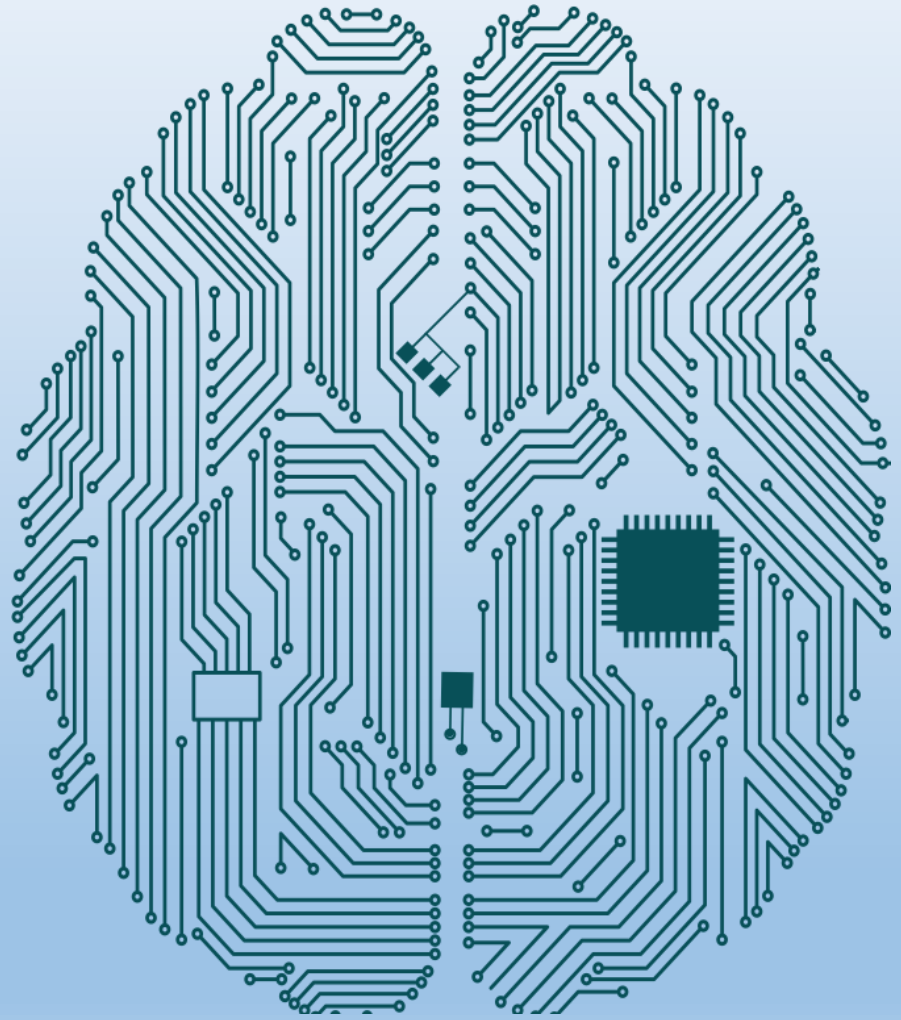


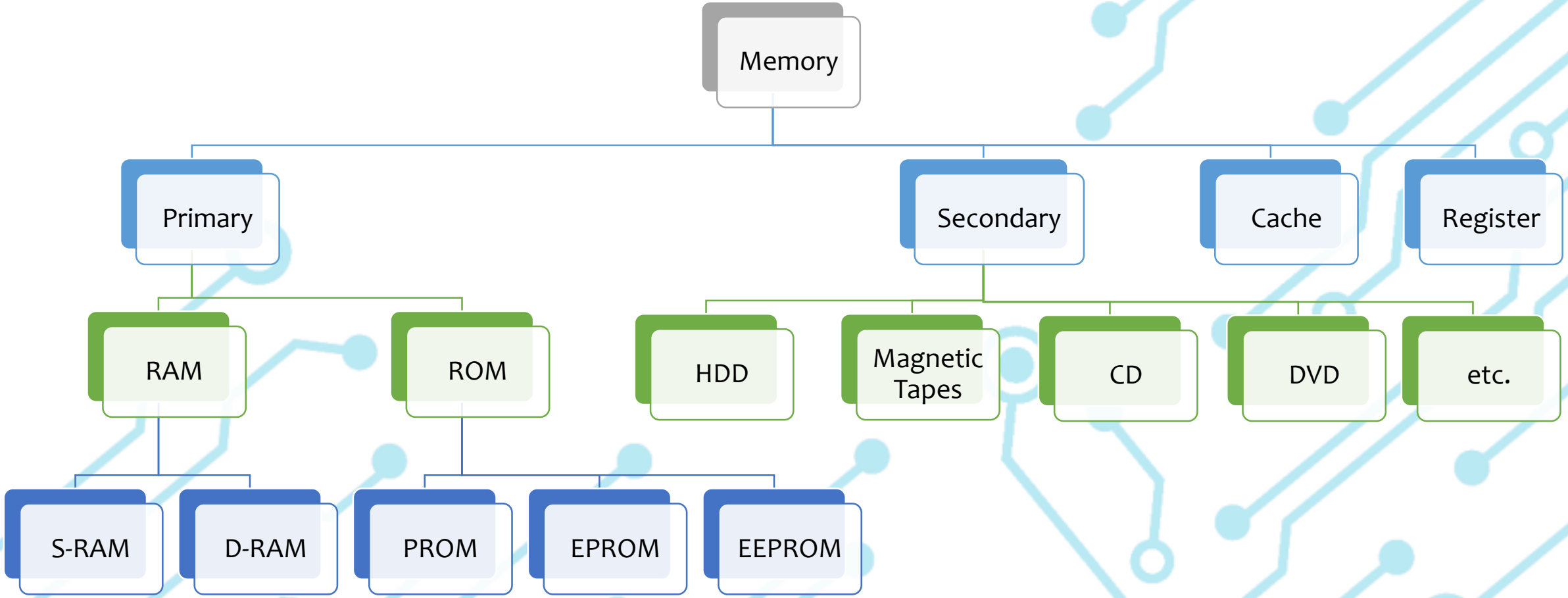
Memory Organization

Unit - IV

Topics to be Covered

- Memory Classifications
- Memory Hierarchy
- Various Types of Memory
  - RAM
  - ROM
  - PROM
  - EPROM
  - EEPROM
  - Associative Memory
- Various Types of Auxiliary Memory
  - Magnetic Tapes
  - Floppy Disks
  - Hard Disks
  - Flash Memory
- Cache Memory
- Virtual Memory

