

UNIT-I

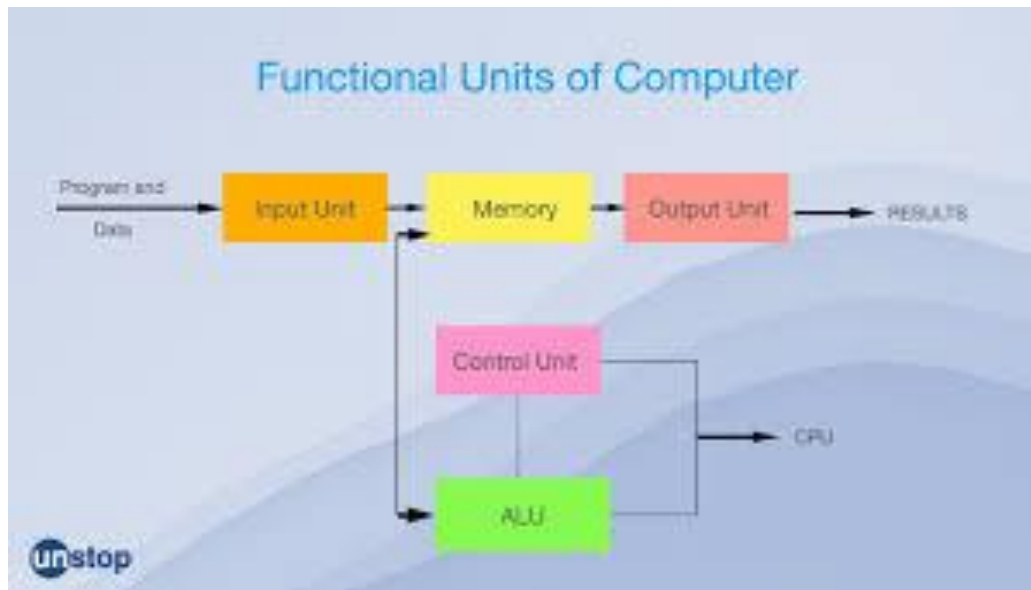
1. Role of Abstraction

- **Definition:** Abstraction in computer science involves hiding complex implementation details and exposing only the necessary and relevant parts of a system to the user or developer.
- **Role:**
 - **Simplification:** Reduces complexity by allowing users to interact with high-level interfaces without dealing with the underlying intricacies. For example, file systems abstract the details of data storage and retrieval.
 - **Modularity:** Encourages breaking down systems into smaller, manageable components or modules. Each module interacts with others through well-defined interfaces.
 - **Reusability:** Abstracted components can be reused across different systems or applications. For instance, a database management system (DBMS) abstracts data storage details, allowing various applications to use it for data manipulation.
 - **Maintainability:** Makes it easier to update or modify system components without affecting other parts of the system. For example, updating an application's user interface does not require changes to the underlying data processing logic.

2. Basic Functional Units of a Computer

- **Central Processing Unit (CPU):**
 - **Arithmetic Logic Unit (ALU):** Performs arithmetic operations (e.g., addition, subtraction) and logical operations (e.g., AND, OR) on data. It is crucial for executing instructions and processing data.
 - **Control Unit (CU):** Directs the operation of the CPU by fetching instructions from memory, decoding them, and executing them. It manages the flow of data between the CPU and other components.

- **Registers:** Small, fast storage locations within the CPU that hold temporary data and instructions. Examples include the accumulator, instruction register, and program counter.



- **Memory:**

- **RAM (Random Access Memory):** Provides fast, temporary storage for data and instructions that are actively being used or processed by the CPU. Volatile memory is lost when power is off.
- **ROM (Read-Only Memory):** Stores firmware and system software that is permanently programmed during manufacturing. Non-volatile, retaining data even when power is off.

