काशी हिन्दू विश्वविद्यालय

BANARAS HINDU UNIVERSITY

Memory Organization

Guided By - Dr. Sarvesh Pandey

Presented By -

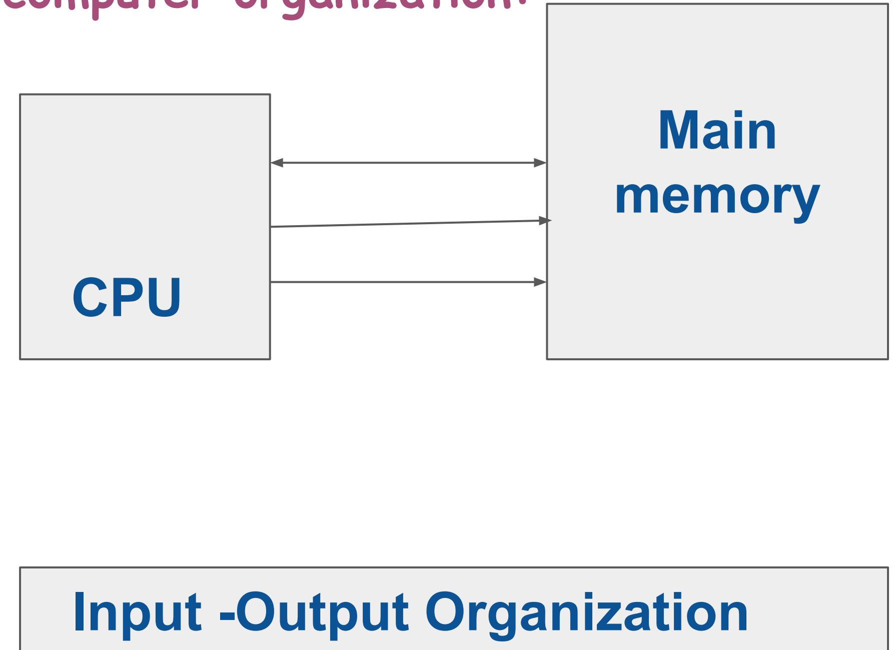
1. Anushka Rai - 20229PHY004
2. Vaishnavi Kushwaha - 20229MAT007
3. Nivedita Pandey - 20229CMP004
4. Pragati Agrawal - 20229MAT002
5. Ritu Kumari - 20229MAT003

Memory Organisation

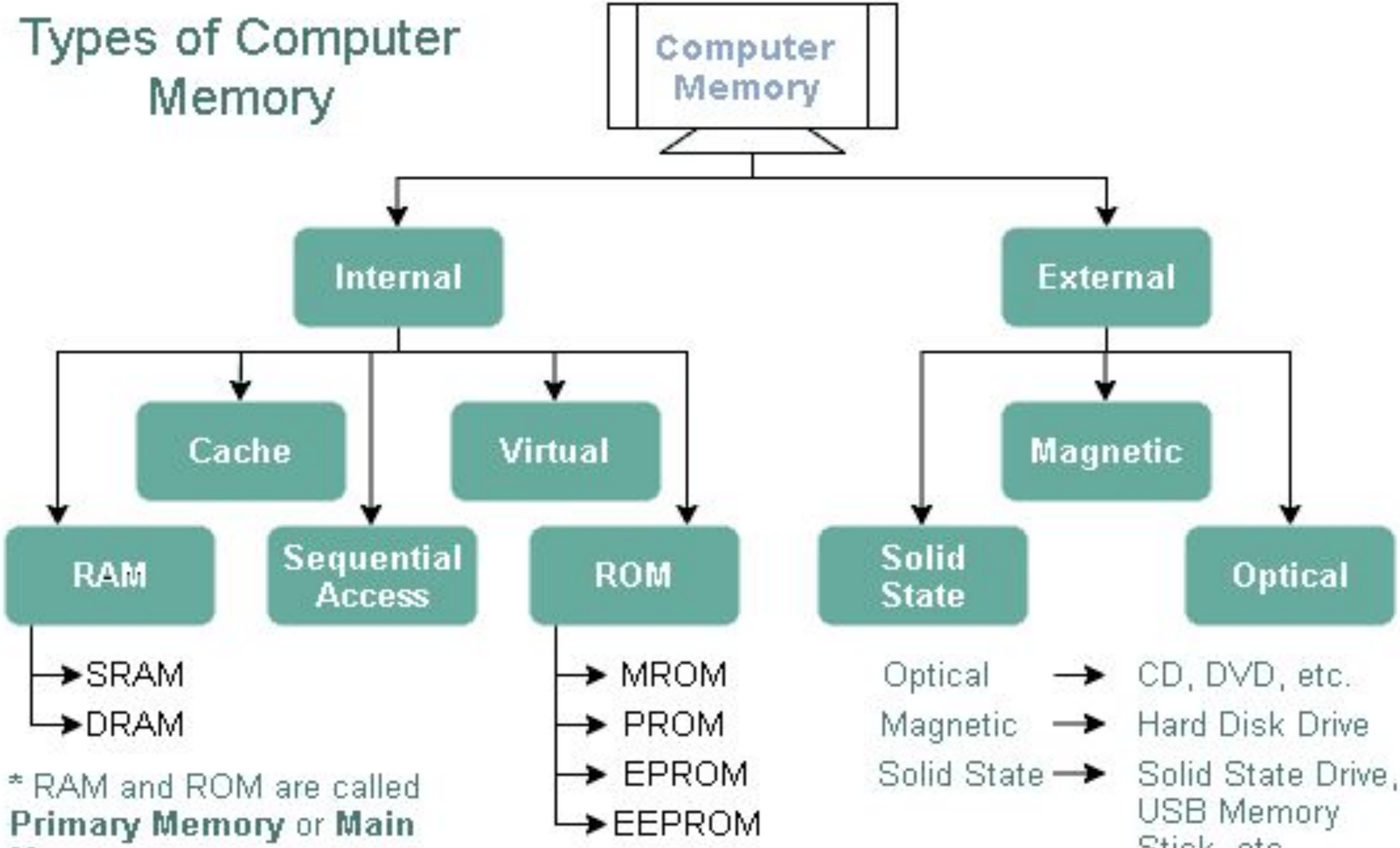
1. Memory Hierarchy
2. Cache Memory
3. Main Memory
4. Secondary Memory
5. Virtual Memory
6. Characteristics of different types of

Introduction

- In computing, memory is a device or system that is used to store information (data or program) for immediate use in a computer or related hardware and digital electronic devices.
- It is one of the main components of computer organization:-
    - CPU
    - Memory
    - Input Output Organization



