Unit-8
Memory Organization

Marwadi University

Department of Computer Engineering

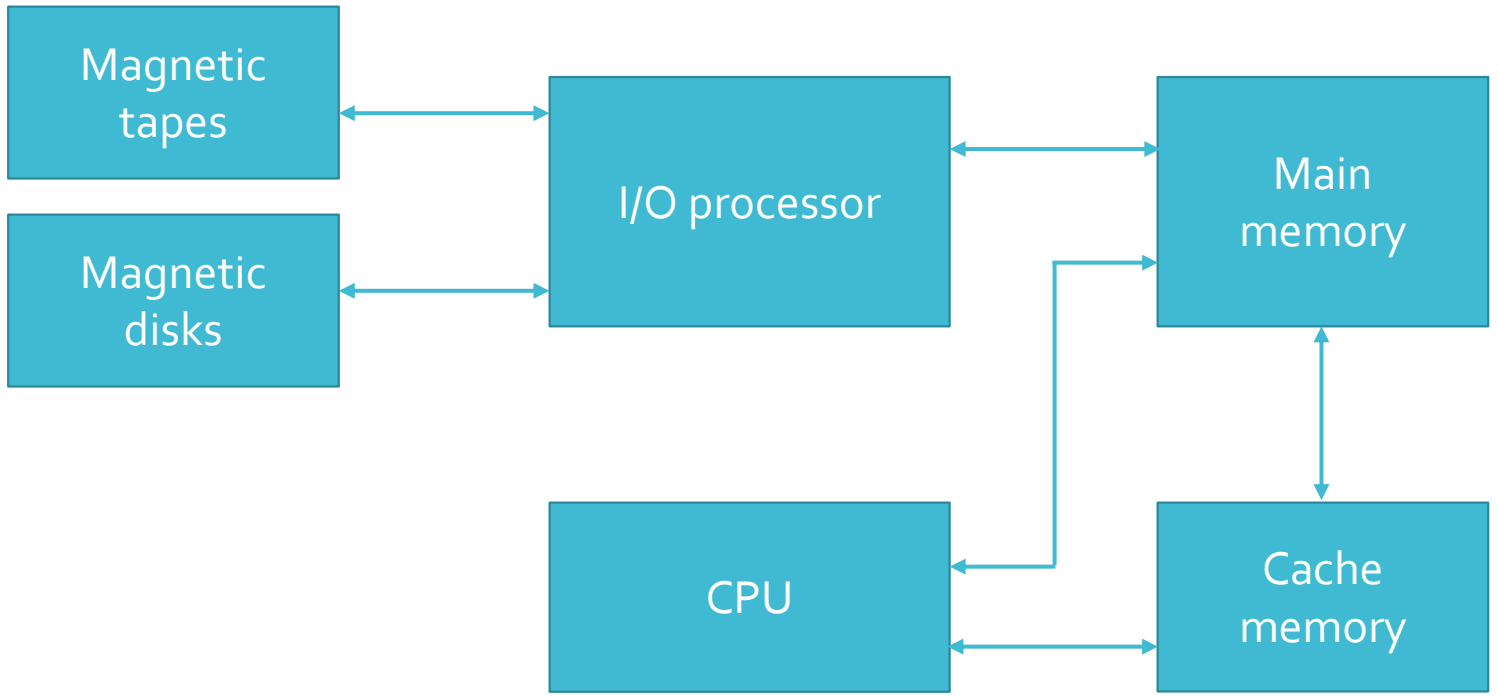
Computer Organization and Architecture
01CE1402

Prof. Kishan Makadiya

Memory Hierarchy

- The memory unit that communicates directly with CPU is called the main memory.
- Device that provides backup storage are called auxiliary memory.
- A special very-high-speed memory called a cache.
- It is sometimes used to increase the speed of processing by making current programs and data available to the CPU at a rapid rate.

Memory Hierarchy



Main Memory - RAM

- Used in computers for the temporary storage of programs and data.
- Read and write both operations are performed by RAM which requires fast cycle times as not to slow down the computer operation.
- It is volatile and lose all stored information if power is interrupted or turned off.
- RAMs typically come with word capacities of 1K, 4K, 8K, 16K, etc.. and word sizes of 1, 4 or 8-bits.
- It can be expanded by combining several memory chips.

Random access memory (RAM)

- RAM chips are used to communicate with the CPU.
- One chip selection line are not sufficient so we are using two chip selection line which decide when Read and write signal will active.
- It is having 7 bit address line so RAM size is of 128* 8 bit.
- A bidirectional bus can be constructed with help of three state buffer.it placed its output in one of the three possible state: signal equivalent to logic 1, signal equivalent to logic 0 or high impedance state.

SRAM v/s DRAM

Static RAM	Dynamic RAM
1. SRAM has lower access time, so it is faster compared to DRAM.	1. DRAM has higher access time, so it is slower than SRAM.
2. SRAM is costlier than DRAM.	2. DRAM costs less compared to SRAM.
3. SRAM requires constant power supply, which means this type of memory consumes more power.	3. DRAM offers reduced power consumption, due to the fact that the information is stored in the capacitor.
4. Due to complex internal circuitry, less storage capacity is available compared to the same physical size of DRAM memory chip.	4. Due to the small internal circuitry in the one-bit memory cell of DRAM, the large storage capacity is available.
5. SRAM has low packaging density.	5. DRAM has high packaging density.
6. No need to refresh periodically.	6. Due to capacitor used as storage element, information may lose over period of time. So, need to refresh periodically.
7. Uses an array of 6 transistors for each memory cell.	7. Uses a single transistor and capacitor for each memory cell.