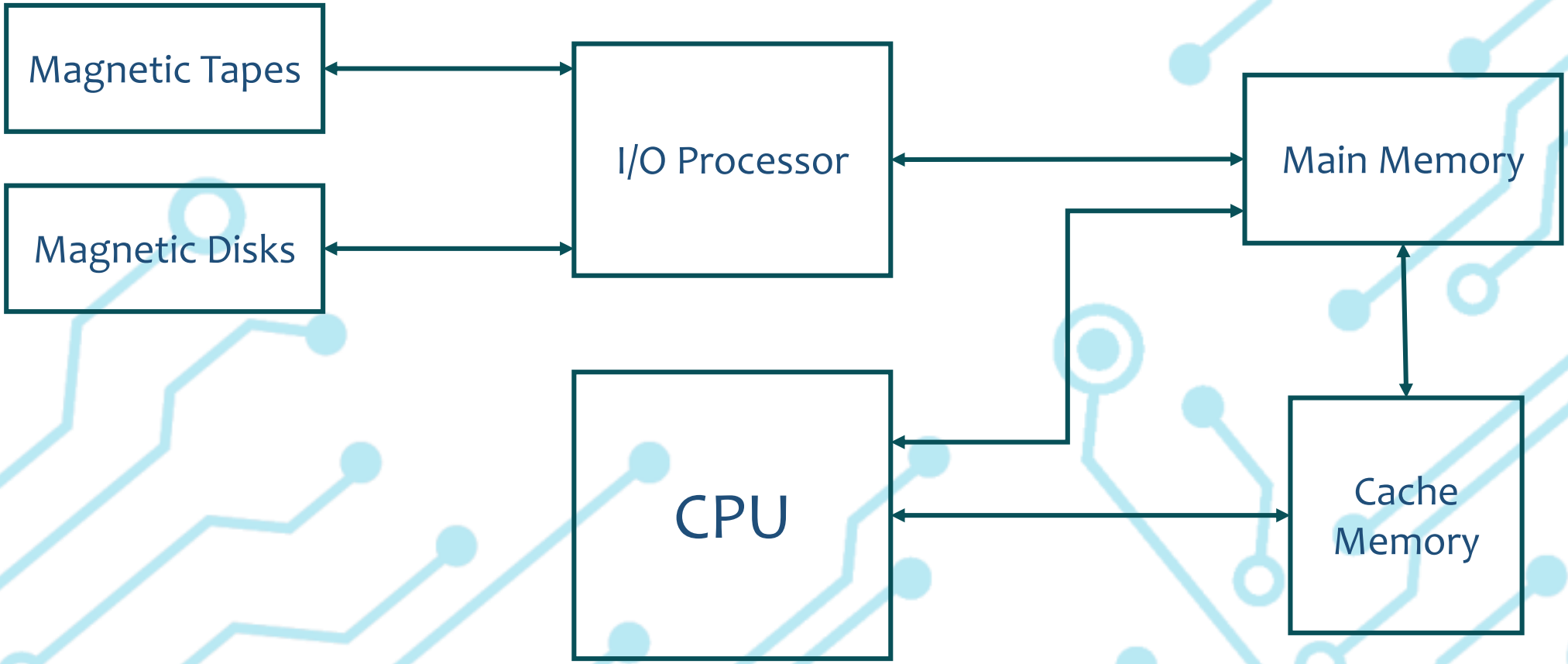
Memory Classification

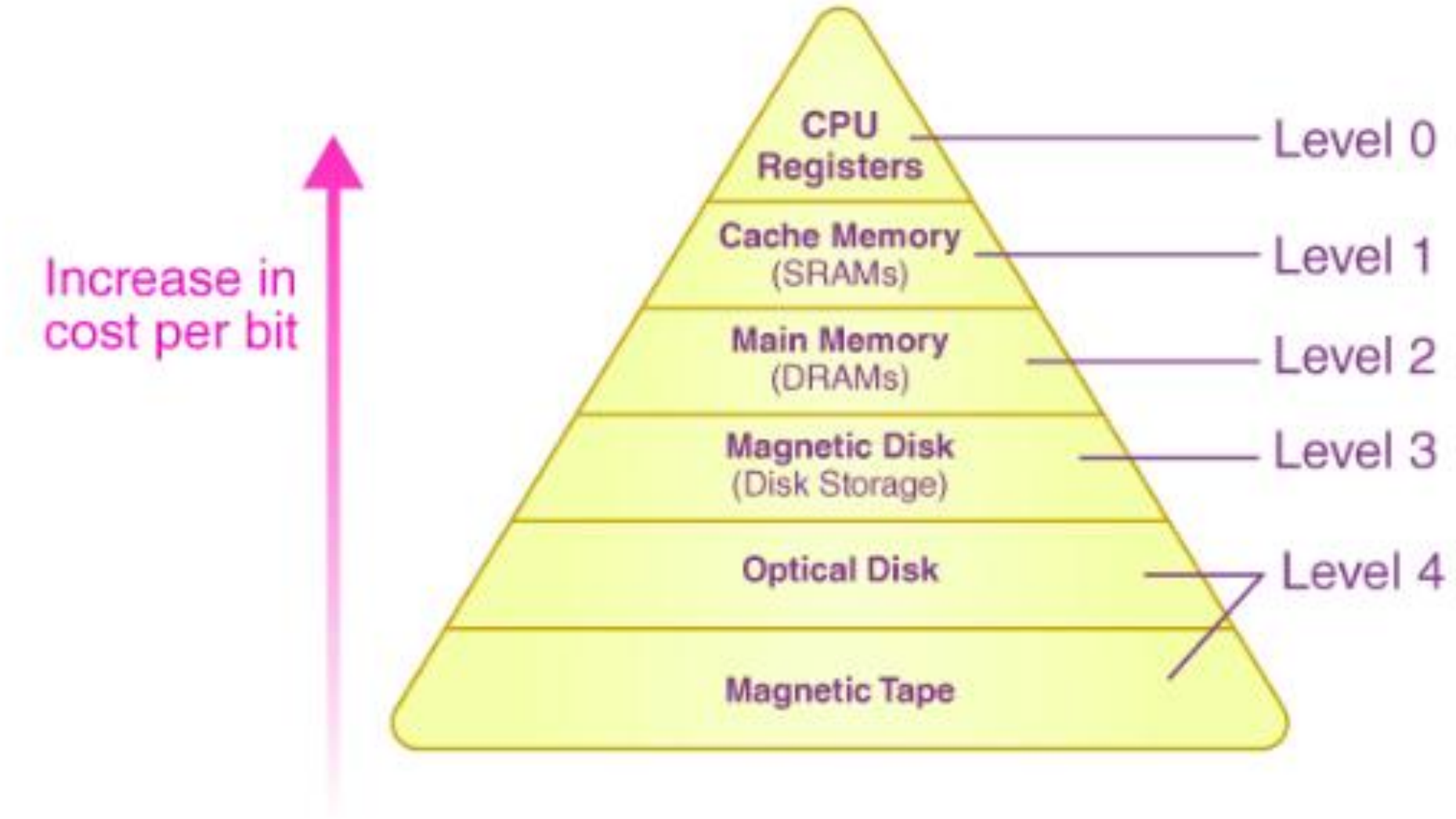


Memory Hierarchy

- The memory hierarchy in computer architecture is a system of organizing different types of memory to optimize performance.
- It arranges memory from fastest and smallest (at the top) to slowest and largest (at the bottom), creating a tiered system that balances speed, cost, and capacity.
- This structure allows the processor to access frequently used data quickly while still having access to larger, slower storage for less frequently used information.