Types of Computer Memory



Computer Memory

- Internal
  - RAM
    - SRAM
    - DRAM
  - Cache
  - Sequential Access
  - Virtual
  - ROM
    - MROM
    - PROM
    - EPROM
    - EEPROM
- External
  - Solid State
  - Magnetic
  - Optical

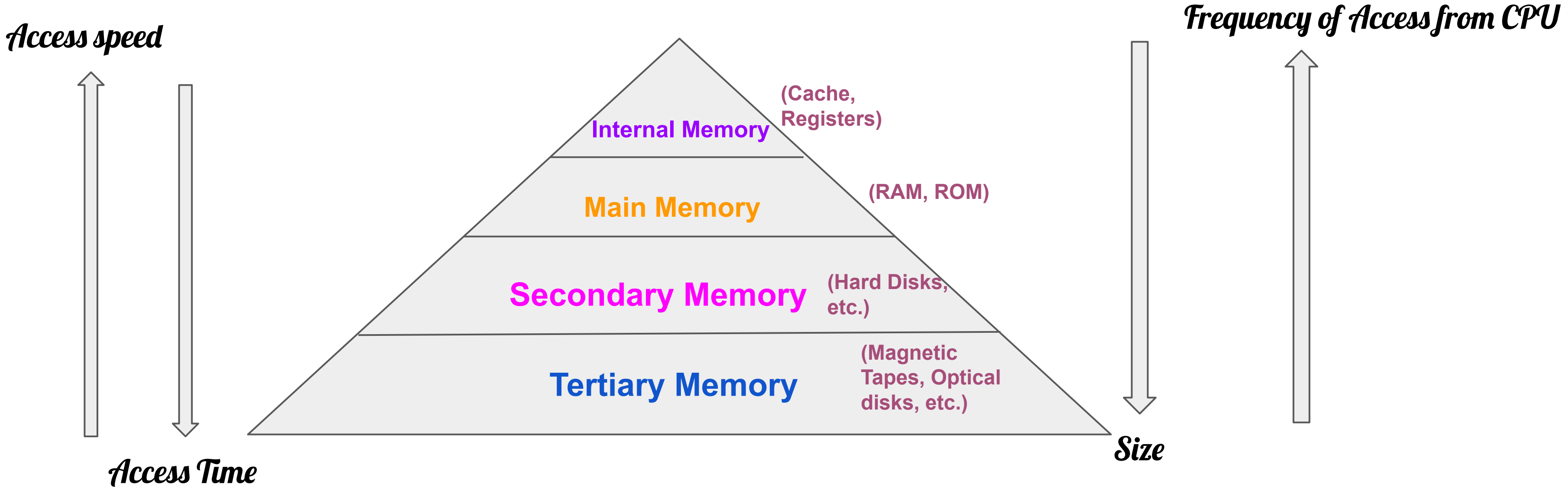
Optical	→	CD, DVD, etc.
Magnetic	→	Hard Disk Drive
Solid State	→	Solid State Drive, USB Memory Stick, etc.

* RAM and ROM are called **Primary Memory** or **Main Memory**..

Memory Hierarchy

What is Memory Hierarchy?

The Memory in a computer is divided into four hierarchies based on the speed as well as use. The processor moves from one level to another based on its requirements. The four hierarchies in the memory are **internal memory, main memory, secondary memory, and tertiary memory**.



Memory Units in Computer System:

1. Cache memory temporarily stores frequently used instructions and data for quicker processing by CPU

2. Primary memory is the computer memory that is directly accessible by CPU. It is comprised of DRAM. It holds the data and instructions that the processor is currently working on.

3. Secondary memory is the memory with high storage capacity and data and programs are not lost from it when system is turned off.

4. Tertiary storage comprises high-capacity data archives designed to incorporate vast numbers of removable media, such as tapes or optical discs

Why there is a need of Memory Hierarchy?

CPUs have always been faster than memories

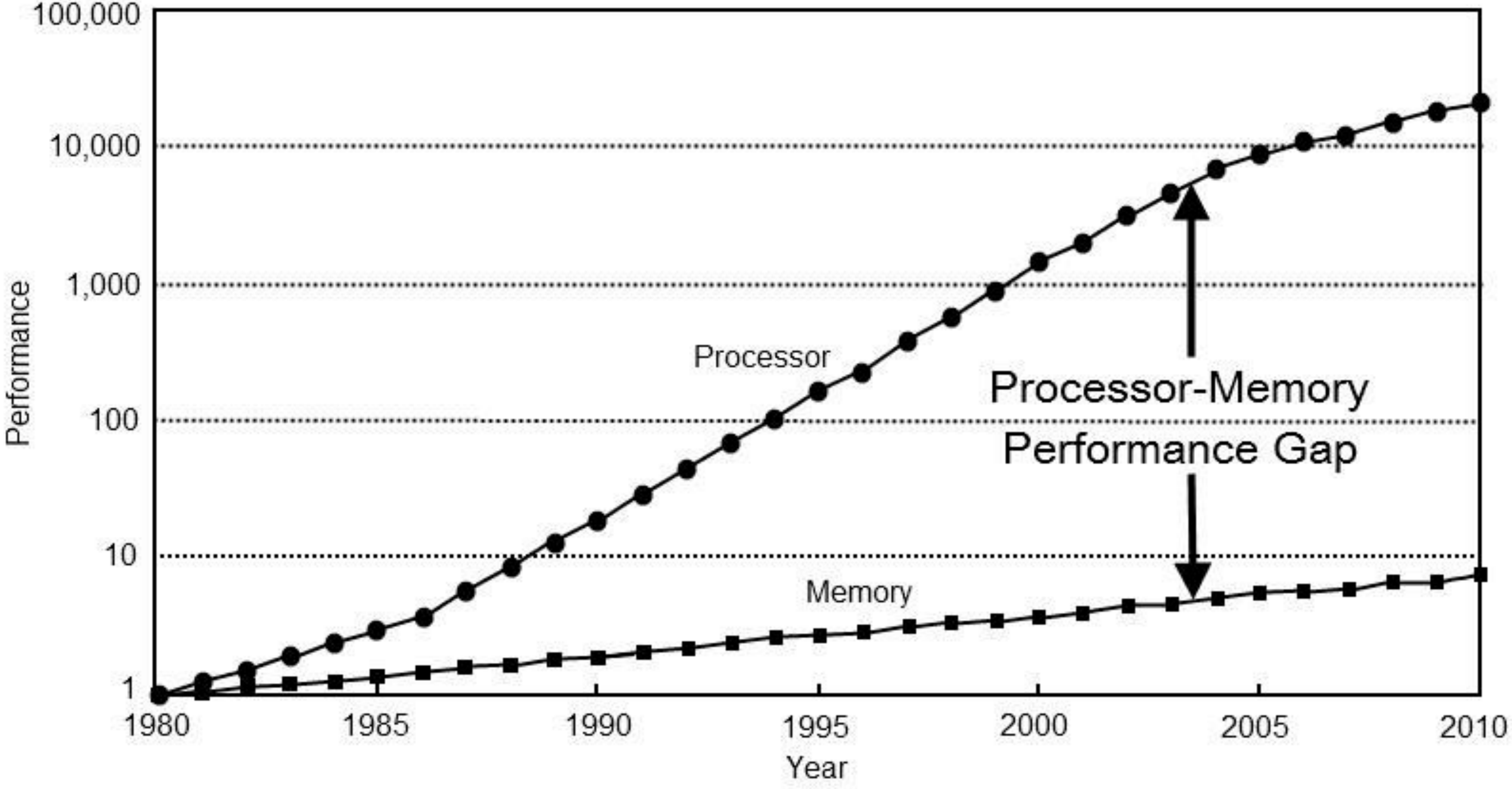
As it has become possible to place more and more circuits on a chip over time, CPU has got even faster.

