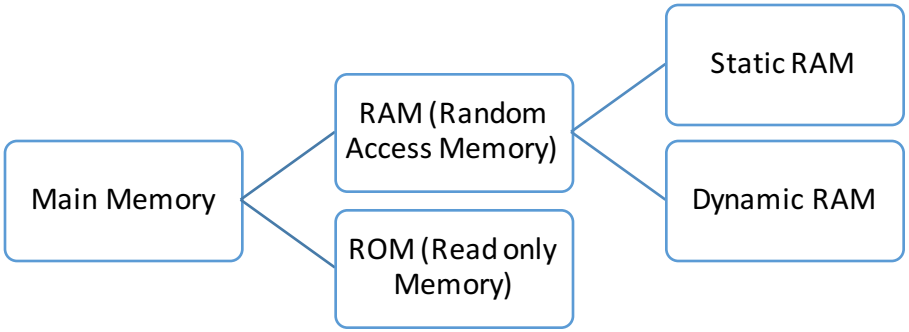
Access Methods

- **Sequential Access**: If storage locations can be accessed only in a certain predetermined sequence, the access method is known as serial or sequential access.
  - Opposite of RAM: **Serial Access Memory** (SAM). SAM works very well for memory **buffers**, where the data is normally stored in the order in which it will be used (a good example is the texture buffer memory on a video card , magnetic tapes, etc.).

- **Direct Access**: In this access information is stored on tracks and each track has a separate read/write head. This features makes it a semi random mode which is generally used in magnetic disks.

Main Memory

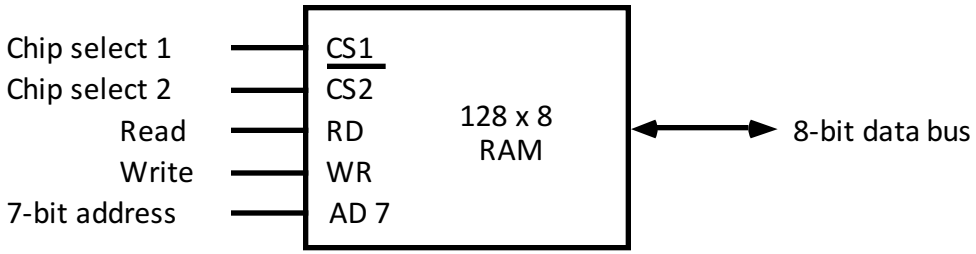
- Most of the main memory in a general purpose computer is made up of **RAM** integrated circuits chips, but a portion of the memory may be constructed with **ROM** chips.
- **RAM**– Random Access memory
  - Integrated RAM are available in two possible operating modes, **Static** and **Dynamic.**
- **ROM**– Read Only memory



Random Access Memory (RAM)

- RAM is used for storing bulk of programs and data that is subject to change.

Typical RAM chip



CS1	$\overline{CS2}$	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

Types of RAM

- Static RAM (SRAM)
  - Each cell stores bit with a six-transistor circuit.
  - Retains value indefinitely, as long as it is kept powered.
  - Faster (8-16 times faster) and more expensive (8-16 times more expensive as well) than DRAM.
- Dynamic RAM (DRAM)
  - Each cell stores bit with a capacitor and transistor.
  - Value must be refreshed every 10-100 ms.
  - Slower and cheaper than SRAM. Has reduced power consumption, and a large storage capacity.

  In contrast to , SRAM and DRAM:
    - Non Volatile RAM (NVRAM)
      - retains its information when power is turned off (non volatile).
      - best-known form of NVRAM memory today is flash memory.

  Virtually all desktop or server computers since 1975 used DRAMs for main memory and SRAMs for cache.

Read Only Memory (ROM)

- It is non-volatile memory, which retains the data even when power is removed from this memory. Programs and data that can not be altered are stored in ROM.

- ROM is used for storing programs that are **PERMANENTLY** resident in the computer and for tables of constants that do not change in value once the production of the computer is completed.

