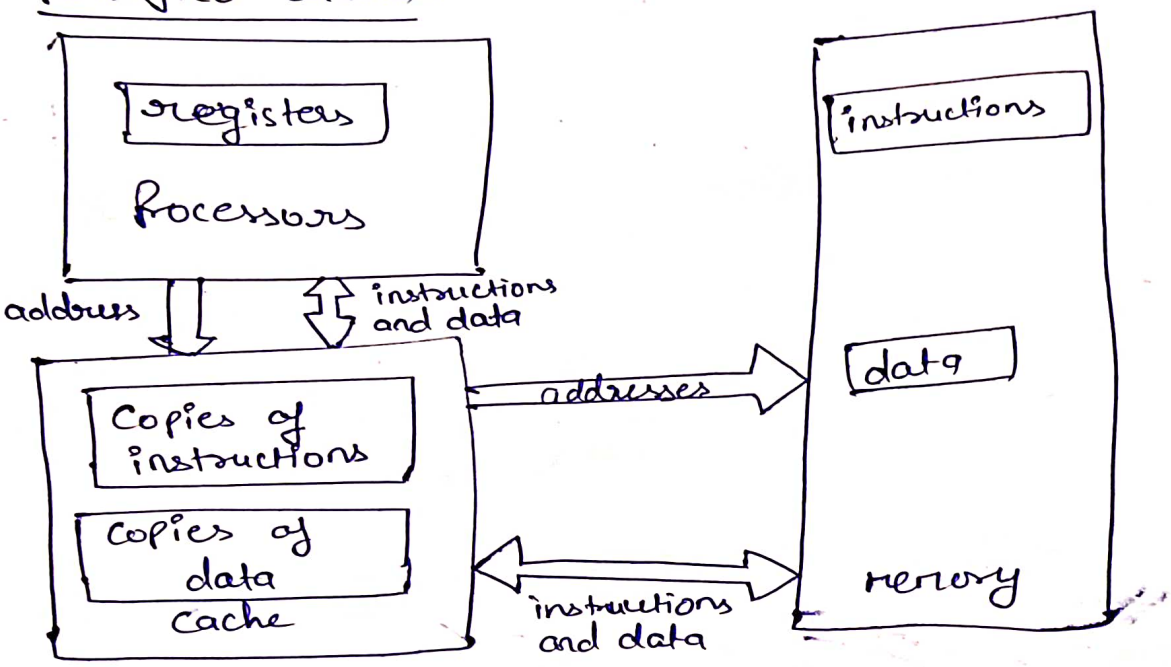
Introduction to different types of Caches:-

(1) Unified Cache:-



→ Cache for data and instructions are combined.
→ The Code and the data accesses can affect each other.
→ It can impact the performance because of cache misses.
→ load / store instructions can cause structural hazards (because cache is common for the instruction and data).

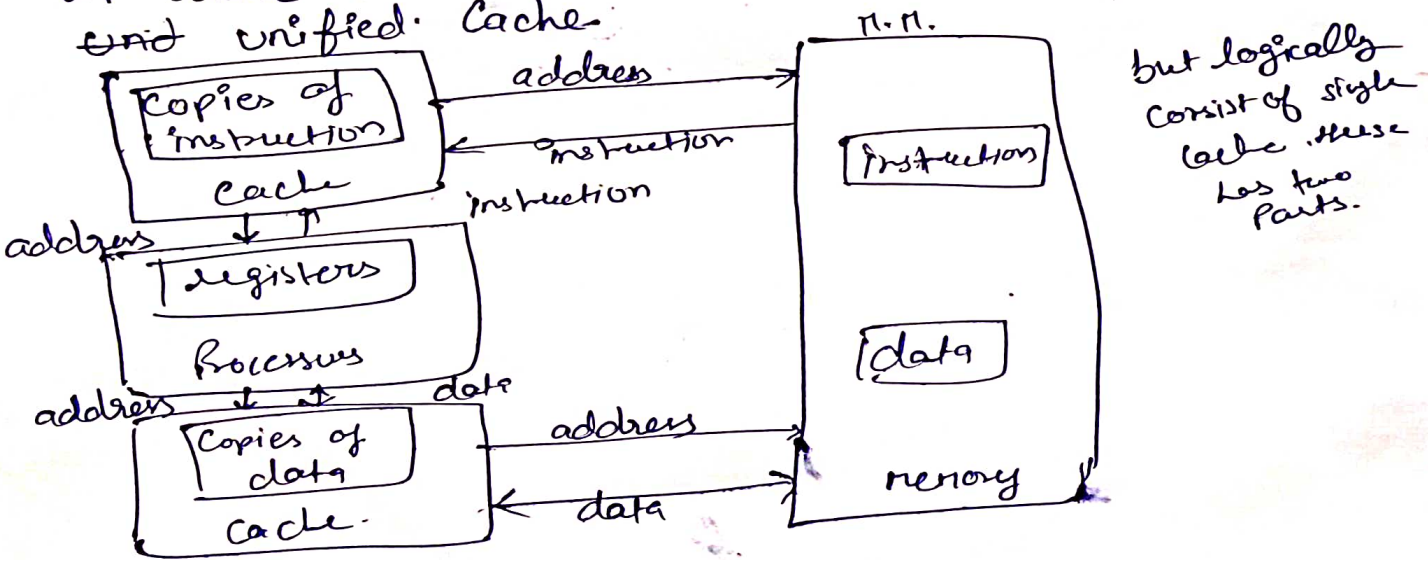
Split Cache or I and D Cache

→ A split cache is a cache that consists of two physically separate parts, where one part, called the instructions cache, is dedicated for holding instructions and the other, called the data cache, is dedicated for holding data.