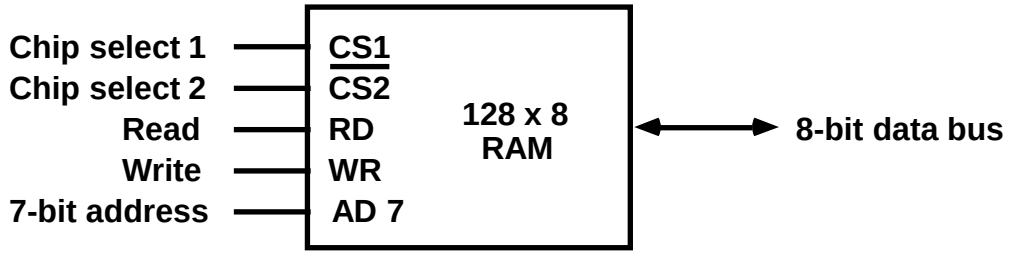
Read-Only Memory (ROM)

- A read-only memory (ROM) is essentially a memory device in which permanent binary information is stored.
- The binary information must be specified by the designer and is then embedded in the unit to form the required interconnection pattern.
- Once the pattern is established, it stays within the unit when the power is turned off and on again.
- A ROM which can be programmed is called a PROM. The process of entering information in a ROM is known as programming.
- ROMs are used to store information which is of fixed type, such as tables for various functions, fixed data and instructions.
- ROMs can be used for designing combinational logic circuits.

Main Memory

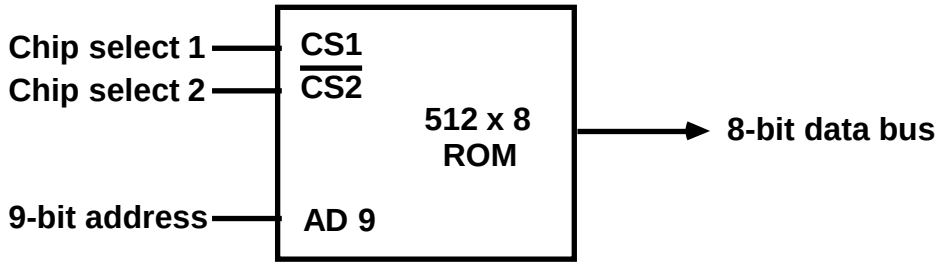
RAM and ROM Chips

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

Typical ROM chip



Memory Address Map

- Address space assignment to each memory chip
- Example: 512 bytes RAM and 512 bytes ROM

Component	Hexa address	Address bus									
		10	9	8	7	6	5	4	3	2	1
RAM 1	0000 - 007F	0	0	0	x	x	x	x	x	x	x
RAM 2	0080 - 00FF	0	0	1	x	x	x	x	x	x	x
RAM 3	0100 - 017F	0	1	0	x	x	x	x	x	x	x
RAM 4	0180 - 01FF	0	1	1	x	x	x	x	x	x	x
ROM	0200 - 03FF	1	x	x	x	x	x	x	x	x	x