Memory Hierarchy




Increase in cost per bit
CPU Registers
Level 0
Cache Memory (SRAMs)
Level 1
Main Memory (DRAMs)
Level 2
Magnetic Disk (Disk Storage)
Level 3
Optical Disk
Level 4
Magnetic Tape

1. Registers

- Registers are small, high-speed memory units located in the CPU. They are used to store the most frequently used data and instructions. Registers have the fastest access time and the smallest storage capacity, typically ranging from 16 to 64 bits.

2. Cache Memory

- Cache memory is a small, fast memory unit located close to the CPU. It stores frequently used data and instructions that have been recently accessed from the main memory. Cache memory is designed to minimize the time it takes to access data by providing the CPU with quick access to frequently used data.

3. Main Memory

- Main memory, also known as RAM (Random Access Memory), is the primary memory of a computer system. It has a larger storage capacity than cache memory, but it is slower. Main memory is used to store data and instructions that are currently in use by the CPU.
- **Types of Main Memory**
- **Static RAM:** Static RAM stores the binary information in flip flops and information remains valid until power is supplied. Static RAM has a faster access time and is used in implementing cache memory.
- **Dynamic RAM:** It stores the binary information as a charge on the capacitor. It requires refreshing circuitry to maintain the charge on the capacitors after a few milliseconds. It contains more memory cells per unit area as compared to SRAM.

4. Secondary Storage

- Secondary storage, such as hard disk drives (HDD) and solid-state drives (SSD) , is a non-volatile memory unit that has a larger storage capacity than main memory. It is used to store data and instructions that are not currently in use by the CPU. Secondary storage has the slowest access time and is typically the least expensive type of memory in the memory hierarchy.

5. Magnetic Disk

- Magnetic Disks are simply circular plates that are fabricated with either a metal or a plastic or a magnetized material. The Magnetic disks work at a high speed inside the computer and these are frequently used.

6. Magnetic Tape

- Magnetic Tape is simply a magnetic recording device that is covered with a plastic film. Magnetic Tape is generally used for the backup of data. In the case of a magnetic tape, the access time for a computer is a little slower and therefore, it requires some amount of time for accessing the strip.

Main Memory

- The main memory is the central storage unit in a computer system.
- It is a relatively large and fast memory used to store programs and data during the computer operation
- Integrated circuit RAM chips are available in two possible operating modes, **static and dynamic**.
- 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**.

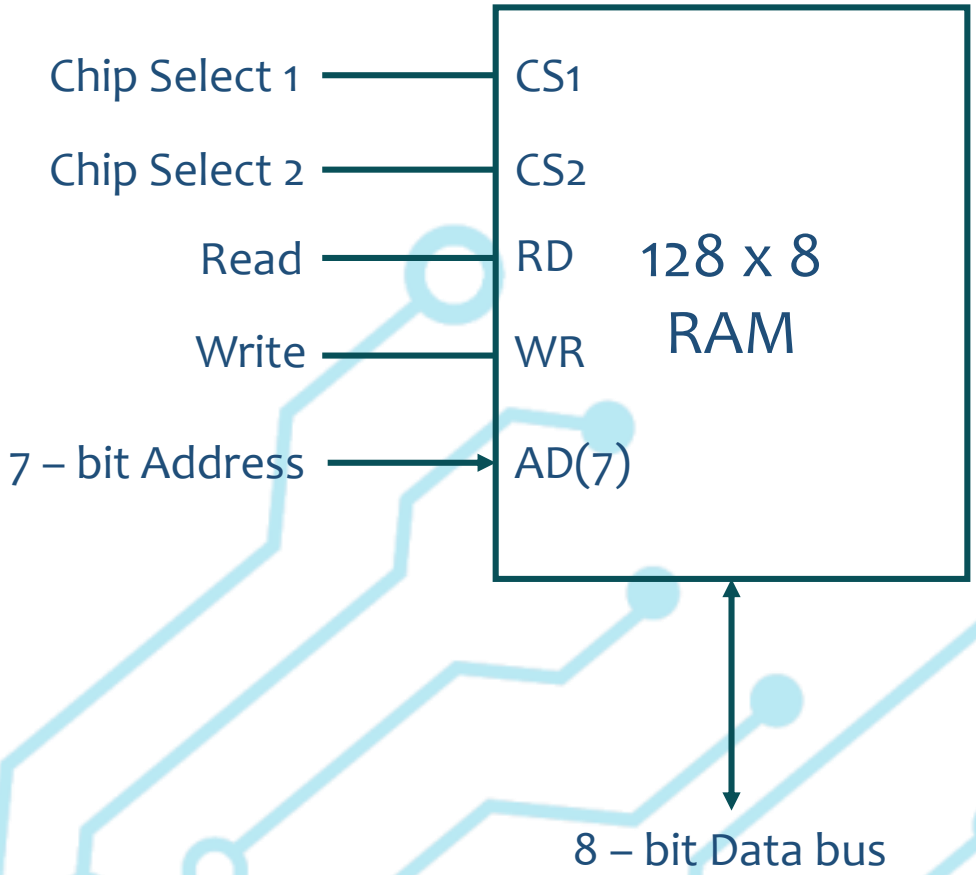
RAM: Static RAM v/s Dynamic RAM

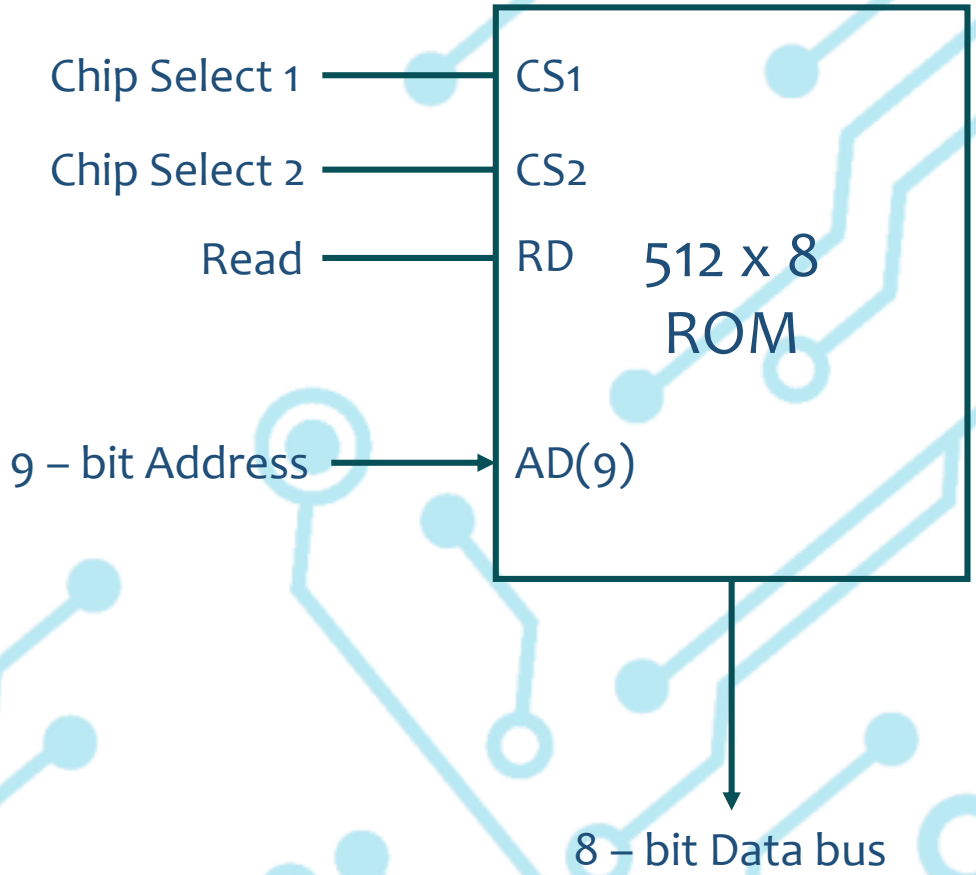
	Static RAM	Dynamic RAM
Technique of Data Storage	The **static RAM** consists essentially of internal flip-flops that store the binary information. The stored information remains valid as long as power is applied to the unit.	The **dynamic RAM** stores the binary information in the form of electric charges that are applied to capacitors. The capacitors are provided inside the chip by MOS transistors.
Refreshing	No need to Refresh	The stored charge on the capacitors tend to discharge with time and the capacitors must be periodically recharged by refreshing the dynamic memory. Refreshing is done by cycling through the words every few milliseconds to restore the decaying charge
Significant Characteristic	The static RAM is easier to use and has shorter read and write cycles.	The dynamic RAM offers reduced power consumption and larger storage capacity in a single memory chip.
Application	One of the major applications of the static RAM is in implementing the cache memories	The dynamic RAMs are used for implementing the main memory.