- The ROM portion of main memory is needed for storing an initial program called **bootstrap loader**, which is to start the computer operating system when power is turned on.

Auxiliary Memory

- Also called as Secondary Memory, used to store large chunks of data at a lesser cost per byte than a primary memory for backup.
- It does not lose the data when the device is powered down—it is non-volatile.
- It is not directly accessible by the CPU, they are accessed via the input/output channels.
- The most common form of auxiliary memory devices used in consumer systems is flash memory, optical discs, and magnetic disks, magnetic tapes.

Types of Auxiliary Memory

- Flash memory: An electronic non-volatile computer storage device that can be electrically erased and reprogrammed, and works without any moving parts. Examples of this are **USB flash drives** and **solid state drives**.

- Optical disc: Its a storage medium from which data is read and to which it is written by lasers. There are three basic types of optical disks: CD-ROM (read-only), WORM (write-once read-many) & EO (erasable optical disks).

Types of Auxiliary Memory

- Magnetic tapes: A magnetic tape consists of electric, mechanical and electronic components to provide the parts and control mechanism for a magnetic tape unit.

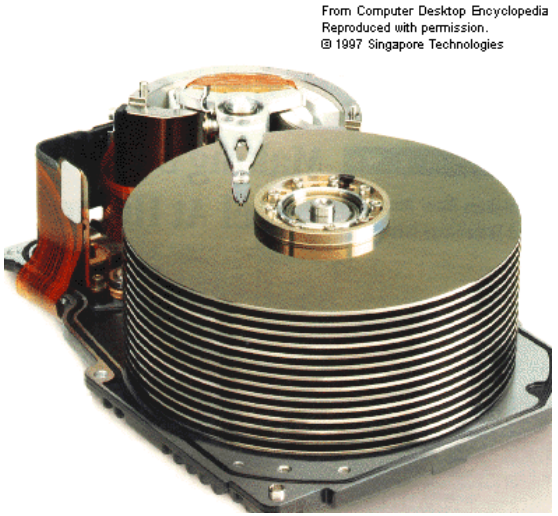
- The tape itself is a strip of plastic coated with a magnetic recording medium. Bits are recorded as magnetic spots on tape along several tracks called **RECORDS.**

- Each record on tape has an identification bit pattern at the beg. and the end.

Types of Auxiliary Memory

Magnetic Disk:

- A magnetic disk is a circular plate constructed of metal or plastic coated with magnetized material.
- Both sides of the disk are used and several disks may be stacked on one spindle with read/write heads available on each surface.
- Bits are stored in magnetized surface in spots along concentric circles called tracks. Tracks are commonly divided into sections called sectors.
- Disk that are permanently attached and cannot removed by occasional user are called hard disks.



Associative Memory

- A memory unit accessed by contents 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.

Read/Write operation in associative memory

- Write operation:
  - When a word is written in in an associative memory, no address is given.
  - The memory is capable of finding an unused location to store the word.

- Read operation:
  - When a word is to be read from an associative memory, the contents of the word, or a part of the word is specified.
  - The memory locates all the words which match the specified content and marks them for reading.

Disadvantage

- An associative memory is more expensive than a random access memory because each cell must have an extra storage capability as well as logic circuits for matching its content with an external argument.

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 cache memory
- The cache is the fastest component in the memory hierarchy and approaches the speed of CPU component

Basic Operations of Cache

- When CPU needs to access memory, the cache is examined.
- If the word is found in the cache, it is read from the cache 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.
- If the cache is full, then a block equivalent to the size of the used word is replaced according to the replacement algorithm being used.