- **Input/Output Devices:**

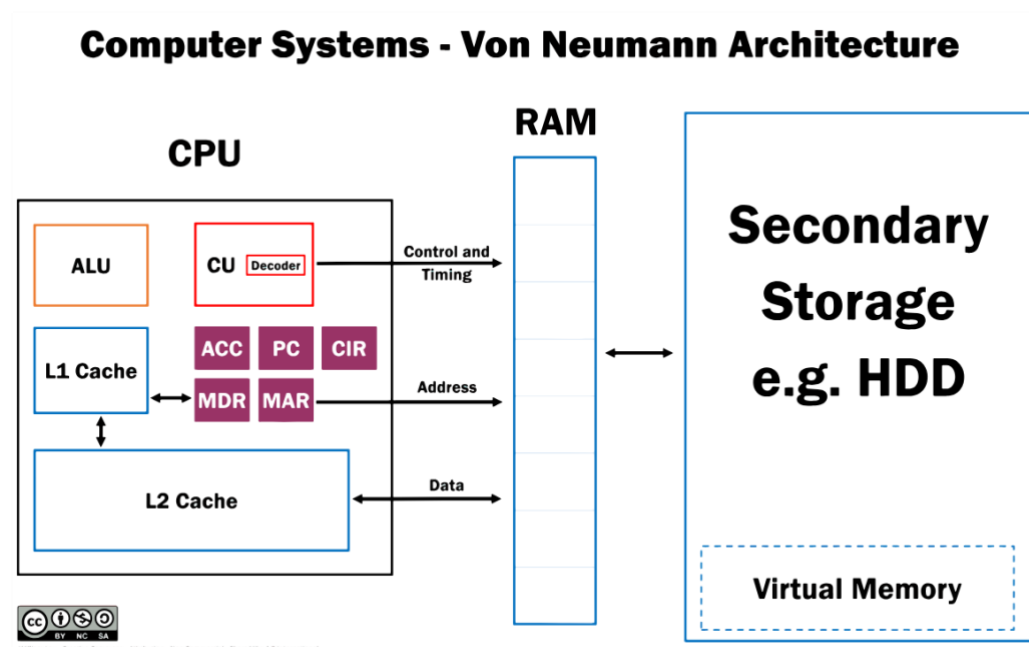
- **Input Devices:** Include keyboards, mice, scanners, and microphones, which allow users to enter data into the computer.
- **Output Devices:** Include monitors, printers, and speakers, which provide data from the computer to the user.

- **Storage:**

- **Hard Drives:** Use spinning platters and magnetic heads to store data magnetically. Provide large storage capacity at relatively low cost.
- **Solid-State Drives (SSDs):** Use flash memory to store data, offering faster access speeds and better durability than hard drives.

3. Von Neumann Model of Computation

- **Overview:** The Von Neumann architecture is a design model for digital computers that describes a system where the CPU, memory, and I/O devices are interconnected and managed through a shared data bus.
- **Components:**
 - **Memory:** Holds both data and instructions in a single, unified memory space. This allows for the stored-program concept, where programs and data are treated similarly.
 - **CPU:** Executes instructions from memory and processes data. The CPU fetches instructions from memory, decodes them, and executes them.
 - **Input/Output (I/O):** Interfaces for communication between the computer and the external environment. I/O operations are managed by the CPU using system calls and interrupts.
 - **Control Unit:** Coordinates the activities of the CPU and other components by managing the execution of instructions and handling the flow of data.



- **Principle:** The architecture operates on the principle that a single set of instructions can be used to perform different types of operations, making it versatile and adaptable for various computational tasks.

4. Moore's Law

- **Definition:** Gordon Moore, co-founder of Intel, observed that the number of transistors on a microchip doubles approximately every two years, leading to an exponential increase in computing power and a decrease in cost per transistor.
- **Impact:**
 - **Performance Improvement:** As the number of transistors increases, processors become more powerful, enabling faster and more complex computations.
 - **Cost Reduction:** The cost per transistor decreases as technology advances, making computing more affordable.
 - **Innovation:** Drives continuous technological innovation and development in computer hardware, influencing design, efficiency, and capability of computing devices.

5. Form Notion and Performance

- **Form Notion:**
 - **Physical Design:** Refers to the physical layout and design of computer components, including size, shape, and packaging. For example, the form factor of a motherboard affects its compatibility with other hardware components.
 - **Ergonomics:** Ensures that computer systems are designed for user comfort and efficiency. This includes considerations like keyboard layout, monitor placement, and cooling solutions.
- **Performance:**
 - **Clock Speed:** Measured in GHz (gigahertz), it represents the frequency at which the CPU executes instructions. Higher clock speeds generally mean faster processing.

- **Instruction Set Architecture (ISA):** Defines the set of instructions that the CPU can execute. The efficiency and capabilities of the ISA impact the performance of software.
- **Number of Cores:** More cores allow a CPU to handle multiple tasks simultaneously, improving multitasking and parallel processing capabilities.
- **Memory Hierarchy:** Includes various levels of memory, such as L1, L2, and L3 caches, RAM, and storage. Faster, smaller caches are closer to the CPU and provide quicker access to frequently used data.