ROM

- As discussed earlier 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**.
- Originally, **RAM** was used to refer to a random access memory, but now it is used to designate a **read/write** memory to distinguish it from a **read-only memory**, although **ROM** is also random access.
- **RAM** is used for storing the **bulk of the programs and data that are subject to change**. **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.
- ROM is needed for storing an initial program called a **bootstrap loader**.
- ROM remain unchanged after power is turned off and on again.

RAM & ROM Chips



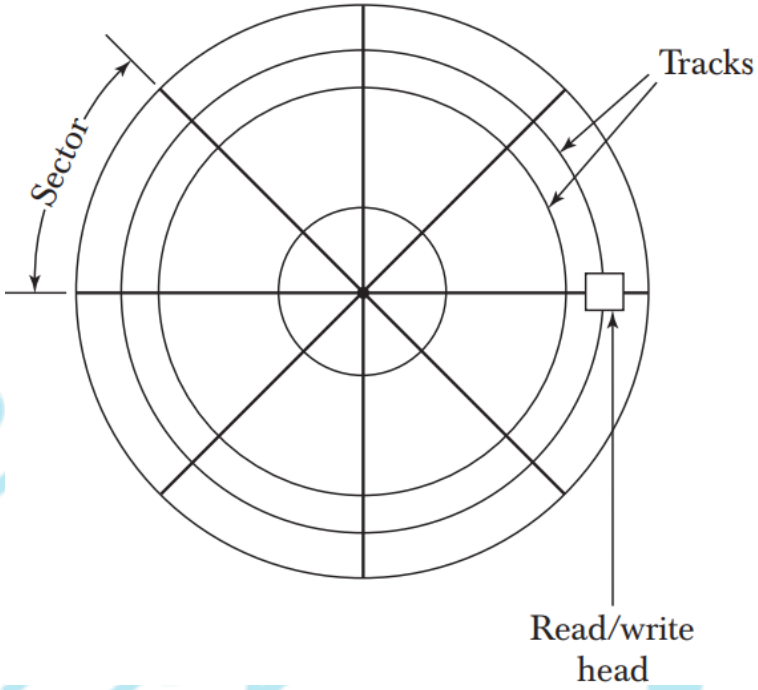


Auxiliary Memory

- The most common auxiliary memory devices used in computer systems are **magnetic disks and tapes**. Other components used, but not as frequently, are magnetic drums, magnetic bubble memory, and optical disks.
- Their logical properties can be characterized and compared by a few parameters. The important characteristics of any device are its access mode.
- access time, transfer rate, capacity, and cost.
- The average time required to reach a storage location in memory and obtain its contents is called the **access time**.
- The access time consists of a **seek time** required to position the read-write head to a location and a **transfer time** required to transfer data to or from the device.

Magnetic Disks

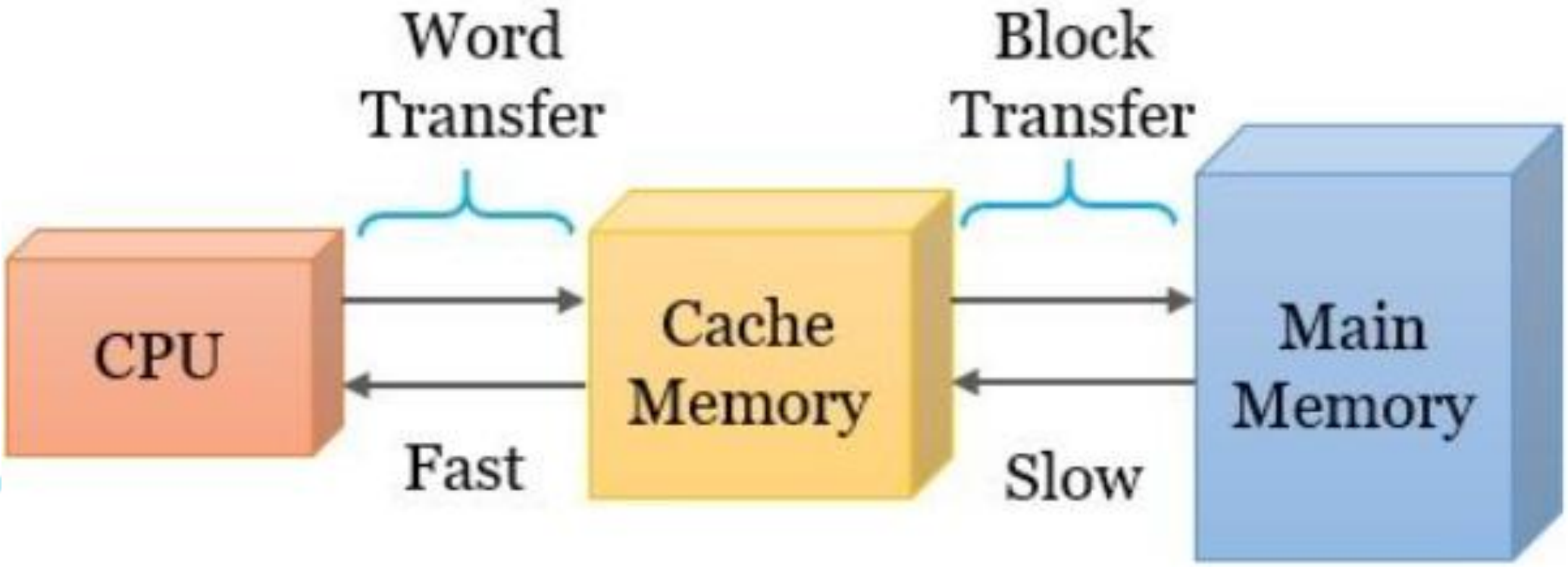
- A magnetic disk is a circular plate constructed of metal or plastic coated with magnetized material.
- Often both sides of the disk are used and several disks may be stacked on one spindle with read/write heads available on each surface.