Hit Ratio

- When the CPU refers to memory and finds the word in cache, it is said to produce a **hit**
- Otherwise, it is a **miss**

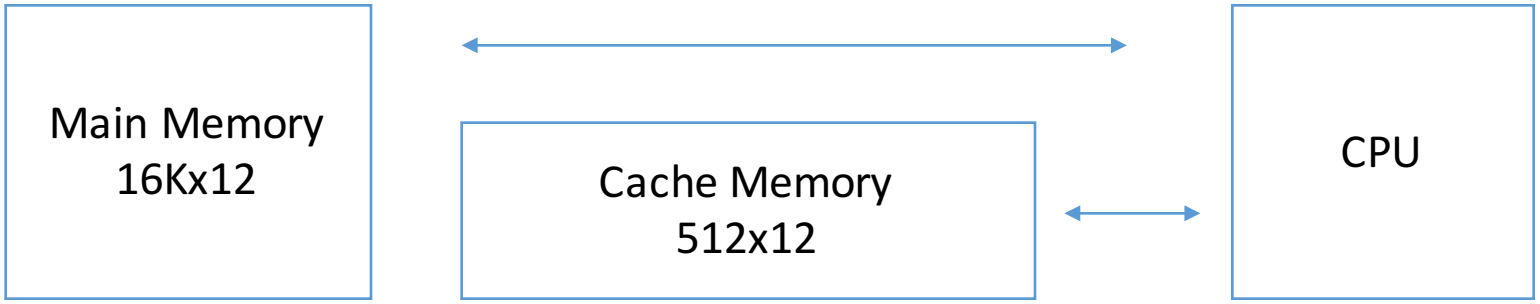
- The performance of cache memory is frequently measured in terms of a quantity called **hit ratio**

Hit ratio = hit / (hit+miss)

Mapping Process

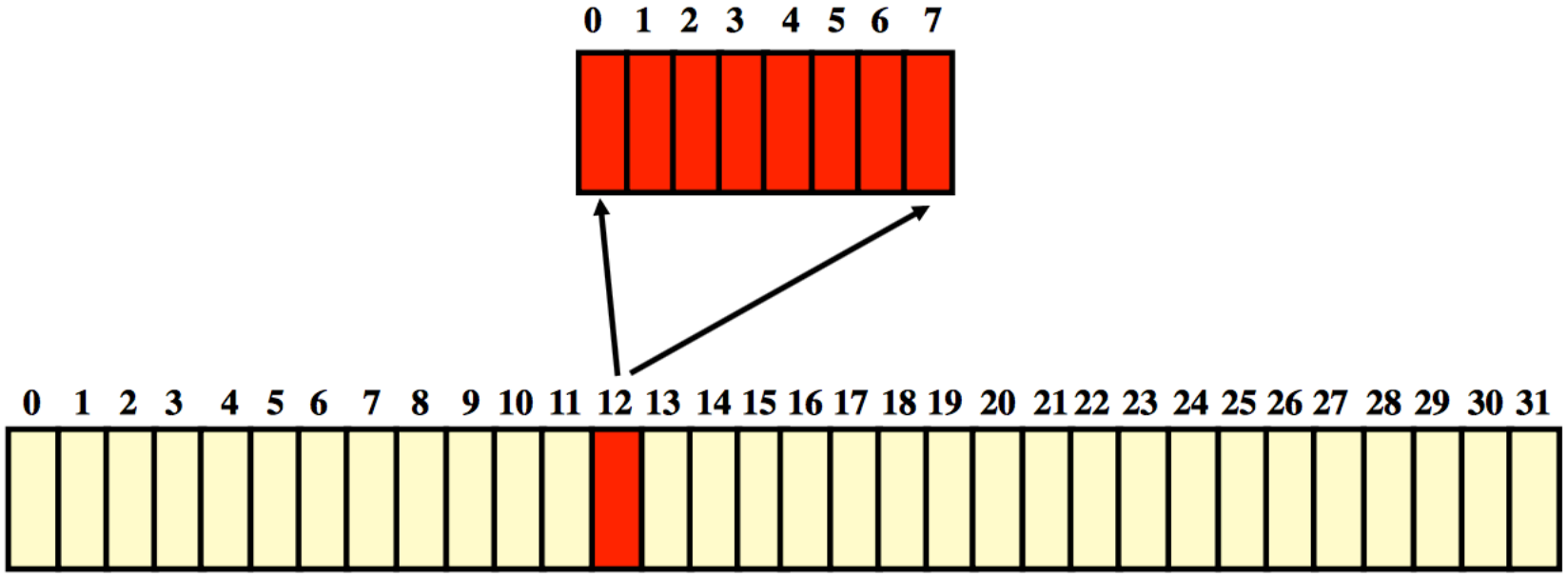
- The transformation of data from main memory to cache memory is referred to as a **mapping** process, there are three types of mapping:
  - Associative mapping
  - Direct mapping
  - Set-associative mapping