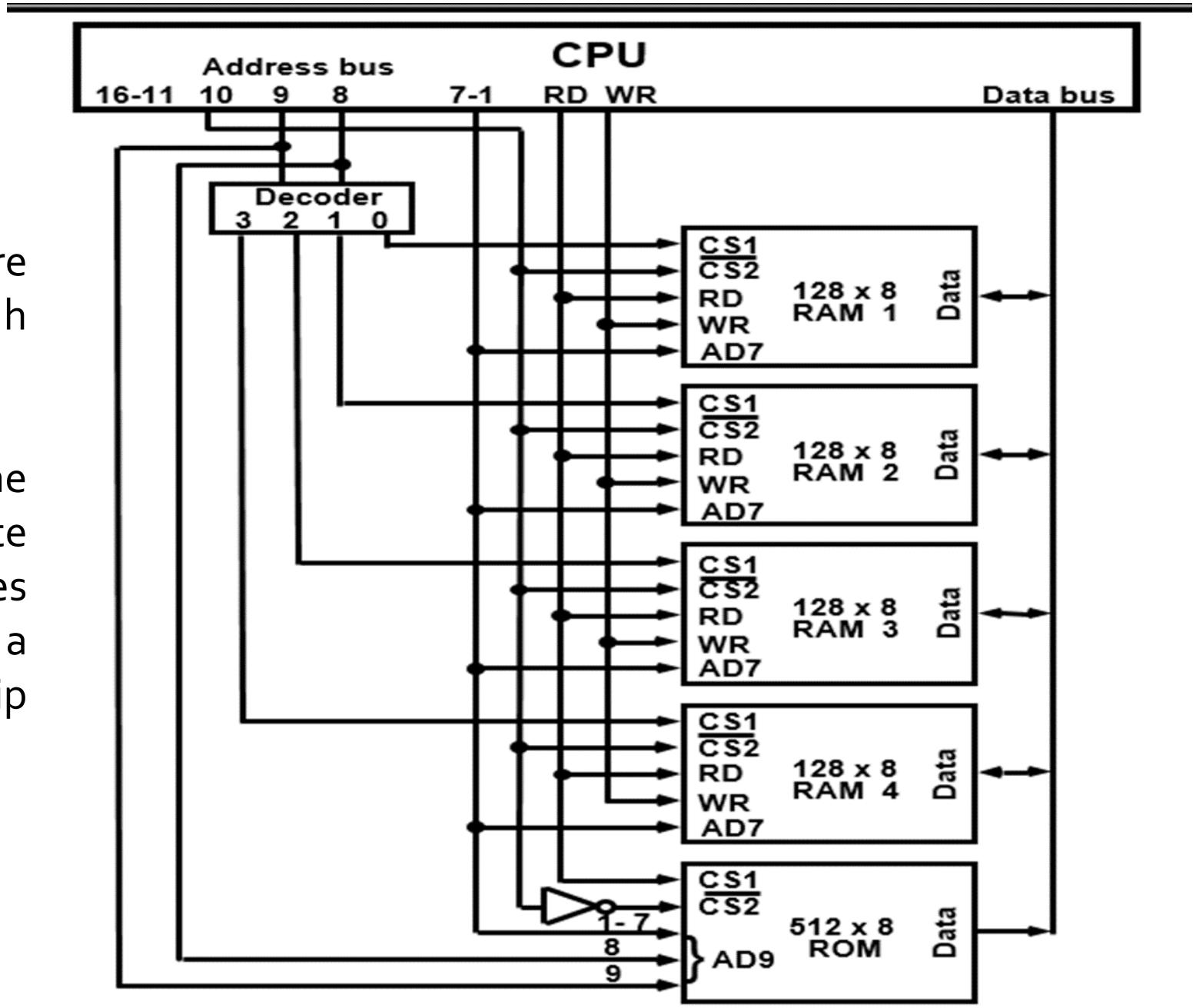
Memory Address Map

- The table called a memory address map, is a representation of assigned address space for each chip in the system.
- To demonstrate with a particular example, assume that a computer system needs 512 bytes of RAM and 512 bytes of ROM.
- The component column specifies whether a RAM or a ROM chip
- Hexadecimal address column specifies range of address.
- Here only we have specified 10 lines as other 6 lines are assumed to be a zero.
- To distinguish between four RAM chips we have choose line no 8 and 9 for various binary combination.

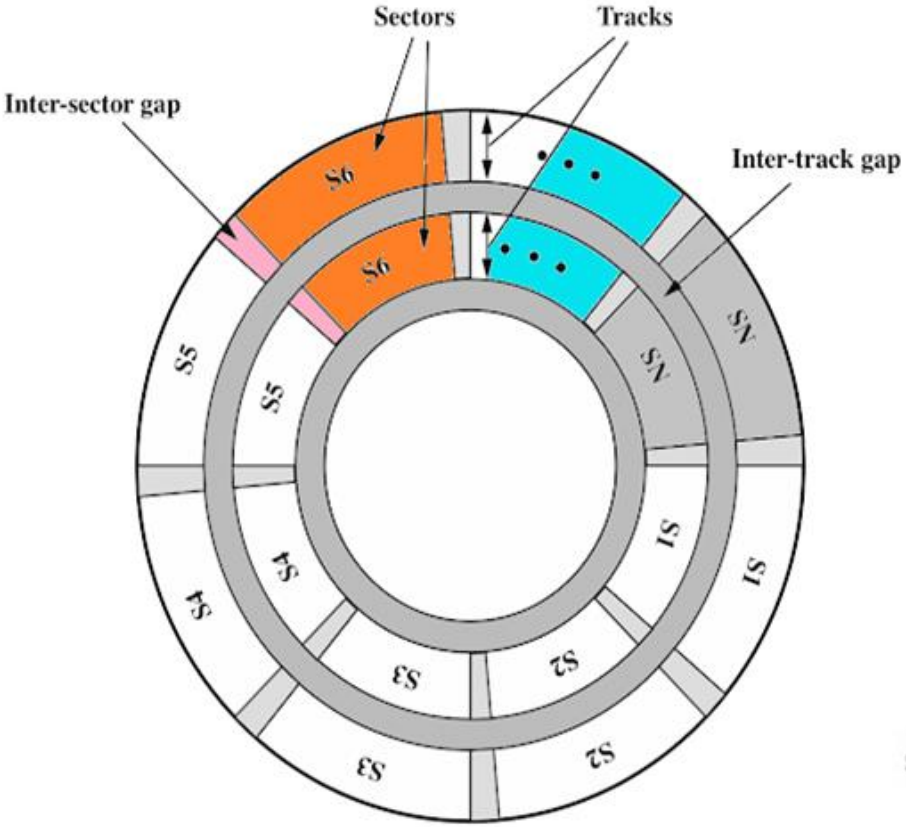
Memory Connection to CPU

- RAM and ROM chips are connected to a CPU through the data and address buses

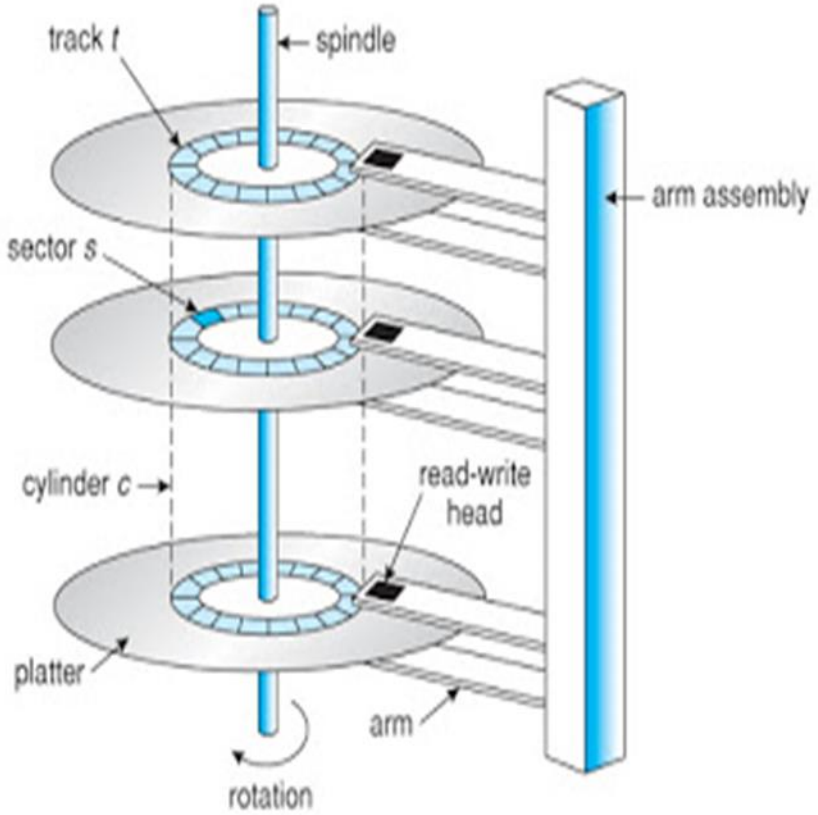
- The low-order lines in the address bus select the byte within the chips and other lines in the address bus select a particular chip through its chip select inputs