On the other hand, new technology is being used to increase the capacity of memory instead of improving the speed.

This factor has caused the increased gap between main memory and CPU

Hence when a CPU issues a command that requests memory access, it does not get the memory unit it wants right away. Instead, it has to wait for some number of CPU cycles for the memory to serve the request. Thus reducing the system's performance.

Hence, to bridge this gap scientists have introduced a smaller and faster memory unit between CPU and RAM termed as *Cache Memory*. Thus making a hierarchy of memory.


100,000
10,000
1,000
100
10
1
Performance
1980
1985
1990
1995
2000
2005
2010
Year
Processor
Memory
Processor-Memory
Performance Gap

Arrangement of Memory Units in System:

- At the bottom, there are cheap storage devices (Tertiary Memory) with large amount of memory, like the optical disks or the magnetic tape. Their access time is quite large and access speed is quite slow.
- A level up, there is Secondary memory. It has much larger storage and lesser cost as compared to primary memory. Example - Hard Disks. Their access speed is slow and access time is large.
- The access time of secondary memory is usually 1000 times that of main memory.
- Higher up, there is a Primary memory {RAM and ROM}, which has medium capacity and speed. Their access speed is faster than secondary memory.
- At the top rests cache and registers, both of which are very fast but has small capacities. Their access speed is fastest.
- The access time between cache memory and main memory is about 1 is to 7~10

CACHE MEMORY

Cache Memory is a special very high speed memory.It is used to speed up and synchronizing with high speed CPU.

Cache memory is costlier than main memory but economical than CPU registers.

Cache memory is an extremely fast memory type that acts as a buffer between RAM and the CPU.

It holds frequently requested data and instructions so that they are immediately available to the CPU when needed.

Cache memory is used to reduce the average time to access data from the Main memory.

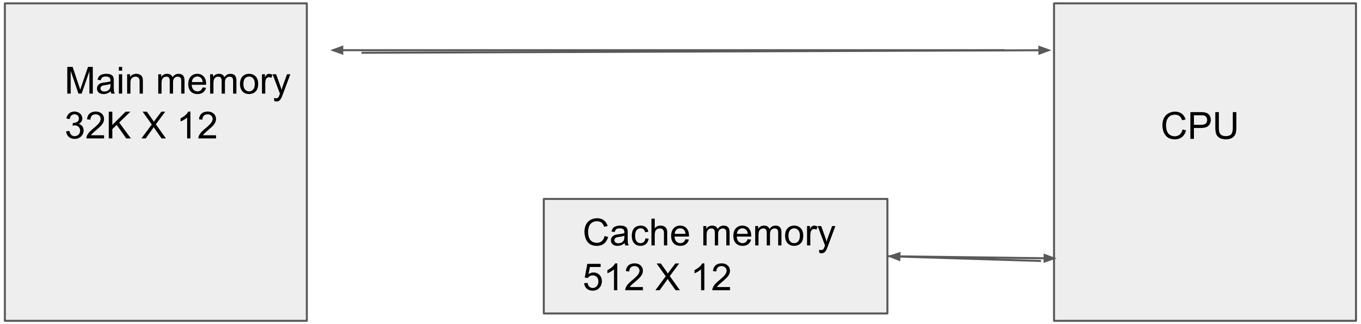
The cache is a smaller and faster memory which stores copies of the data from frequently used main memory locations.

When the CPU need to access the memory it first search in cache.If found it is read.

If the word is not found it is read from main memory and a block of data is transferred from main memory to cache which contains the word.

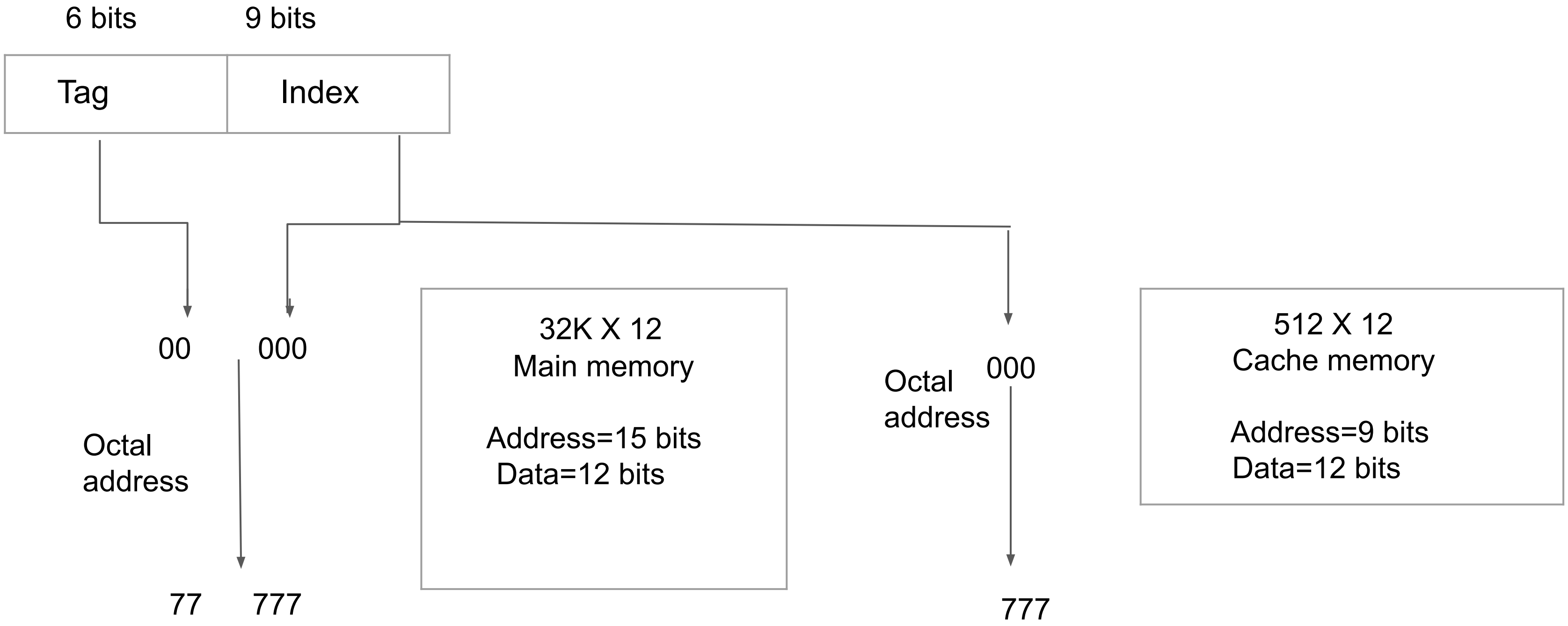
If the word is found in cache,it is said hit. If the word is not found it is called miss.