- To help understand the mapping procedure, we have the following example:



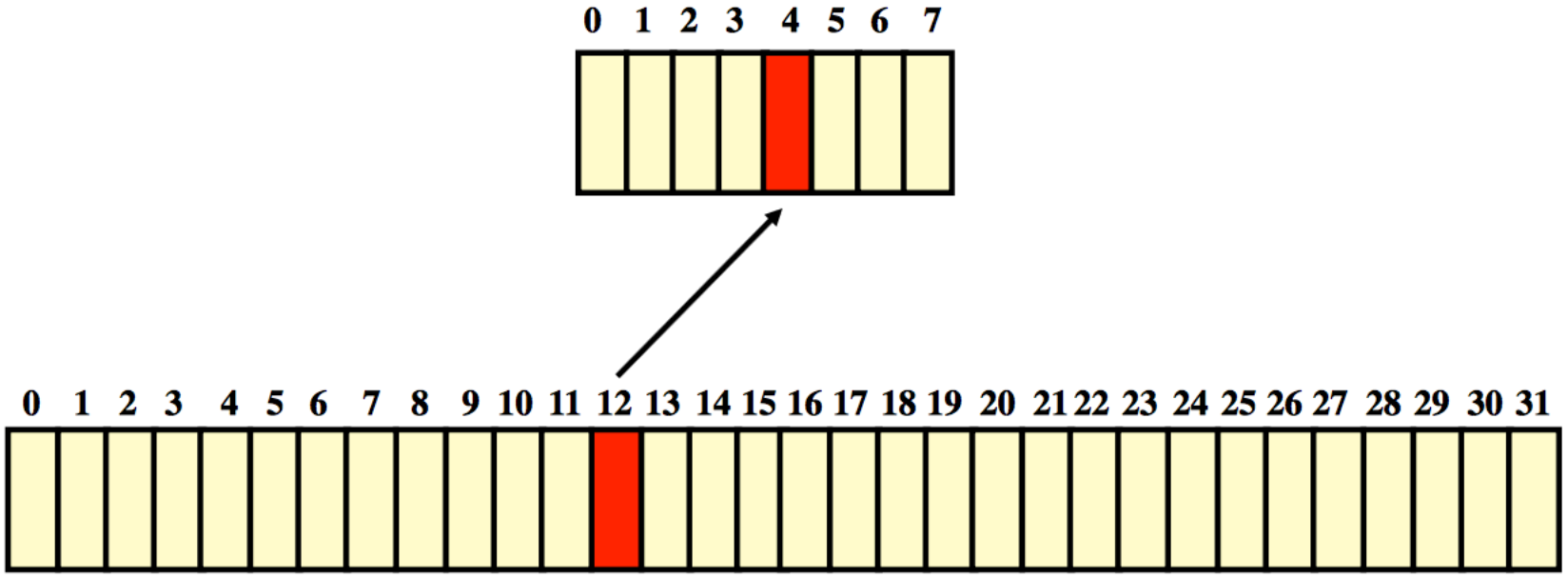
Associative Mapping

- Each block mapped to any cache location
  - any block from main memory can be placed any where in the cache. After being placed in the cache, a given block is identified uniquely by its main memory block number, referred to as the tag, which is stored inside a separate tag memory in the cache.