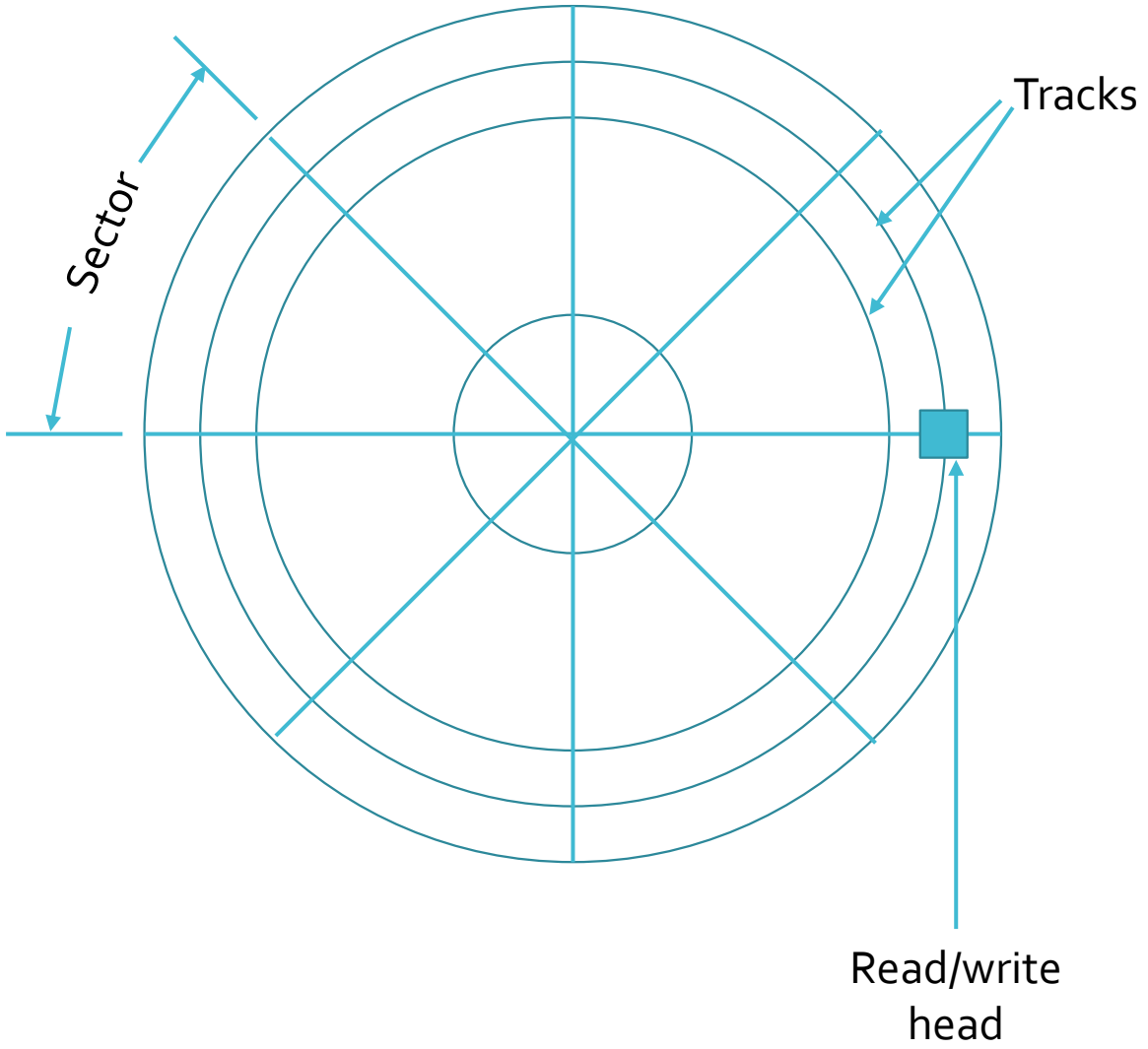
Auxiliary Memory



Magnetic Disks



Magnetic Disks



- Circular plate 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 for access purposes.
- Bits are stored in the magnetized surface in spots along concentric circles called track.
- The tracks are commonly divided into sections called sectors.