Performance of cache is measured in terms of hit ratio which is ratio of total hit to total memory access by CPU.



Example of cache memory

Addressing relationships between main and cache memories



Mapping Techniques

The transformation of data from main memory to cache is known as mapping process.Three types of mapping process are:

1.Associative mapping

2.Direct mapping

3.Set-associative mapping

Associative Mapping

Fastest and most flexible cache organization uses associative memory.

It stores both address and content of memory word.

Address is placed in argument registers and memory is searched for matching address.

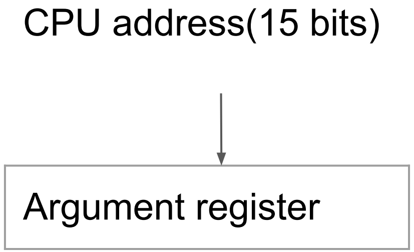
If address is found corresponding data is read.

If address is not found, it is read from main memory and transferred to cache.

If the cache is full,an address-word pair must be displayed.

Various algorithm are used to determine which pair to displace.some of them are FIFO(First In First Out),LRU(Least Recently Used)etc.

Associative mapping cache



CPU address(15 bits)

Argument register

Address	Data
01000	3450
02777	6710
22345	1234

Direct Mapping

CPU address is divided into two fields tag and index.

Index field is required to access cache memory and total address is used to access main memory.

If there are 2^k words in cache and 2^n words in main memory, then n bit memory address is divided into two parts k bits for index field and n-k bits for tag field.

When CPU generates memory request,index field is used to access the cache.

Tag field of the CPU address is compared with the tag in the word read.If the tag match, there is a hit.

If the tag does not match ,word is read from main memory and updated in the cache.

This example uses the block size of 1.

It can be also implemented for block size of 8 words.

The index field is divided into two parts: block field and word field.

In 512 word cache there are 64 blocks of 8 words each

Every time miss occur entire block of word is transferred from main memory to cache.

Direct mapping cache organization

Memory address	Memory data
00000	1220
00777	2340
01000	3450
01777	4560
02000	5670
02777	6710

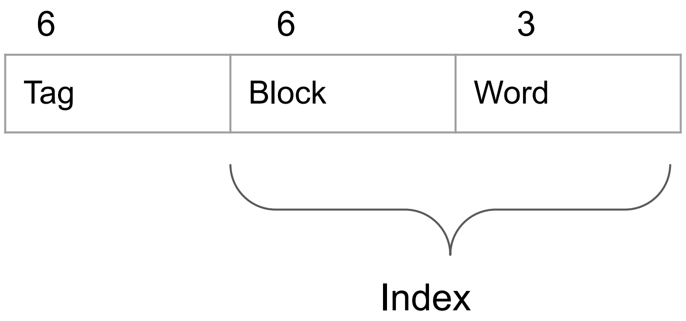
(a) Main memory

Index Address	Tag	Data
000	00	1220
777	02	6710

(b) Cache memory

Direct mapping cache with block size of 8 words

	Index	Tag	Data
Block 0	000	01	3450
	007	01	6758
Block 1	010		
	017		
Block 63	770	02	
	777	02	6710



6	6	3
Tag	Block	Word

Index

Set Associative Mapping

In direct mapping two words with the same index in their address but with different tag values cannot reside in cache memory at the same time.

In this mapping each data word is stored together with its tag and number of tag data items in one word of cache is said to form a set.

In general a set associative cache of set size k will accommodate k words of main memory in each word of cache.

When a miss occur and the set is full, one of the tag data item is replaced with new value using various algorithm.

Two way set associative mapping cache

Index	Tag	Data		Tag	Data
000	01	3450		02	5670
777	02	6710		00	2340

Writing into Cache

Writing can be done in two ways:

1. Write through

2. Write back

In the write through, whenever wrote option is performed in cache memory main memory is also updated in parallel with the cache.

In write back we only cache is updated and marked by the flag.When the word is removed from cache flag is checked if it is set the corresponding address in main memory is updated.

Cache Initialization

When power is turned on cache contain invalid data indicated by valid bit value 0.

Valid bit of word is set whenever the word is read from main memory and updated to cache.

If valid bit is 0,new word automatically replace the invalid data.