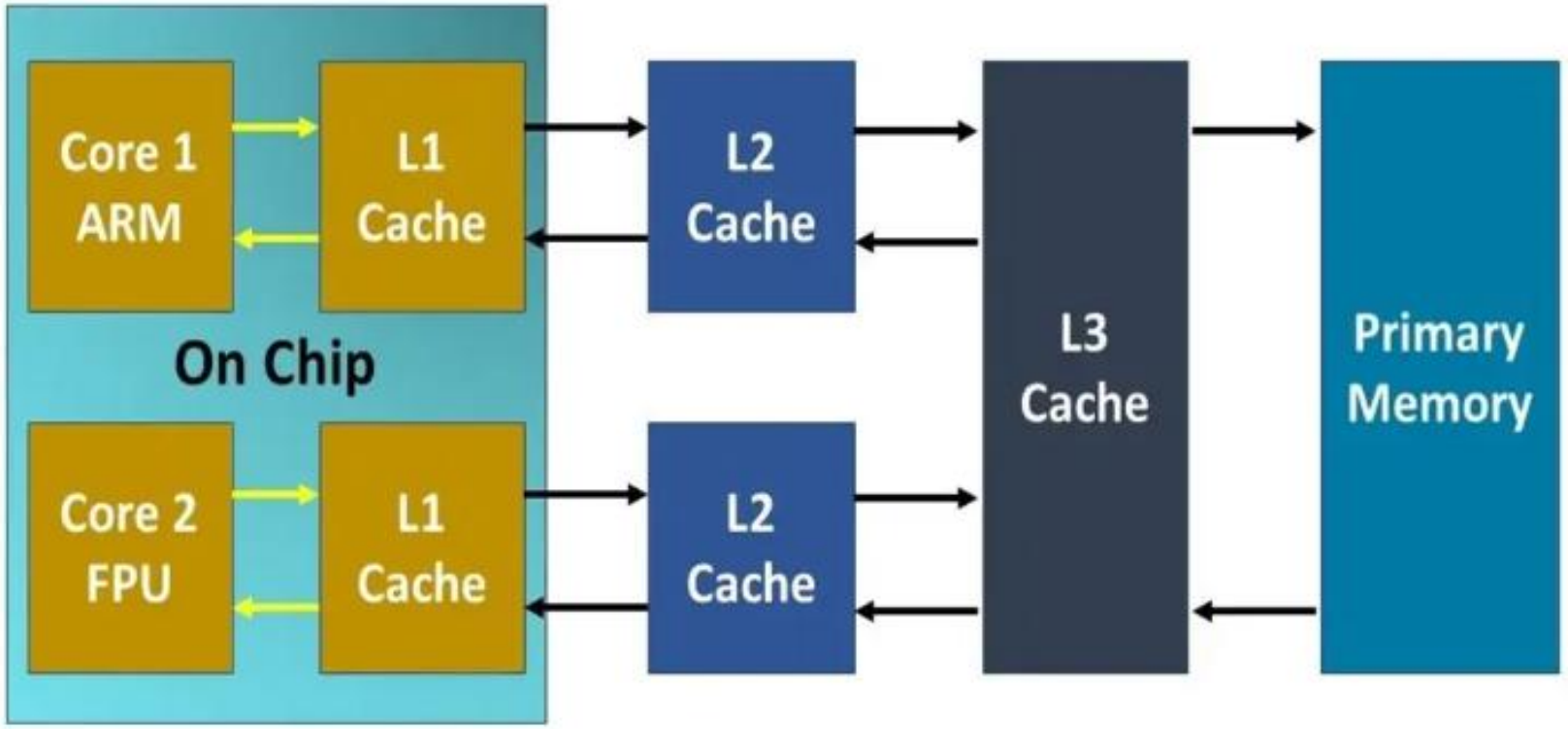
Magnetic Tapes

- A 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. Usually, seven or nine bits are recorded simultaneously to form a character together with a parity bit. Read/write heads are mounted one in each track so that data can be recorded and read as a sequence of characters.
- Magnetic tape units can be stopped, started to move forward or in reverse, or can be rewound.

Cache Memory



Cache Memory



DIFFERENCES BETWEEN CACHE AND RAM

CACHE MEMORY	RAM MEMORY
Fast access	Slower access
Low capacity	High capacity
Expensive	Cheaper
LEVEL 1,2&3	SRAM and DRAM

- Cache memory is a high-speed memory used to store frequently accessed data and instructions, improving computer performance by reducing the time it takes for the CPU to retrieve information.
- It acts as a buffer between the CPU and main memory (RAM), providing faster access to data compared to retrieving it directly from RAM.
- What it is:
- Cache memory is a smaller, faster memory than main memory (RAM).
- It's designed to store copies of data and instructions that the CPU uses most often.
- By storing frequently accessed information closer to the CPU, it reduces the need to access slower main memory, leading to faster processing.

- How it works:
- **1. Cache Hit:**
- When the CPU needs data, it first checks the cache. If the data is found (a "cache hit"), it's retrieved quickly.
- **2. Cache Miss:**
- If the data is not in the cache (a "cache miss"), the CPU needs to access the slower main memory to retrieve it.
- **3. Data is copied:**
- In the case of a cache miss, the data from main memory is not only sent to the CPU but also copied to the cache for future use.

- Levels of Cache:
- Cache memory is often organized into levels (L1, L2, and sometimes L3).
- **L1 cache:** The fastest and smallest cache, usually located directly on the CPU core.
- **L2 cache:** Larger and slightly slower than L1, may be located on the CPU or nearby.
- **L3 cache:** The largest and slowest of the cache levels, shared by all CPU cores.

- Benefits of Cache Memory:
- **Improved performance:**
- By reducing the time it takes to access frequently used data, cache memory significantly speeds up overall system performance.
- **Reduced latency:**
- Faster access to data translates to lower latency, meaning applications respond more quickly.
- **Efficient use of resources:**
- Cache memory helps optimize the use of system resources by reducing the need to access slower main memory constantly.