Magnetic Disks

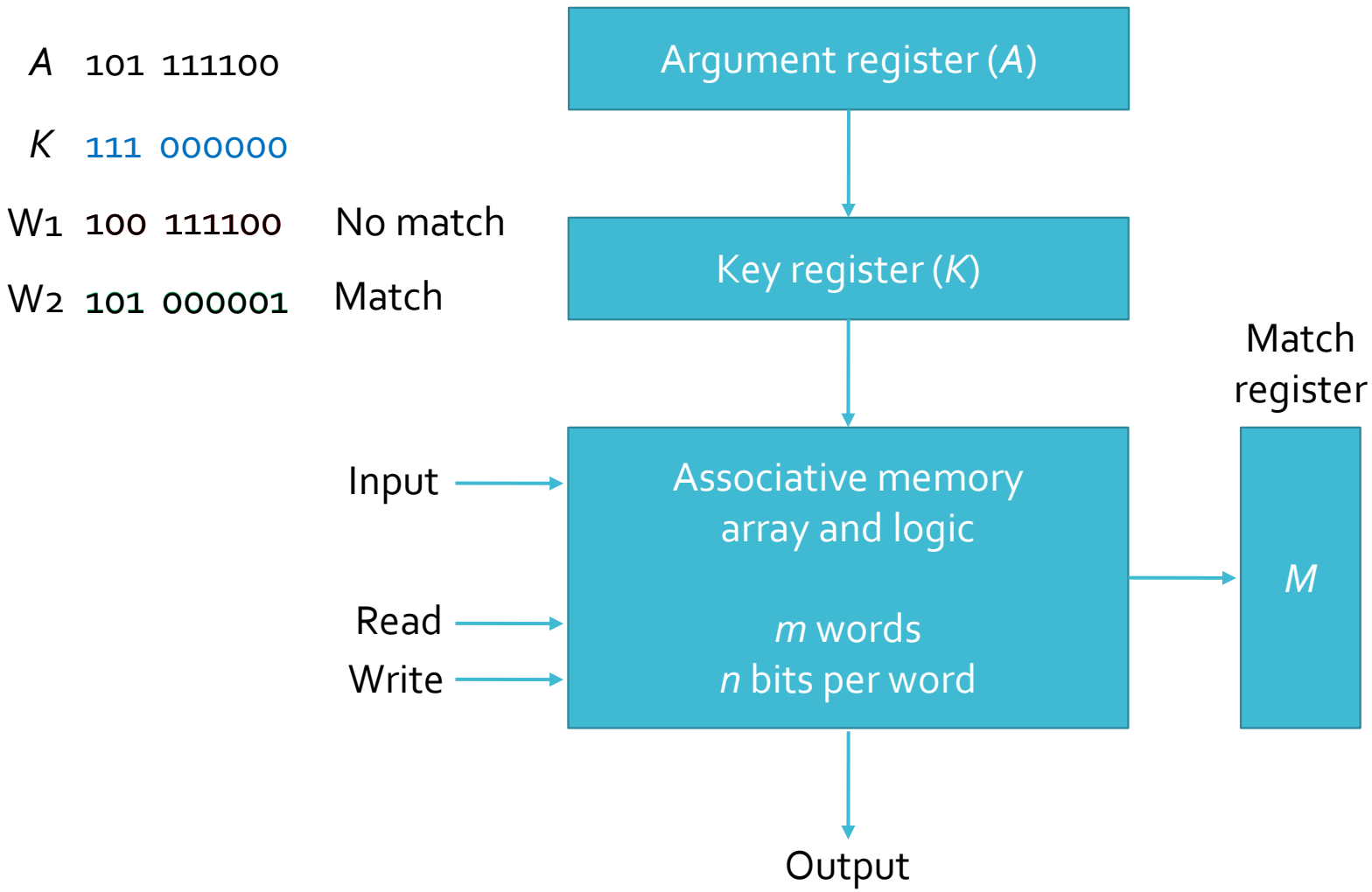
- Primary computer storage device.
- One or more platters in the form of disks covered with magnetic media.
- Each disk platter has a flat circular shape.
- Each platter has two working surfaces that stores data.
- Digital information is stored on magnetic disks in the form of microscopically small, magnetized needles.
- Data is stored on either or both surfaces of discs in concentric rings called tracks.
- Each track is divided into a number of sectors.
- To read information, the arm is positioned over the correct track.
- Data is read and written by a disk drive which rotates the discs and positions the read/write heads over the desired track

Magnetic Tape

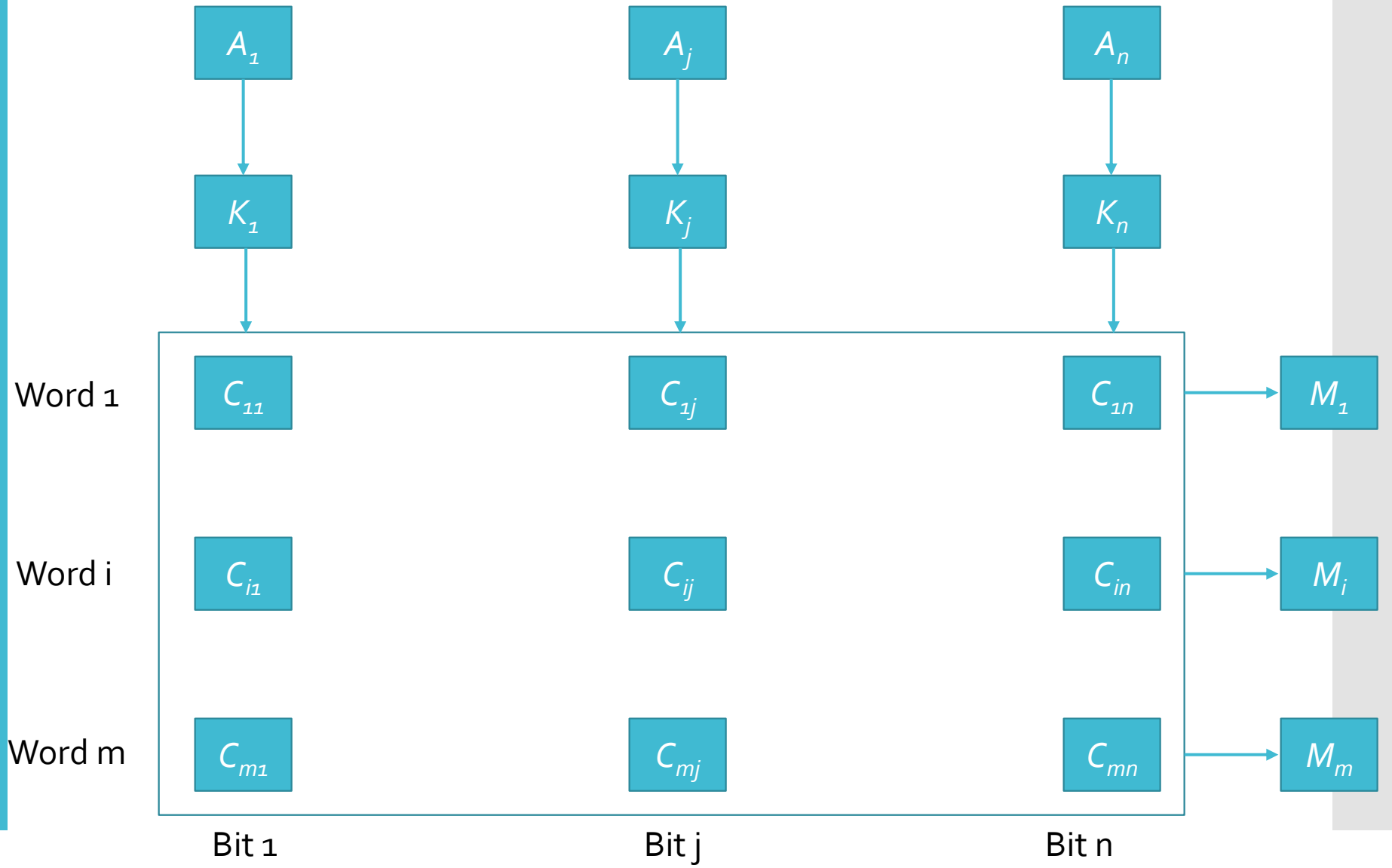
- Magnetic tape transport consists of the electrical, mechanical and electronic components to provide the parts and control mechanism for a magnetic-tape unit.
- The tape itself is a strip of plastic coated with a magnetic recording medium.
- Bits are recorded as magnetic spots on the tape along several tracks.
- 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.
- A tape unit is addressed by specifying the record number and the number of characters in the record.