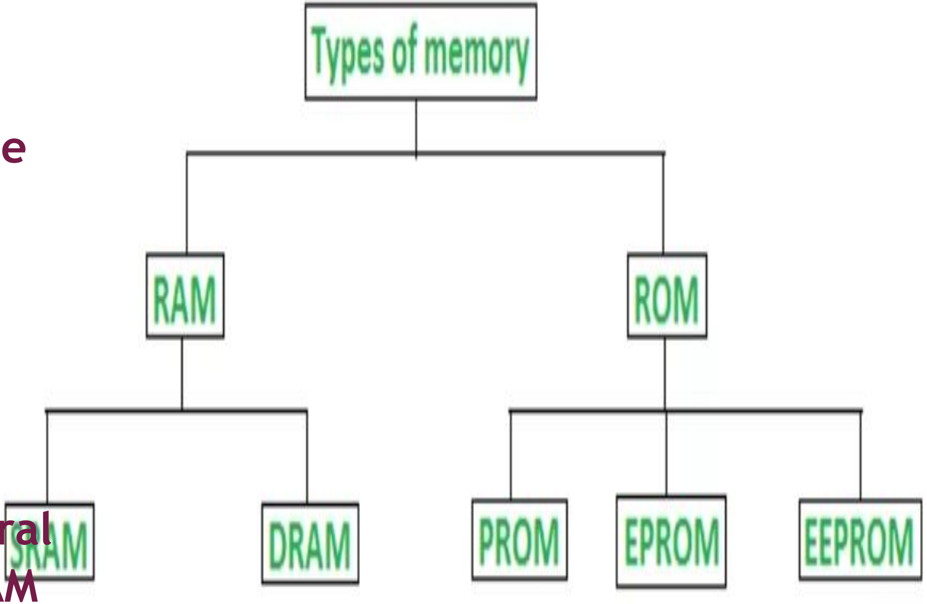
MAIN MEMORY

- Main memory (sometimes called primary storage) refers to storage locations that are directly accessible by the processor.
- Types of main memory:-

  RAM (Random Access Memory)

  ROM(Read Only Memory)

- Most of the main memory in a general purpose computer is made up of RAM integrated circuit chips,but a portion of the memory may be constructed with ROM chips.



RAM(Random Access Memory)

- It’s called random access because the data can be quickly read and modified in any order. Compare this with older storage media like CD-RWs, where data is accessed in a fixed sequence that's slower.
- RAM is computer’s short-term memory. For example, when we start our operating system — the applications we need like our audio or our antivirus software are copied into our computer’s memory for our processor to access readily. But when we turn off or restart our computer, the RAM clears.So it is a volatile memory.

Types Of RAM:-

SRAM	DRAM
• A type of semiconductor memory that uses bi- stable latching circuitry (flip flop) to store each bit.	• A type of random access semiconductor memory that stores each bit of data in separate tiny capacitor within an integrated circuit
• Stands for Static Random access memory	• Stands for Dynamic Random Access Memory.
• Each cell stores bit with a six transistor circuit.	• Each cell stores bit with a Capacitor and a transistor.
• Retains value indefinitely, as long as it is kept powered.	• Value must be refreshed every 10 -100ms.

● Does not require fresh cycles to retain data.	● Require periodic refresh cycles to retain data.
● Requires refreshing, it has more complex circuitry and timing requirements.	● Not as complex as SRAM.
● Used as CPU cache.	● Used for computer’s main memory.
● Requires minimum time to access data.	● Requires more time to access data.
● Very fast(8 to 16 times faster).	● Not as fast as SRAM.

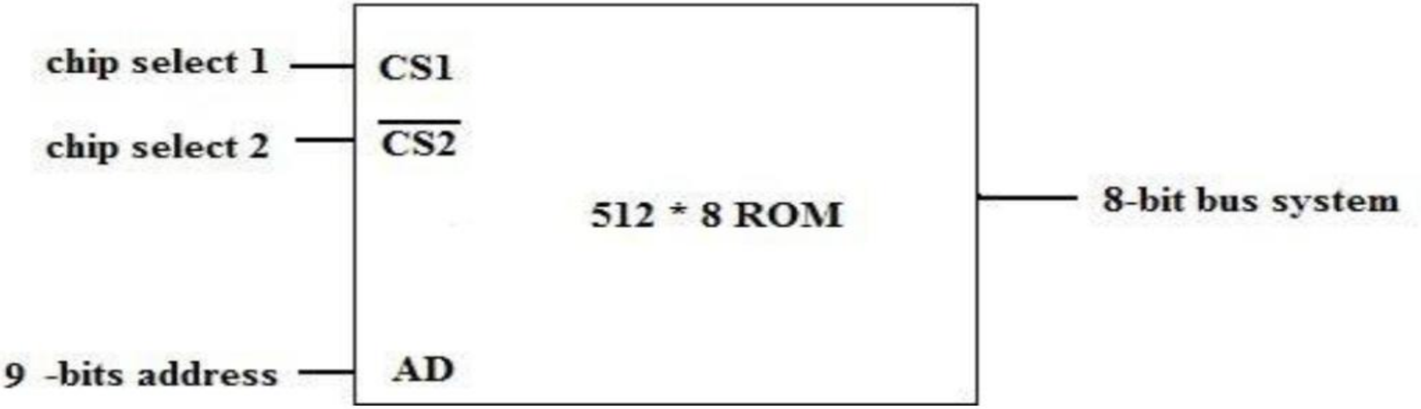
Complex structure - has flip- flop.	Simple structure - has transistor and a capacitor.
Has a lower density.	Has a higher density.
Expensive(8 to 16 times more expensive than DRAM)	Less Expensive
Relatively insensitive to disturbances such as electrical noise.	Sensitive to disturbances.

SRAM Vs DRAM Summary:-

	Time per bit	Access Time	Persist ?	Sensitive?	Cost	Application
SRAM	6	1X	Yes	No	100X	Cache memory
DRAM	1	10X	No	Yes	1X	Main memory, frame buffers.

ROM(Read Only Memory)

- 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 computer is completed.
- The ROM portion of the main memory is needed for storing a initial program called BOOTSTRAP LOADER, which is to start the computer software operating when power is turned off.



Types of ROM

PROM (Programmable read-only memory)	It can be programmed by the user. Once programmed, the data and instructions in it cannot be changed.
EPROM (Erasable Programmable read-only memory) –	It can be reprogrammed. To erase data from it, expose it to ultraviolet light. To reprogram it, erase all the previous data.
EEPROM (Electrically erasable programmable read-only memory)	The data can be erased by applying an electric field, with no need for ultraviolet light. We can erase only portions of the chip.
MROM(Mask ROM) –	is masked off at the time of production. Like other types of ROM, mask ROM cannot enable the user to change the data stored in it. If it can, the process would be difficult or slow.

VIRTUAL MEMORY

DEFINITION:-

- Virtual memory is a memory management technique where secondary memory can be used as if it were a part of the main memory.

What is virtual memory?

- Virtual memory is a memory management technique where secondary memory can be used as if it were a part of the main memory.

- Virtual memory is a common technique used in a computer's operating system (OS).

- Virtual memory uses both hardware and software to enable a computer to compensate for physical memory shortages, temporarily transferring data from random access memory (RAM) to disk storage.

- Mapping chunks of memory to disk files enables a computer to treat secondary memory as though it were main memory.

- Virtual memory frees up RAM by swapping data that has not been used recently over to a storage device, such as a hard drive or solid-state drive (SSD).