6. Data Representation and Basic Operations

- **Data Representation:**

- **Binary System:** The fundamental representation system for computers, using binary digits (bits) to encode data. Each bit can be 0 or 1.
- **Data Types:** Different types of data, such as integers (whole numbers), floating-point numbers (numbers with decimals), and characters (letters and symbols), are represented using binary encoding.
- **Encoding:** Techniques for representing data in binary form, such as ASCII (American Standard Code for Information Interchange) for text and IEEE 754 for floating-point numbers.

- **Basic Operations:**

- **Arithmetic Operations:** Performed by the ALU, including basic operations like addition, subtraction, multiplication, and division. These operations are fundamental for data processing and manipulation.
- **Logical Operations:** Include operations such as AND, OR, NOT, and XOR, used for binary data manipulation and decision-making processes.
- **Data Movement:** Involves transferring data between registers, memory, and I/O devices. Operations like loading (moving data from memory to a register) and storing (moving data from a

register to memory) are crucial for managing data flow within the system.

These concepts provide a comprehensive understanding of how computers operate, from basic principles and hardware architecture to data representation and system performance.

4o mini

UNIT-II

Instruction Set Architecture (ISA) Overview

1. CPU Registers

- **Definition:** Small, fast storage locations within the CPU used to hold data and instructions temporarily.
- **Types:**
 - **General-Purpose Registers:** Used for arithmetic and logical operations, data manipulation, and temporary storage. Examples in RISC-V include x0 to x31.
 - **Special-Purpose Registers:** Include the Program Counter (PC), Stack Pointer (SP), and Status Register. These manage the flow of instructions and system state.

2. Instruction Format and Encoding

- **Instruction Format:** Defines the layout of an instruction in binary, specifying how various fields are arranged.
- **Fields:** Common fields include opcode (operation code), source and destination registers, immediate values, and function codes.
- **Encoding:** The process of translating an instruction into binary code. For example, in RISC-V, instructions are 32 bits long, with fields for opcode, register addresses, and immediate values.

3. Addressing Modes

- **Immediate Addressing:** Uses a constant value directly in the instruction.

- **Register Addressing:** Uses register values as operands.
- **Direct Addressing:** Specifies the exact memory address where the data is located.
- **Indirect Addressing:** Uses a register to hold the address of the data in memory.

4. Instruction Set

- **Definition:** A collection of instructions that a CPU can execute, defining the operations and their formats.
- **Examples:** Arithmetic operations (add, sub), logical operations (and, or), data transfer (load, store), and control flow (branch, jump).

5. Instruction Types

- **Arithmetic Instructions:** Perform operations like addition, subtraction, multiplication, and division.
- **Logical Instructions:** Perform bitwise operations such as AND, OR, and NOT.
- **Data Movement Instructions:** Transfer data between registers and memory, such as load and store instructions.
- **Control Instructions:** Change the flow of execution, such as branch and jump instructions.

6. Instruction Decoding and Execution

- **Instruction Decoding:** The process of interpreting the binary instruction to determine the operation to be performed. The CPU's control unit decodes the instruction based on the opcode and function codes.
- **Instruction Execution:** The actual execution of the decoded instruction, involving the ALU, registers, and memory as needed.

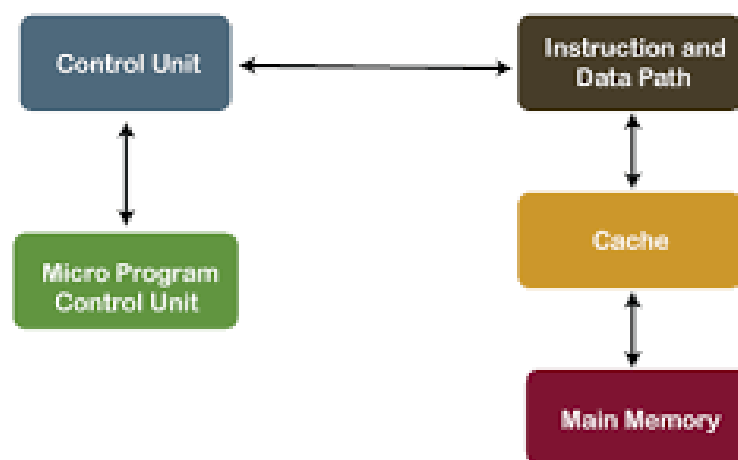
7. Basic Instruction Cycle

- **Fetch:** Retrieve the instruction from memory based on the address in the Program Counter (PC).
- **Decode:** Interpret the fetched instruction to determine the operation and operands.