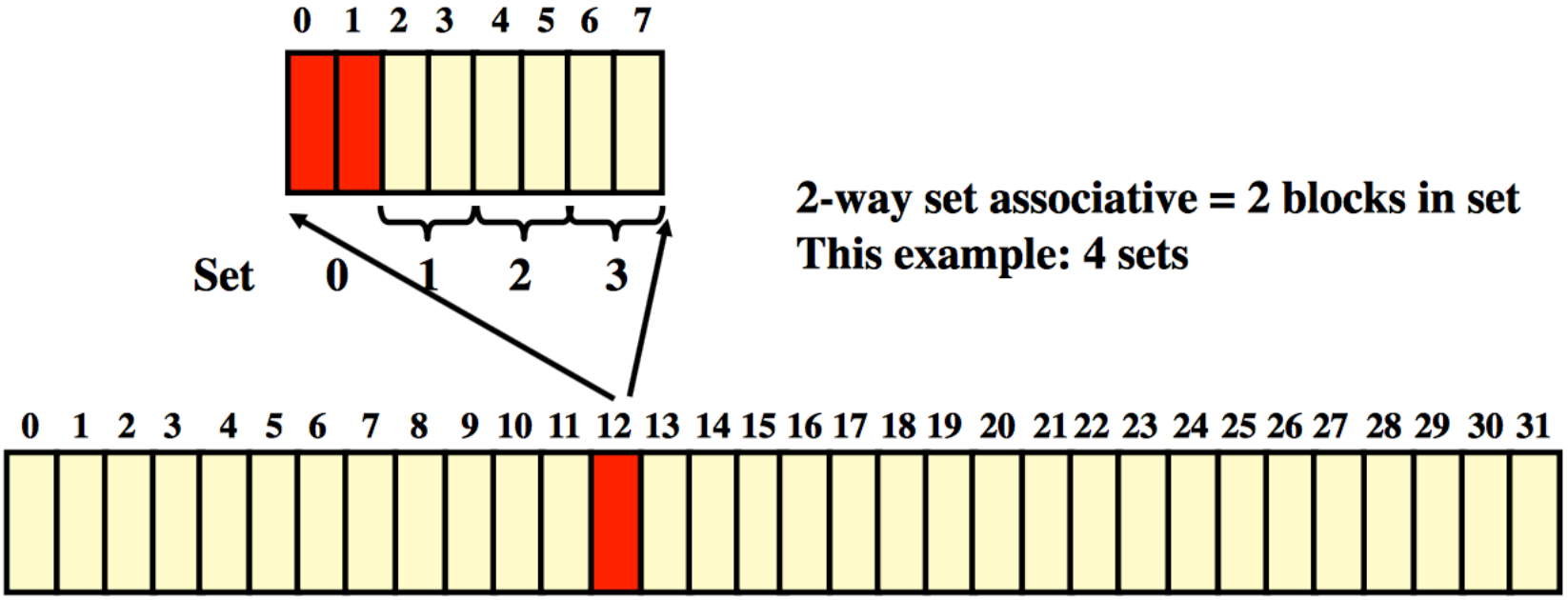
Direct Mapping

- Each memory block is mapped to exactly one block in the cache



Set – Associative Mapping

- Each block mapped to subset of cache locations
- Is a hybrid between a fully associative cache, and direct mapped cache



Replacement Algorithms

- Optimal replacement algorithm – find the block for replacement that has minimum chance to be referenced next time.

- Two algorithms:
  - FIFO: Selects the item which has been in the set the longest.
  - LRU: Selects the item which has been least recently used by the CPU.

FIFO (First In First Out)

- The first-in block in the cache is replaced first.
- In the other word, the block that is in the cache longest is replaced.
- Advantage: Easy to implement.
- Disadvantage: In some condition, blocks are replaced too frequently.

LRU (Least Recently Used)

- Replaced the least recently used block in the cache.
- To determine where is LRU block, a counter can be associated with each cache block.
- Advantage: This algorithm follows locality principle, so it limits number of times the block to be replaced.
- Disadvantage: Implementation is more complex.