Associative Memory (Content Addressable 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.



Associative Memory



Cache Memory

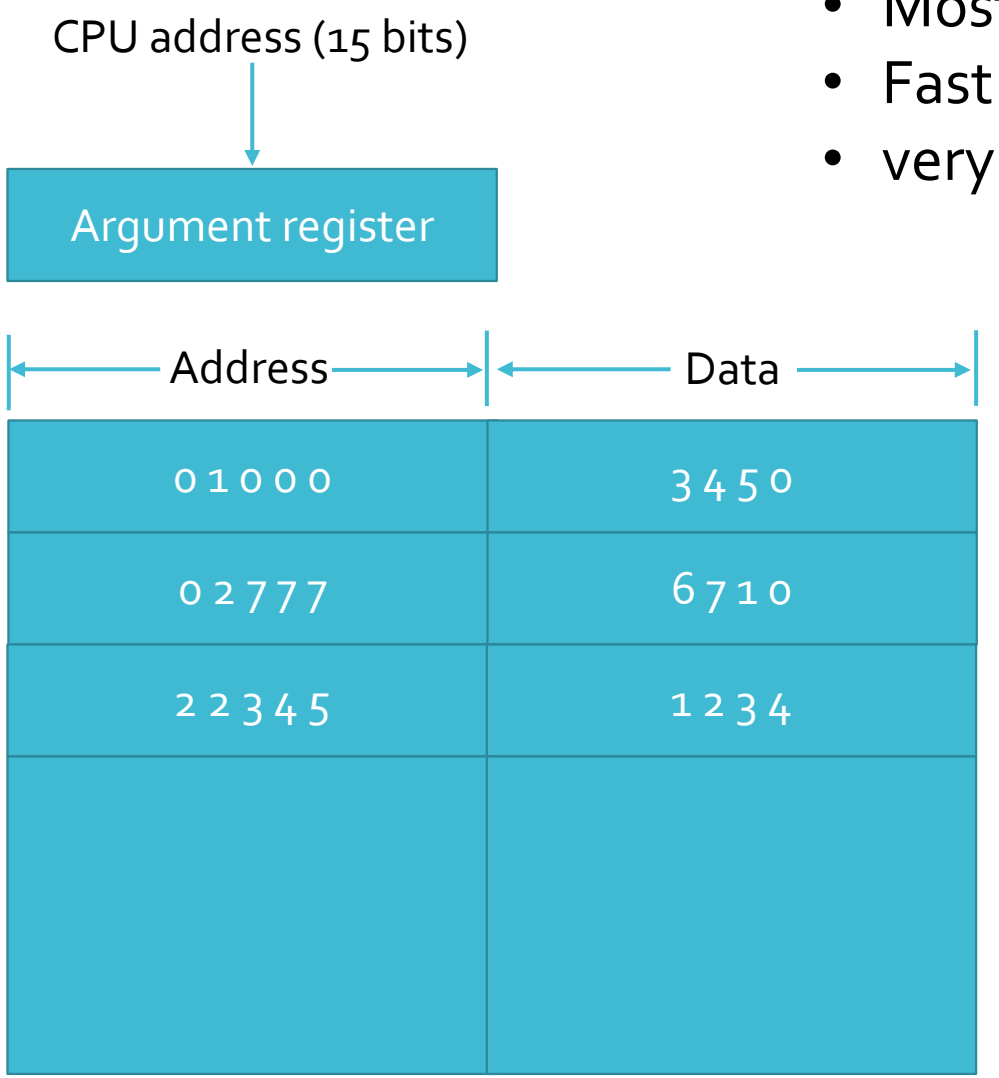
- Cache is a fast small capacity memory that should hold those information which are most likely to be accessed.
- The basic operation of the cache is, when the CPU needs to access memory, the cache is examined.
- If the word is found in the cache, it is read from the fast memory. If the word addressed by the CPU is not found in the cache, the main memory is accessed to read the word.
- The transformation of data from main memory to cache memory is referred to as a mapping process.

Cache Memory

- 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 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*.
- Hit ratios of 0.9 and higher have been reported.