→ Both of the instruction cache and data cache are logically considered to be a single cache, described as a split cache, because both are hardware managed caches for the same physical address space at the same level of the memory hierarchy.

→ Instruction fetch request are handled only by the Instruction cache and memory operand read and write requests are handled only by the data cache.

→ A cache that is not split is called a ~~and~~ unified cache.



but logically consist of single cache. these has two parts.

The Harvard vs von Neumann :- architecture distinction originally applies to main memory. However, most Modern Computer System Implemented the Modified Harvard architecture. where the L1 cache implements the Harvard architecture and the rest of the memory hierarchy implements the von-Neuman architecture. ~~and the rest of the memory systems~~, the Harvards von neumann distinction Mostly applies to the L1 cache design. (That's why the split cache design is also called the Harvard cache design and the unified cache design is also called Von Neumann.

→ Split I and D → Harvard architecture.
→ specially used for load & store ~~instruction or~~ microprocessors.

Advantage :- The split design is useful for Pipelined processors where the instructions fetch unit and the memory access unit are physically located on diff parts of the chip.

→ When the unified design it is impossible to place the cache simultaneously close to the instructions fetch unit and memory unit, resulting in high cache access latency from one or both units.

The split design enables us to place the instructions cache close to the instruction fetch unit and the data cache close to the memory unit. Hereby simultaneously reducing the latency of both. This is the primary advantage of the split design over the unified design.

(2) Another advantage of the split design is that it allows instruction and data access to occur in parallel without contention. split cache can have double bandwidth of a unified cache. This improves the performance in ~~parallel~~ pipelined processors because instructions and data accesses can occur in the same cycle in diff stage of the pipeline.

Disadvantage :- 1) the combined space of the instruction and data caches may not be efficiently utilized.

2) A cache line that contains both instructions and data may exist in both cache of the same time.

On chip cache:-

On chip cache are a popular technique to fight with the speed mismatch b/w memory and CPU. As integrated circuit become denser, designers have more chip area that can be developed to on-chip caches. To increase the cache size as the available ~~area~~ area also increase may not be the best solution, however. Since the larger a cache, the longer its access time, thus using cache hierachies (two or more level) is potential solutions.

On chip components are the components built on the chip (IC) itself. on the same silicon substrate like transistors, resistors, capacitors, coils etc.

~~off chip~~ on chip components can provide good matching and doesn't consume size (not bulky) and reduce connections required on the PCB.

Unit → 4

Physical Memory (also known as random access memory) is a form of very fast, but volatile data storage. RAM Modules are typically measured in nanoseconds (10^{-3}), and physical disk are typically measured in milliseconds (10^{-1}) This makes Physical Memory roughly 100,000 times faster than a Common Physical disk. ~~Physical Memory~~ Therfore, when possible, Windows and windows Server keep the most frequently accessed Pages of memory in Physical Memory and rely on a disk only if needed. When a system is low on Physical Memory, it often leads to system-wide delays. or, in extreme cases, a Complete hang of the system.

Physical Memory (RAM Modules) Consumes a significant amount of electricity. So reducing the amount of Physical Memory can be a way of saving on Power costs and (or using the RAM for other systems that need it more. The best way to determine if any of the Physical Memory is not serving any use is to see how much memory is free. not be confused with "available" memory.

free memory is the sum of the free page list and the zero page list. free page list is a list of Physical Memory Pages must be cleared and the zero pages that are already clean.

Memory Module:- A memory module is a circuit board that holds memory chips.

Types of Memory Module:-

- SIMMs
- DIMMs

SIMMs:- A Single in-line Memory Module is the most common type of Memory Module. SIMMs can have either 30 or 72 pins, but the 72-pins module is more common in new computers. A bank of SIMMs usually consist of two sockets. So you can install two SIMMs at a time.

DIMMs:- A Dual-in line Memory Module is smaller to SIMMs but has 168 pins. A bank of DIMMs consist of only one socket, so you only need to install one DIMM at a time.

Difference blw SIMMs and DIMMs ?

SIMMs	DIMMs
1) It support 32 bit channels for data transfering	1) It support 64 bit channel for data transfering.
2) SIMMs consumes 5 volts of power	2) DIMMs consumes 3.3 volts of power.
3) SIMMs provides the storage 4MB to 64MB.	3) DIMM provides the storage 32 MB to 1 GB.
4) SIMMs are the older.	4) DIMMs are the replacement of the SIMMs.
5) SIMMs are installed in pairs at a time	5) DIMMs are installed one at a time.
6) SIMMs are used by 486 CPU as well as early Pentium Computers	6) DIMMs are used by Modern Computers.
7) The length and width of SIMMs are 4.25 and 1" inch.	7) The length and width of DIMMs are 1.67" to 5.25" and 1" to 1.75".
8) There are single notches on SIMMs.	8) There are two notches in DIMMs.

Memory Buffer:- A memory buffer register or memory data register is the register in a computer's CPU that stores the data being transferred to and ~~fro~~ from the immediate access storage. It contains the copy of the value in the memory location specified by the memory address register. It acts as a buffer, allowing the processor and ~~as a~~ memory units to act independently without being affected by minor differences in operation. A data item will be copied to the MBR ready for use at the next clock cycle, when it can be either used by the processor for reading and writing, or stored in main memory.

This Register holds the content of the memory which are to be transfered from memory to other components or vice versa.

Memory buffer Register is a two-way-register. When data is fetched from memory and placed into the MDR. it is written to go in one direction. when there is a write instruction. the data to be written is placed into the MDR from another CPU register, which then puts the data into memory.

address space:- Address space is the amount of memory.

The addresses programmers use are called virtual addresses. A set of these virtual address is called address space.

The place where the address was saved in the main memory is know as the location. The group of locations is called memory space.