- Virtual memory is important for improving system performance, multitasking and using large programs. However, users should not overly rely on virtual memory, since it is considerably slower than RAM. If the OS has to swap data between virtual memory and RAM too often, the computer will begin to slow down – this is called thrashing.

How virtual memory works?

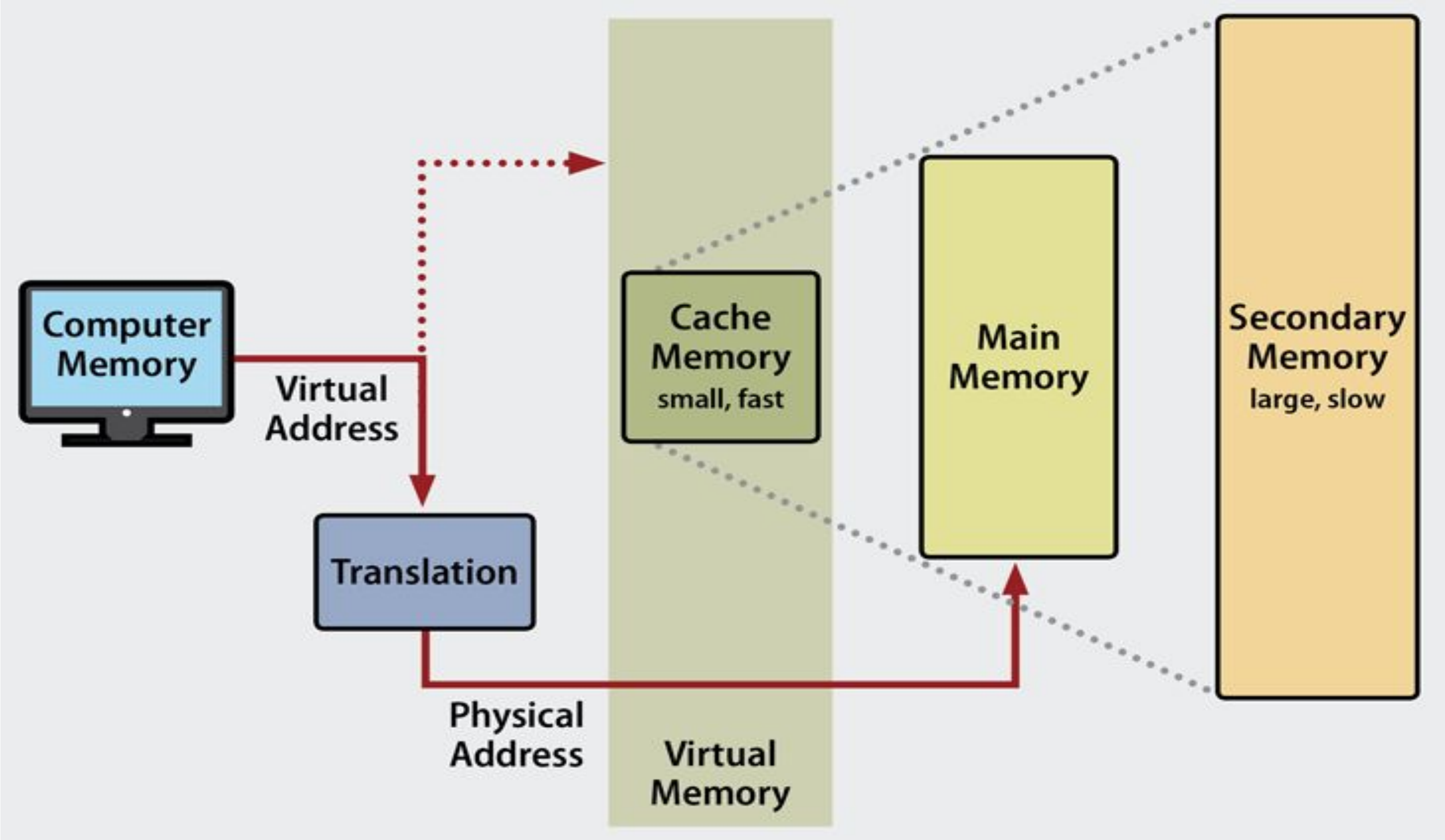
- Virtual memory uses both hardware and software to operate.

- When an application is in use, data from that program is stored in a physical address using RAM.

- A memory management unit (MMU) maps the address to RAM and automatically translates addresses. The MMU can, for example, map a logical address space to a corresponding physical address.

- If, at any point, the RAM space is needed for something more urgent, data can be swapped out of RAM and into virtual memory.

- The computer’s memory manager is in charge of keeping track of the shifts between physical and virtual memory. If that data is needed again, the computer’s MMU will use a context switch to resume execution


Computer Memory
Virtual Address
Translation
Physical Address
Cache Memory
small, fast
Virtual Memory
Main Memory
Secondary Memory
large, slow

While copying virtual memory into physical memory, the OS divides memory with a fixed number of addresses into either page files or swap files. Each page is stored on a disk, and when the page is needed, the OS copies it from the disk to main memory and translates the virtual addresses into real addresses.

Example:

A business owner might use their computer's virtual memory system when running multiple applications at once. For example, the user might try to load their email in their browser window while also running a word processing software, a shift scheduling software and a content management system at the same time. Since the computer needs to run several programs at once, it might adjust its memory usage to optimize its ability to open the user's email application while maintaining the operations of the other software programs.

To open the user’s email, the computer’s operating system (OS) may have to initiate its memory management unit (MMU) to search for the page or segment table containing the virtual or physical address for the process that can open the email application. Once located, the OS can either move the application to the computer’s RAM to open the application, or it can access the application if it’s already stored in RAM. If the RAM is near its limit, the computer may move another file from the RAM to another storage space to reduce its RAM usage.