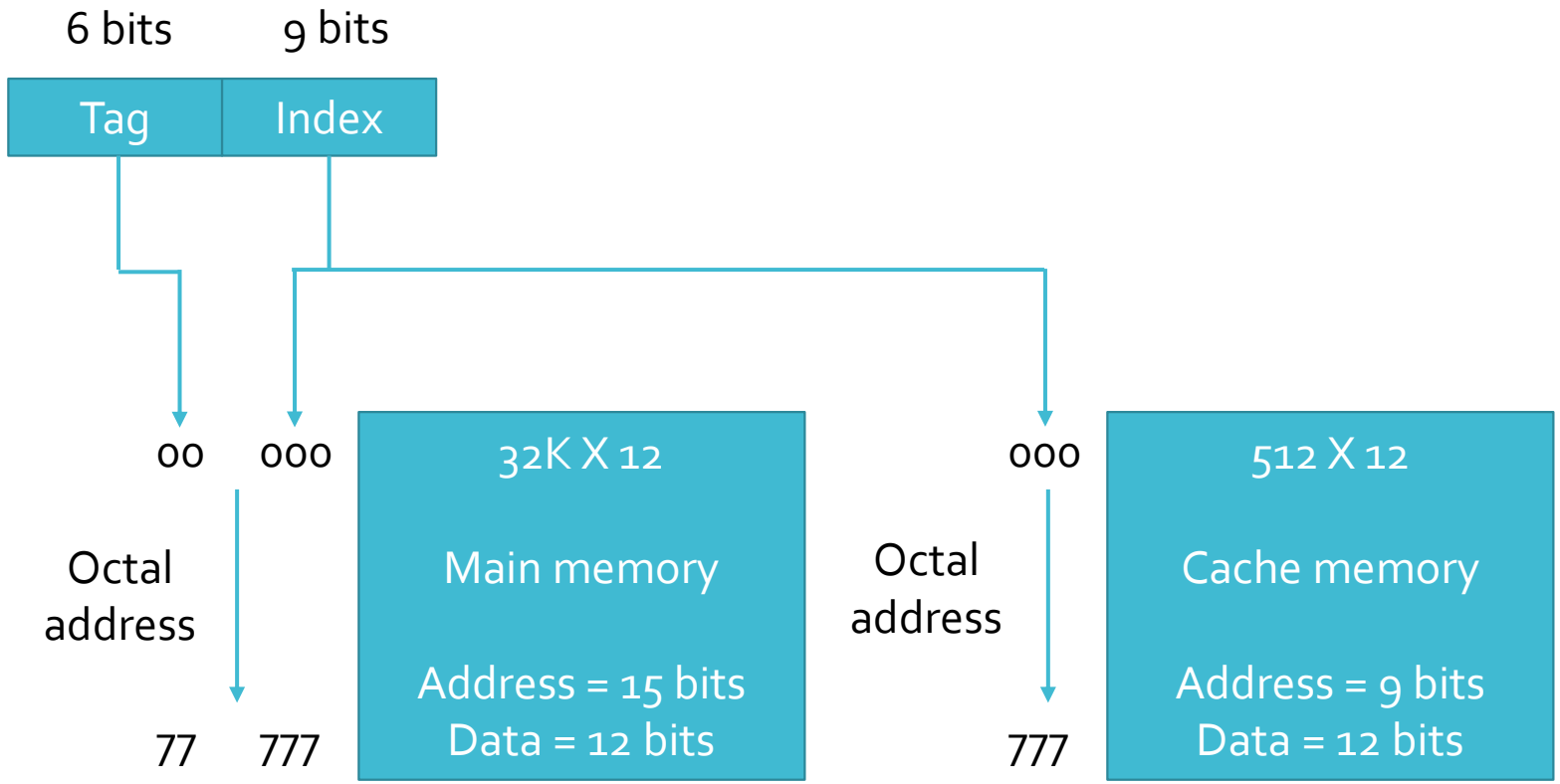
Associative Mapping

- Mapping Table stores both address and the content of the memory word
  - Most flexible
  - Fast
  - very Expensive