- **Execute:** Perform the operation specified by the instruction using the ALU or other components.
- **Write Back:** Store the result of the operation back into a register or memory.
- **Update PC:** Update the Program Counter to point to the next instruction.

8. Reduced Instruction Set Computer (RISC)

- **Definition:** A type of ISA characterized by a small, highly optimized set of instructions that are executed in a single cycle.
- **Features:**
 - **Simple Instructions:** Each instruction performs a simple operation.
 - **Fixed Instruction Length:** Instructions have a uniform size, which simplifies decoding.
 - **Register-Based Operations:** Most operations are performed between registers rather than memory.

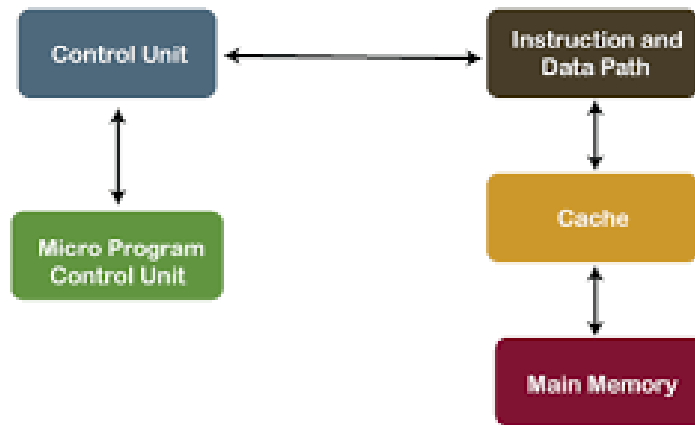


CISC Architecture

9. Complex Instruction Set Computer (CISC)

- **Definition:** A type of ISA with a larger set of instructions, including complex instructions that can perform multiple operations in a single instruction.
- **Features:**

- **Complex Instructions:** Instructions can perform multiple operations or access memory directly.
- **Variable Instruction Length:** Instructions vary in size, which can complicate decoding.

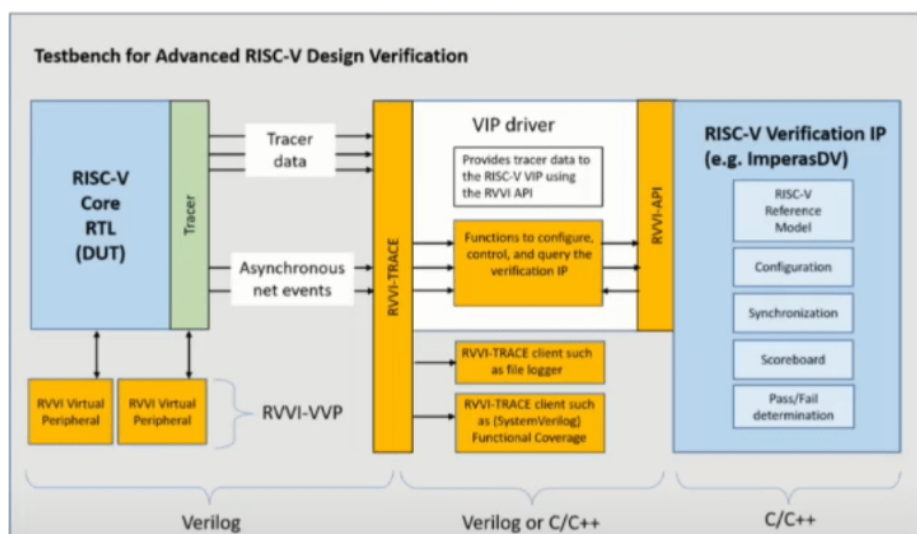


CISC Architecture

-
- **Memory-to-Memory Operations:** Some instructions can directly operate on data in memory.

10. RISC-V Instructions

- **Overview:** An open-source, RISC-based ISA designed for modern processors. It provides a standard set of instructions and is highly extensible.
- **Instruction Types:** Includes arithmetic, logical, load/store, and control instructions.