Snapshot of a virtual memory management system :-

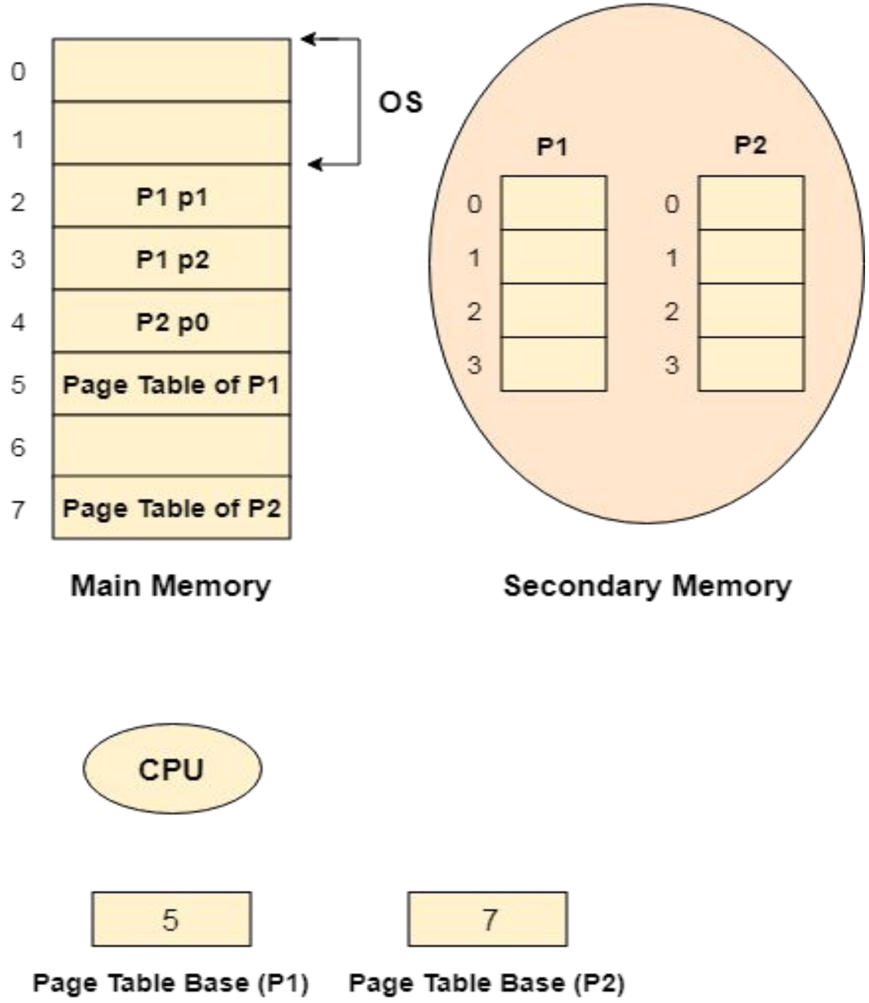
Let us assume 2 processes, P1 and P2, contains 4 pages each. Each page size is 1 KB. The main memory contains 8 frame of 1 KB each. The OS resides in the first two partitions. In the third partition, 1st page of P1 is stored and the other frames are also shown as filled with the different pages of processes in the main memory.

The page tables of both the pages are 1 KB size each and therefore they can be fit in one frame each. The page tables of both the processes contain various information that is also shown in the image.

The CPU contains a register which contains the base address of page table that is 5 in the case of P1 and 7 in the case of P2. This page table base address will be added to the page number of the Logical address when it comes to accessing the actual corresponding entry.

	Frame	Present / Absent	D Bit	Reference bit	Protection
0		0	0	0	
1	2	1	0	1	
2	3	1	1	0	
3		0	0	0	

Page Table of P1



	Frame	Present / Absent	D Bit	Reference bit	Protection
0	4	1	0	1	
1		0	0	0	
2		0	0	0	
3		0	0	0	

Page Table of P2

Types of virtual memory:- Paging and Segmentation

Virtual memory can be managed in a number of different ways by a system's operating system, and the two most common approaches are paging and segmentation.

Virtual Memory Paging

- In a system which uses paging, RAM is divided into a number of blocks – usually 4k in size – called pages.

- Processes are then allocated just enough pages to meet their memory requirements.

- That means that there will always be a small amount of memory wasted, except in the unusual case where a process requires exactly a whole number of pages.

- During the normal course of operations, pages (i.e. memory blocks of 4K in size) are swapped between RAM and a page file, which represents the virtual memory.

Virtual Memory Segmentation

- Segmentation is an alternative approach to memory management, where instead of pages of a fixed size, processes are allocated segments of differing length to exactly meet their requirements.

- That means that unlike in a paged system, no memory is wasted in a segment.

- Segmentation also allows applications to be split up into logically independent address spaces, which can make them easier to share, and more secure.

- But a problem with segmentation is that because each segment is a different length, it can lead to memory “fragmentation.”

  This means that as segments are allocated and de-allocated, small chunks of memory can be left scattered around which are too small to be useful.

  - As these small chunks build up, fewer and fewer segments of useful size can be allocated.

  - And if the OS does start using these small segments then there are a huge number to keep track of, and each process will need to use many different segments, which is inefficient and can reduce performance.

The benefits of using virtual memory:-

The advantages to using virtual memory include:-

- It can handle twice as many addresses as main memory.
- It enables more applications to be used at once.
- It frees applications from managing shared memory and saves users from having to add memory modules when RAM space runs out.
- It has increased speed when only a segment of a program is needed for execution.
- It has increased security because of memory isolation.
- It enables multiple larger applications to run simultaneously.

- Allocating memory is relatively inexpensive.

The limitations of using virtual memory:-

Although the use of virtual memory has its benefits, it also comes with some trade-offs worth considering, such as:-