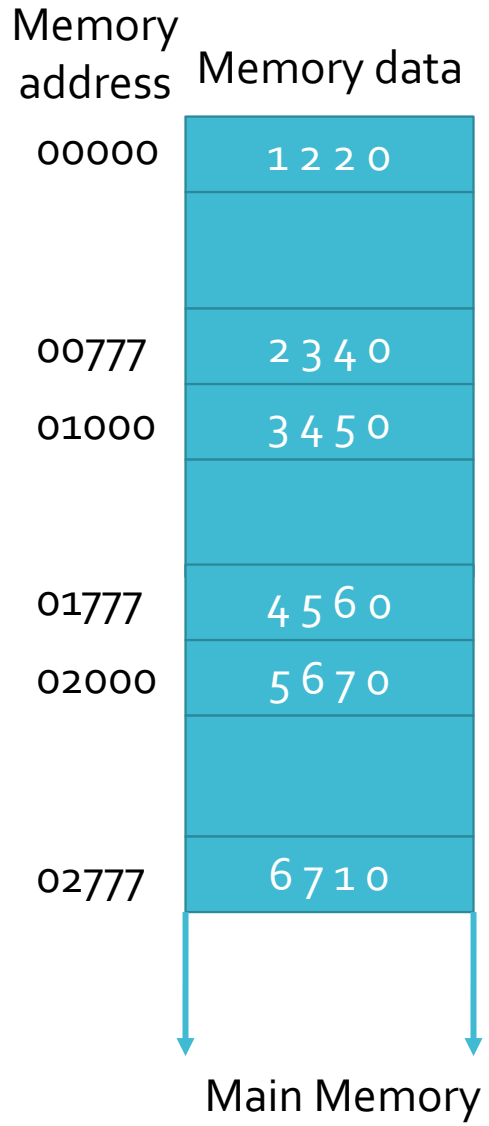


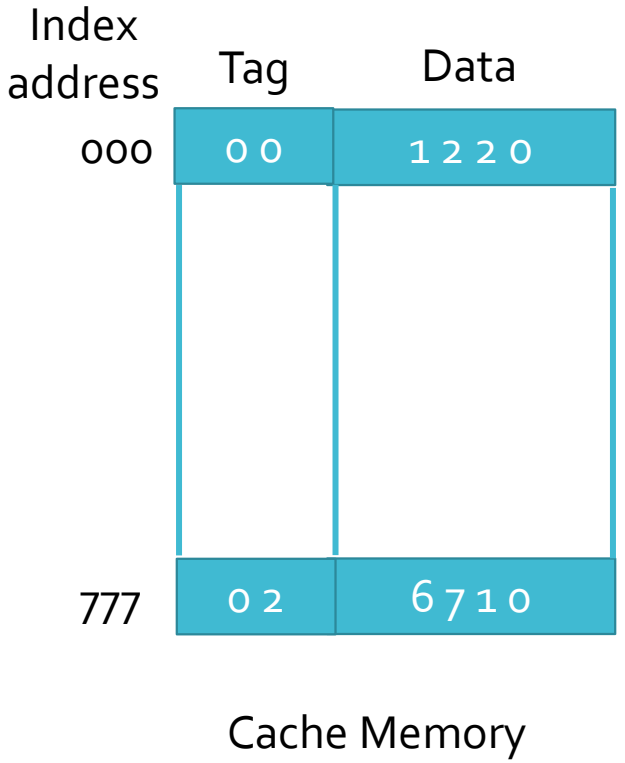
Direct Mapping

- Each memory block has only one place to load in Cache
- n-bit memory address consists of 2 parts:
  - k bits of Index field and
  - n-k bits of Tag field
- n-bit addresses are used to access main memory and k-bit Index is used to access the Cache



Direct Mapping





Direct Mapping Cache Organization

Set-Associative Mapping

Operation

- CPU generates a memory address (TAG; INDEX)
- Access Cache with INDEX, (Cache word = (tag 0, data 0); (tag 1, data 1))
- Compare TAG and tag 0 and then tag 1
- If tag i = TAG -> Hit, CPU <- data i
- If tag i ≠ TAG -> Miss, Replace either (tag 0, data 0) or (tag 1, data 1)

Virtual Memory

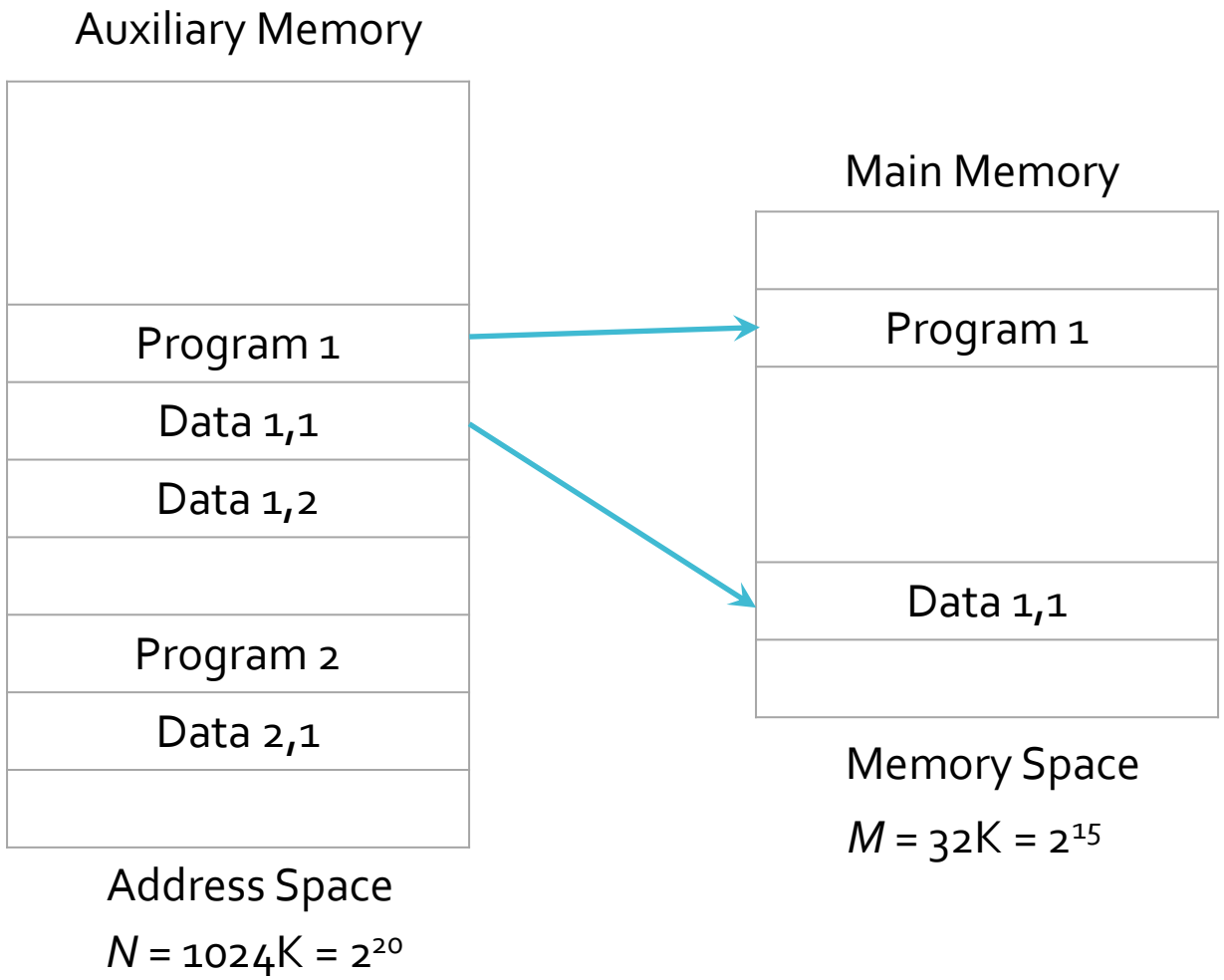
- Virtual memory is used to give programmers the illusion that they have a very large memory at their disposal, even though the computer actually has a relatively small main memory.
- A virtual memory system provides a mechanism for translating program-generated addresses into correct main memory locations.
- **Address space**

  An address used by a programmer will be called a virtual address, and the set of such addresses is known as address space.
- **Memory space**

  An address in main memory is called a location or physical address. The set of such locations is called the memory space.

Virtual Memory

- Relation between Address space & Memory space



Virtual Memory