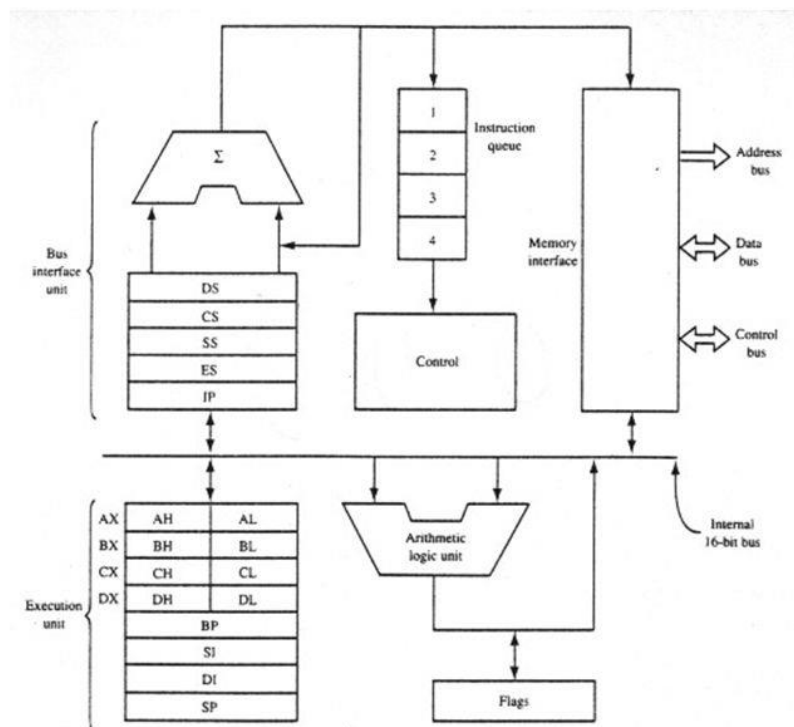


- **Formats:** Includes R-type (register), I-type (immediate), S-type (store), B-type (branch), U-type (upper immediate), and J-type (jump) formats.
- **Extensions:** Supports additional features like floating-point operations (RISC-V-F), atomic operations (RISC-V-A), and vector operations (RISC-V-V).

11. x86 Instruction Set

- **Overview:** A CISC-based ISA used in Intel and AMD processors. It features a wide range of instructions with varying complexities.
- **Instruction Types:** Includes arithmetic, logical, data transfer, control flow, and string manipulation instructions.
- **Formats:** Instructions vary in length and can include complex addressing modes. Features include variable-length instructions and support for memory-to-memory operations.
- **Complexity:** The x86 ISA includes many legacy instructions and complex addressing modes, making it versatile but also more complex to decode and execute.

X86 Architecture



These components and concepts form the foundation of how CPUs execute instructions and manage data, influencing the design and performance of computer systems.

UNIT-III

Processor Concepts Overview

1. Revisiting Clocking Methodology

- **Clocking Methodology:** Refers to how a processor's operations are synchronized using a clock signal.
 - **Clock Signal:** A periodic pulse that coordinates the timing of the processor's operations.
 - **Clock Cycle:** The duration of one cycle of the clock signal, which determines how frequently operations can occur.
 - **Clock Frequency:** The number of clock cycles per second, measured in Hertz (Hz). Higher frequencies mean more operations per second.
- **Clocking Strategies:**
 - **Single-Cycle:** All operations are completed in one clock cycle. Simpler but may require a longer clock cycle to accommodate the slowest operation.
 - **Multi-Cycle:** Different operations are spread across multiple clock cycles. Allows for shorter clock cycles but requires more complex control logic.
 - **Pipelining:** Breaks instructions into stages that can be processed simultaneously, increasing throughput and efficiency.

2. Amdahl's Law

- **Definition:** A principle that estimates the potential speedup of a system when only part of it is improved. It is used to predict the impact of improvements on overall system performance.

- **Formula:** $\text{Speedup} = \frac{1}{(1-P) + \frac{P}{S}}$ Where:
 - P is the fraction of the system that benefits from the improvement.
 - S is the speedup factor of the improved part.
- **Implications:** Highlights diminishing returns on performance improvements. For example, if 90% of a program benefits from an enhancement and the enhancement speeds up that part by 10x, the overall speedup is limited.