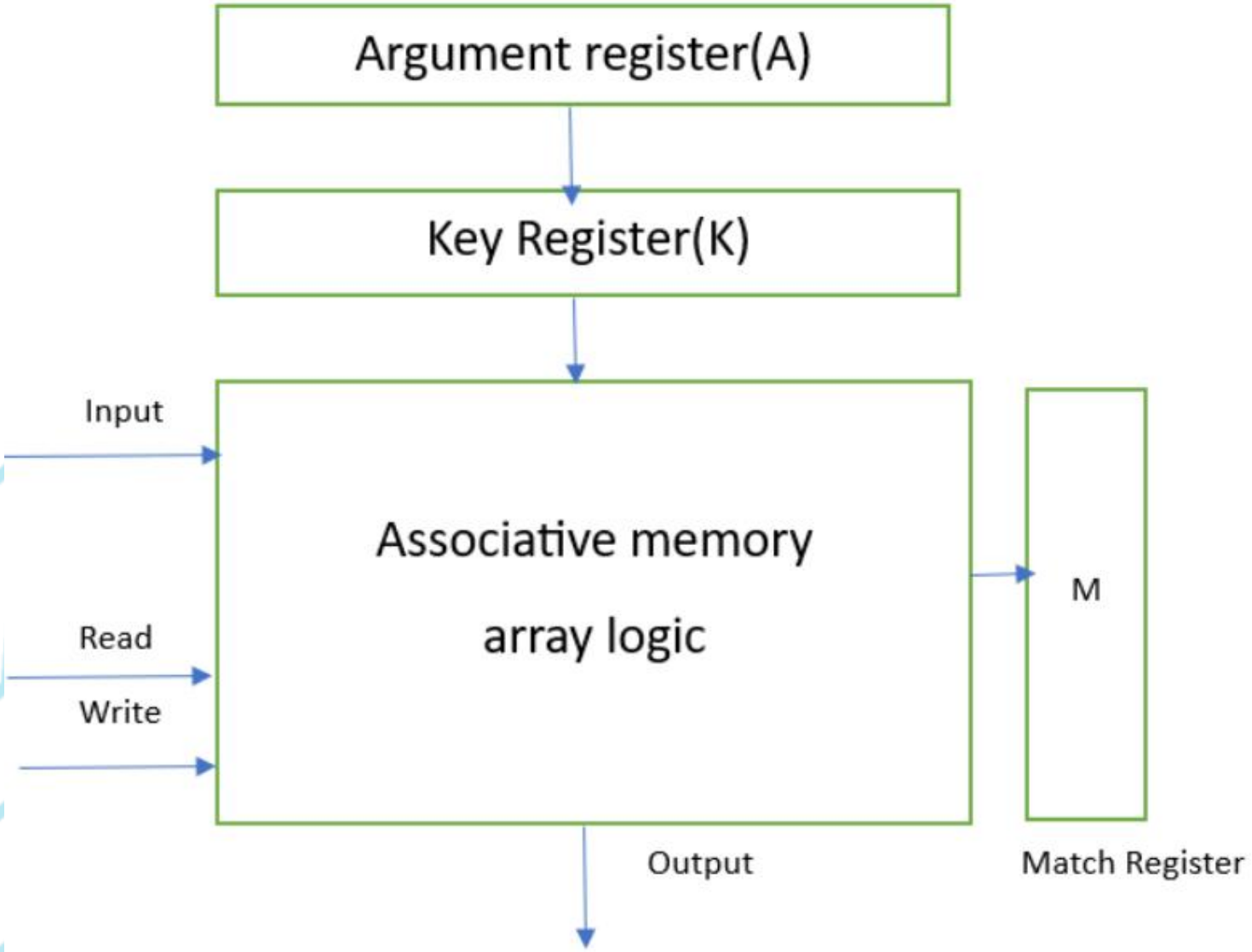
- Examples of Cache Memory in Use:
- **Web Browsing:**
- Browsers use cache to store copies of web pages, making them load faster on subsequent visits.
- **Databases:**
- Databases use cache to store frequently accessed data and query results, improving query performance.
- **Video Streaming:**
- Services like Netflix use cache to store video segments, allowing for smoother playback even with temporary internet interruptions.

Associative Memory

- **Associative memory** is also known as content addressable memory (CAM) or associative storage or associative array. It is a special type of memory that is optimized for performing searches through data, as opposed to providing a simple direct access to the data based on the address.
- It can store the set of patterns as memories when the associative memory is being presented with a key pattern, it responds by producing one of the stored pattern which closely resembles or relates to the key pattern.

How Does Associative Memory Work?

- In conventional memory, data is stored in specific locations, called addresses, and retrieved by referencing those addresses. In associative memory, data is stored together with additional tags or metadata that describe its content. When a search is performed, the associative memory compares the search query with the tags of all stored data, and retrieves the data that matches the query.
- Associative memory is designed to quickly find matching data, even when the search query is incomplete or imprecise. This is achieved by using parallel processing techniques, where multiple search queries can be performed simultaneously. The search is also performed in a single step, as opposed to conventional memory where multiple steps are required to locate the data.


Argument register(A)
Key Register(K)
Input
Associative memory
array logic
Read
Write
M
Output
Match Register

- Argument Register: It contains words to be searched. It contains 'n' number of bits.
- Match Register: It has m-bits, One bit corresponding to each word in the memory array. After the making process, the bits corresponding to matching words in match register are set to '1'.
- Key Register: It provides a mask of choosing a particular field/key in argument register. It specifies which part of the argument word need to be compared with words in memory.
- Associative Memory Array: It combines word in that are to be compared with the arguments word in parallel. It contains 'm' words with 'n' bit per word.