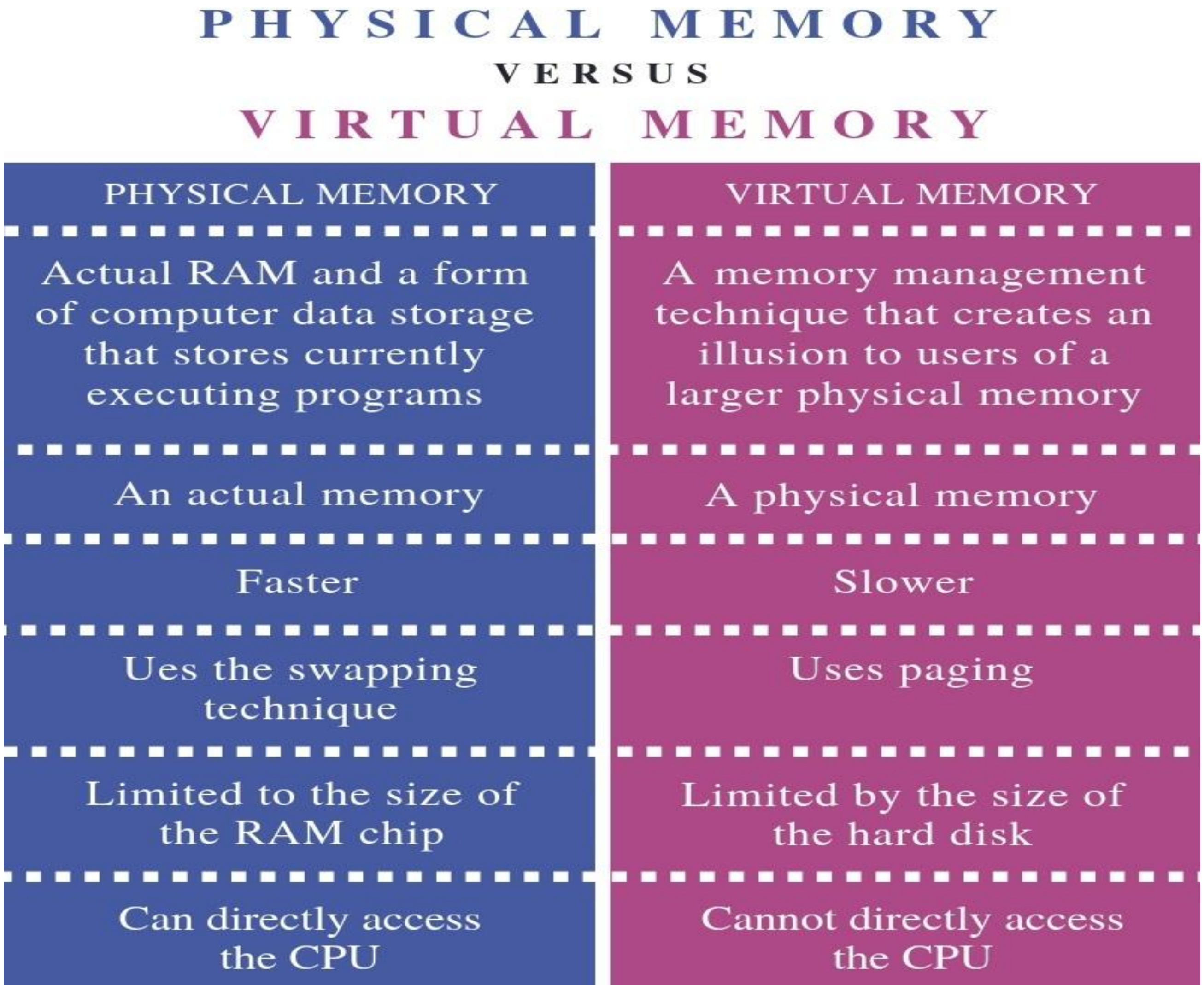
- Applications run slower if they are running from virtual memory
- Data must be mapped between virtual and physical memory, which requires extra hardware support for address translations, slowing down a computer further.
- The size of virtual storage is limited by the amount of secondary storage, as well as the addressing scheme with the computer system.
- Thrashing can occur if there is not enough RAM, which will make the computer perform slower.

- It may take time to switch between applications using virtual memory.
- It lessens the amount of available hard drive space.

Physical memory (RAM) vs. Virtual memory (Virtual RAM):-

When talking about the differences between virtual and physical memory, the biggest distinction commonly made is to speed. RAM is considerably faster than virtual memory. RAM, however, tends to be more expensive.

When a computer requires storage, RAM is the first used. Virtual memory, which is slower, is used only when the RAM is filled.



PHYSICAL MEMORY VERSUS VIRTUAL MEMORY

PHYSICAL MEMORY	VIRTUAL MEMORY
Actual RAM and a form of computer data storage that stores currently executing programs	A memory management technique that creates an illusion to users of a larger physical memory
An actual memory	A physical memory
Faster	Slower
Ues the swapping technique	Uses paging
Limited to the size of the RAM chip	Limited by the size of the hard disk
Can directly access the CPU	Cannot directly access the CPU

The key characteristics of memory devices or memory system are as follows:

1. Location
2. Capacity
3. Unit of Transfer
4. Access Method
5. Performance
6. Physical type
7. Physical characteristics
8. Organization

1. Location:

It deals with the location of the memory device in the computer system. There are three possible locations:

- CPU : This is often in the form of CPU registers and small amount of cache
- Internal or main: This is the main memory like RAM or ROM. The CPU can directly access the main memory.
- External or secondary: It comprises of secondary storage devices like hard disks, magnetic tapes. The CPU doesn’t access these devices directly. It uses device controllers to access secondary storage devices.

2. Capacity:

The capacity of any memory device is expressed in terms of:
i)word size ii)Number of words

- **Word size:** Words are expressed in bytes (8 bits). A word can however mean any number of bytes. Commonly used word sizes are 1 byte (8 bits), 2bytes (16 bits) and 4 bytes (32 bits).
- **Number of words:** This specifies the number of words available in the particular memory device. For example, if a memory device is given as 4K x 16.This means the device has a word size of 16 bits and a total of 4096(4K) words in memory.

3. Unit of Transfer:

It is the maximum number of bits that can be read or written into the memory at a time. In case of main memory, it is mostly equal to word size. In case of external memory, unit of transfer is not limited to the word size; it is often larger and is referred to as blocks.

4. Access Methods:

It is a fundamental characteristic of memory devices. It is the sequence or order in which memory can be accessed. There are three types of access methods:

- **Random Access:** If storage locations in a particular memory device can be accessed in any order and access time is independent of the memory location being accessed. Such memory devices are said to have a random access mechanism. RAM (Random Access Memory) IC’s use this access method.

- **Serial Access:** If memory locations can be accessed only in a certain predetermined sequence, this access method is called serial access. Magnetic Tapes, CD-ROMs employ serial access methods.
- **Semi random Access:** Memory devices such as Magnetic Hard disks use this access method. Here each track has a read/write head thus each track can be accessed randomly but access within each track is restricted to a serial access.

5. Performance:

The performance of the memory system is determined using three parameters

- **Access Time:** In random access memories, it is the time taken by memory to complete the read/write operation from the instant that an address is sent to the memory. For non-random access memories, it is the time taken to position the read write head at the desired location. Access time is widely used to measure performance of memory devices.

- **Memory cycle time:** It is defined only for Random Access Memories and is the sum of the access time and the additional time required before the second access can commence.
- **Transfer rate:** It is defined as the rate at which data can be transferred into or out of a memory unit.

6. Physical type: Memory devices can be either semiconductor memory (like RAM) or magnetic surface memory (like Hard disks).

7.Physical Characteristics:

- **Volatile/Non- Volatile:** If a memory devices continues hold data even if power is turned off. The memory device is non-volatile else it is volatile.

8. Organization:

- **Erasable/Non-erasable:** The memories in which data once programmed cannot be erased are called Non-erasable memories. Memory devices in which data in the memory can be erased is called erasable memory.
  E.g. RAM(erasable), ROM(non-erasable).

References

- M.M Mano,Computer System Architecture, PHI
- https://www.Wikipedia.org
- https://www.geeksforgeeks.org
- https://www.javatpoint.com