3. Building a Data Path and Control

- **Data Path:** The part of the processor that performs data processing operations. It includes components like:
 - **Registers:** Store intermediate values and results.
 - **ALU:** Executes arithmetic and logical operations.
 - **Multiplexers:** Select between different data sources.
 - **Buses:** Transfer data between components.
- **Control:** Manages the operation of the data path by generating control signals based on the instruction being executed.
 - **Control Unit:** Decodes instructions and generates the necessary signals to execute them.

4. Single-Cycle Processor

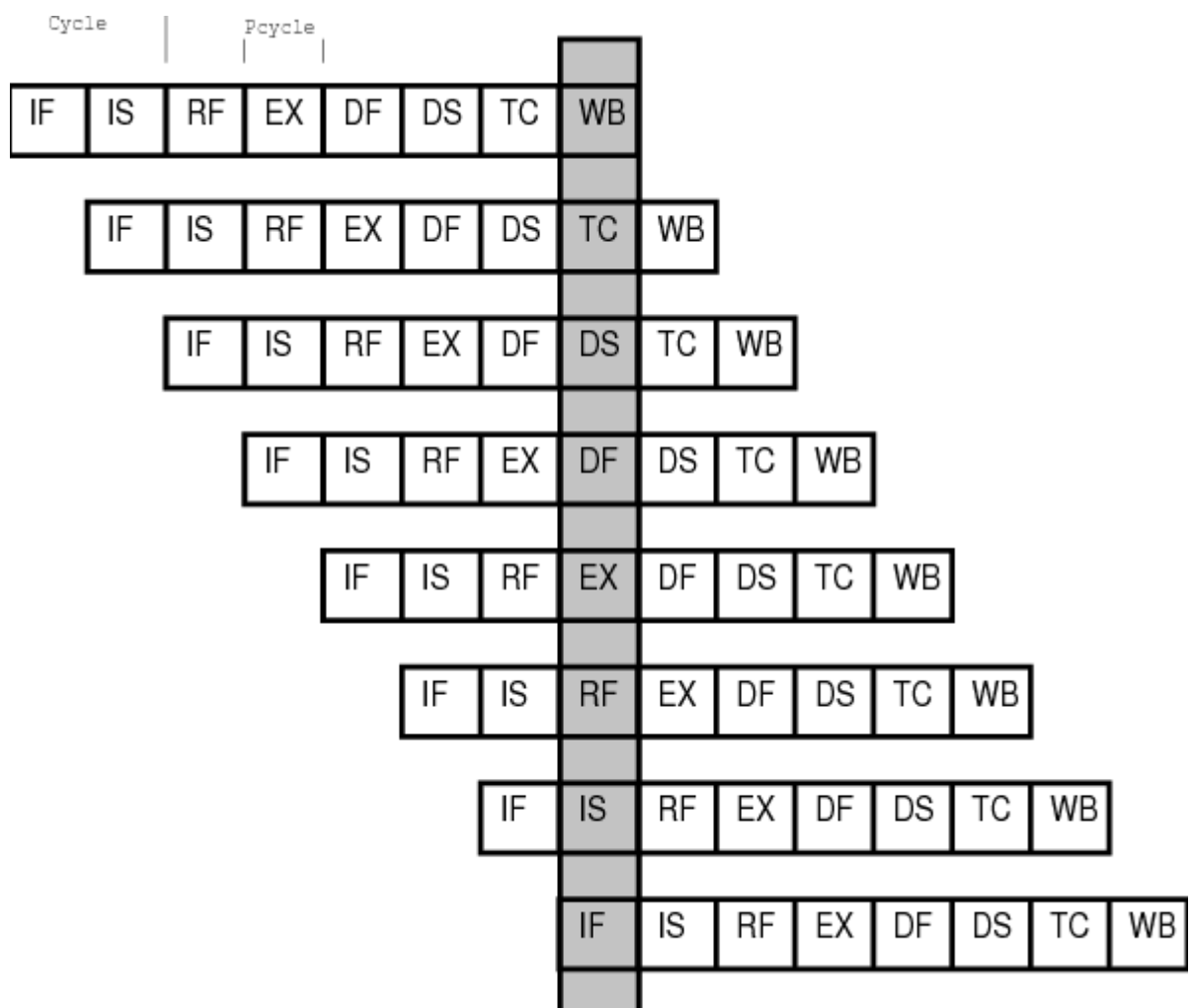
- **Definition:** A processor design where each instruction is executed in a single clock cycle.
- **Characteristics:**
 - **Simpler Design:** Easier to implement as every instruction completes in one cycle.
 - **Longer Clock Cycle:** The clock cycle must be long enough to accommodate the slowest instruction, potentially leading to inefficient utilization of clock cycles.