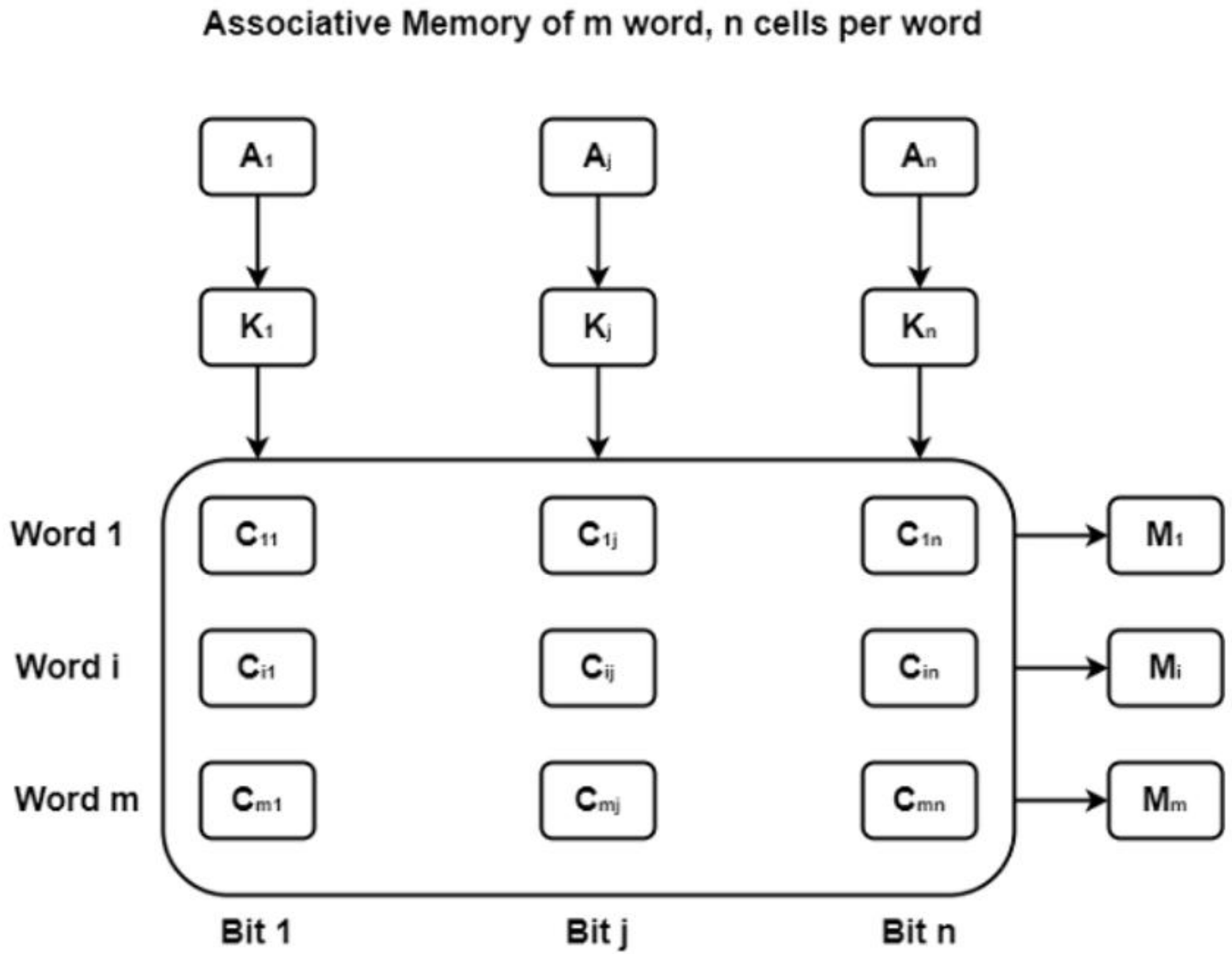
- **Applications of Associative memory :-**
- It can be only used in memory allocation format.
- It is widely used in the database management systems, etc.
- Networking: Associative memory is used in network routing tables to quickly find the path to a destination network based on its address.
- Image processing: Associative memory is used in image processing applications to search for specific features or patterns within an image.
- Artificial intelligence: Associative memory is used in artificial intelligence applications such as expert systems and pattern recognition.
- Database management: Associative memory can be used in database management systems to quickly retrieve data based on its content.

- **Advantages of Associative memory :-**
- It is used where search time needs to be less or short.
- It is suitable for parallel searches.
- It is often used to speedup databases.
- It is used in page tables used by the virtual memory and used in neural networks.
- **Disadvantages of Associative memory :-**
- It is more expensive than RAM
- Each cell must have storage capability and logical circuits for matching its content with external argument.

Difference between main memory and associative or content addressable memory

Main Memory	Associative or Content addressable memory
Memory accessed by using address of contents/words	Memory is accessed by matching contents with word stored in memory
To store a word, we assign the address and write the word at that address	To write a word, memory finds the empty space and writes it there.
To read a word an address is assigned and then a read command given.	To read operation is performed by matching the content with every word in the memory.
It has address decoder	It has a match logic
This type of memory is Cheaper	This memory is expensive
It is slow	Faster
Employed in general purpose computers	Employed in applications where search time very critical and may be very short

The following figure can define the relation between the memory array and the external registers in associative memory.



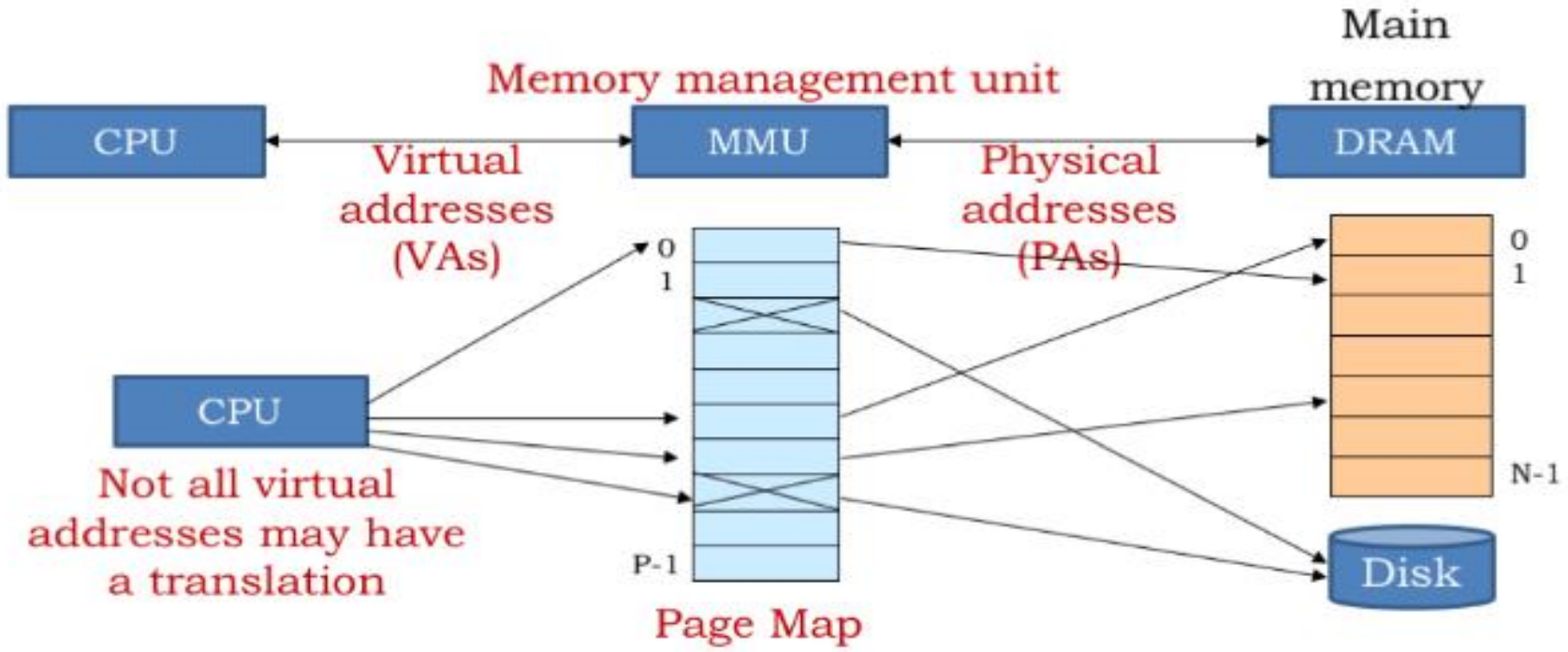
- The cells in the array are considered by the letter C with two subscripts. The first subscript provides the word number and the second determines the bit position in the word. Therefore cell C_{ij} is the cell for bit j in word i.
- A bit in the argument register is compared with all the bits in column j of the array supported that $K_j = 1$. This is completed for all columns j = 1, 2 . . . , n.
- If a match appears between all the unmasked bits of the argument and the bits in word i, the equivalent bit M_i in the match register is set to 1. If one or more unmasked bits of the argument and the word do not match, M_i is cleared to 0.

Virtual Memory

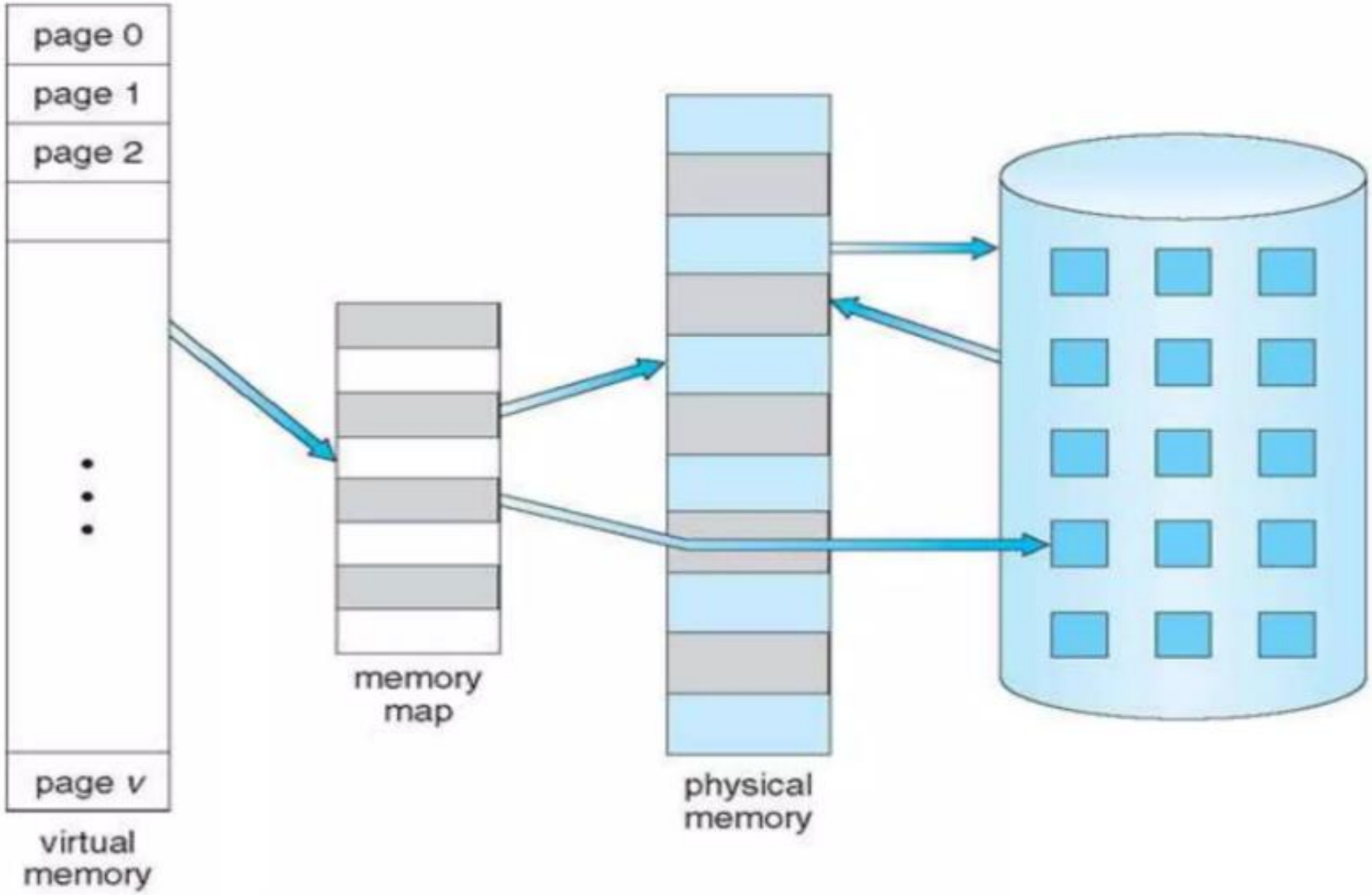
- In Computer Architecture (COA), **virtual memory** is a technique that creates the illusion of a larger, more accessible memory than physically available by using a portion of the secondary storage (like a hard drive or SSD) as an extension of RAM.
- This allows for the execution of programs larger than physical memory, efficient multitasking, and easier programming by abstracting away physical memory limitations.

Virtual Memory

- Two kinds of addresses:
  - CPU uses virtual addresses
  - Main memory uses physical addresses
- Hardware translates virtual addresses to physical addresses via an operating system (OS)-managed table, the page map



Virtual Memory



Virtual Memory

- The simple tactic of breaking a process up into pages led to the development of important concept : Virtual memory
- ***Virtual memory*** **is imaginary memory: it gives you the illusion of a memory arrangement that's not physically there.**

- OS and hardware produce illusion of a disk as fast as main memory
- Virtual memory as an alternate set of memory addresses.
- Programs use these virtual addresses rather than real addresses to store instructions and data.
- When the program is actually executed, the virtual addresses are converted into real memory addresses.
- Process runs when not all pages are loaded in memory
- Only keep referenced pages in main memory
- Keep unreferenced pages on slower, cheaper backing store (disk)
- Bring pages from disk to memory when necessary.

Advantages

- You can run more applications at once.
- Allows you to fit many large programs into a relatively small RAM.
- You don't have to buy more memory RAM
- VM supports Swapping.
- Common data or code may be shared to save memory.
- Process need not be in memory as a whole , Only part of a program needs to be loaded into memory.
- Process may even be larger than all of physical memory.
- Data / code can be read from disk as needed.
- Code can be placed anywhere in physical memory without relocation.
- More processes can be maintained in Main Memory which increases effective use of CPU.
- Don't need to break program into fragments to accommodate memory limitations

Demand paging

- **Fetch Policy** Determines when a page should be brought into main memory. One of common policies is **Demand paging**
- Refines paging by demand paging - each page of a process is brought in only when needed, that is, on demand
- Demand paging
  - Do not require all pages of a process in memory
  - Bring in pages as required
  - Less I/O needed
  - Less memory needed
  - Faster response
  - More users „

Virtual Memory	Cache Memory
increases the capacity of main memory.	increase the accessing speed of CPU.
Virtual memory is not a memory unit, its a technique.	Cache memory is exactly a memory unit.
The size of virtual memory is greater than the cache memory.	While the size of cache memory is less than the virtual memory.
Operating System manages the Virtual memory.	On the other hand hardware manages the cache memory.
In virtual memory, the program with size larger than the main memory are executed.	While in cache memory, recently used data is copied into.
In virtual memory, mapping frameworks is needed for mapping virtual address to physical address.	While in cache memory, no such mapping frameworks is needed.
It is not as speedy as cache memory.	It is a fast memory.
Those data or programs are kept here that are not completely get placed in the main memory.	The frequently accessed data is kept in cache memory in order to reduce the access time of files.
Users are able to execute the programs that take up more memory than the main memory.	The time required by CPU to access the main memory is more than accessing the cache. That is the reason frequently accessed data is stored in cache memory so that accessing time can be minimized.