5. Multi-Cycle Processor

- **Definition:** A processor design where each instruction is broken down into multiple stages, each taking one or more clock cycles.
- **Characteristics:**
 - **Shorter Clock Cycle:** Allows for faster clock cycles since each stage only needs to accommodate part of an instruction's execution.
 - **Complex Control:** Requires more sophisticated control logic to manage the different stages of instruction execution.

6. Instruction Pipelining

- **Definition:** A technique where multiple instructions are overlapped in execution by dividing the process into separate stages.
- **Stages:** Common stages include Instruction Fetch (IF), Instruction Decode (ID), Execute (EX), Memory Access (MEM), and Write Back (WB).



-
- **Benefits:**

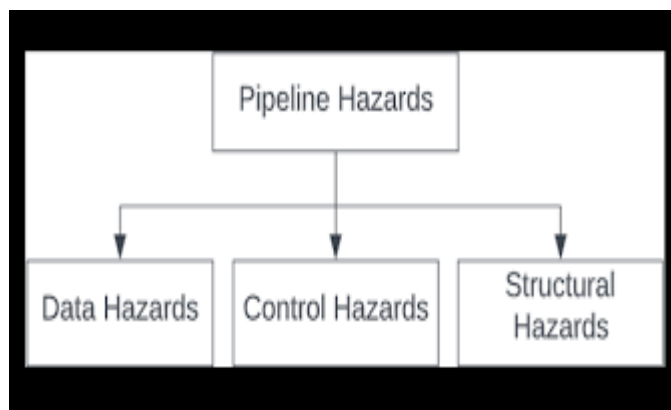
- **Increased Throughput:** More instructions are completed per unit of time.
- **Improved Utilization:** Different stages of multiple instructions can be processed simultaneously.

7. Notion of Instruction-Level Parallelism (ILP)

- **Definition:** The degree to which instructions can be executed simultaneously or in parallel within a processor.
- **ILP Techniques:**
 - **Pipelining:** Executes overlapping stages of multiple instructions.
 - **Superscalar Execution:** Uses multiple execution units to execute several instructions simultaneously.
 - **Out-of-Order Execution:** Executes instructions as resources become available rather than strictly in program order.

8. Data and Control Hazards and Mitigations

- **Data Hazards:** Occur when instructions that are close together in execution depend on the same data.
 - **Types:**
 - **Read-After-Write (RAW):** An instruction needs data that a previous instruction is writing.
 - **Write-After-Read (WAR):** An instruction writes data that a previous instruction is reading.
 - **Write-After-Write (WAW):** Two instructions write to the same location in an incorrect order.



- **Mitigations:**
 - **Stalling:** Pausing the pipeline until the hazard is resolved.
 - **Forwarding:** Using previous results directly to avoid delays.
 - **Reordering:** Adjusting instruction order to minimize hazards.
- **Control Hazards:** Arise from branch instructions that change the flow of control.
 - **Mitigations:**
 - **Branch Prediction:** Guessing the outcome of a branch to keep the pipeline filled.
 - **Delayed Branching:** Scheduling instructions that will execute regardless of the branch outcome.

9. Limits of ILP

- **Limits:**
 - **Hazards:** Data and control hazards limit the extent to which instructions can be executed in parallel.
 - **Dependence Chains:** Long chains of dependent instructions can stall the pipeline and reduce ILP.
 - **Branching:** Frequent branches and unpredictable control flow can reduce the effectiveness of pipelining and ILP techniques.
 - **Resource Conflicts:** Limited execution units and memory bandwidth can constrain ILP.
- **Diminishing Returns:** Increasing ILP beyond a certain point yields diminishing returns due to these inherent limitations, making it challenging to achieve significant performance gains.

These concepts cover various aspects of processor design and performance optimization, providing a comprehensive understanding of how modern processors execute instructions and manage data.

UNIT-IV

Memory Hierarchy Overview

1. SRAM vs. DRAM

- **SRAM (Static Random Access Memory):**
 - **Characteristics:** Faster and more reliable than DRAM, as it does not need to be refreshed. Uses flip-flops to store each bit of data.
 - **Usage:** Commonly used for cache memory due to its speed and reliability.
 - **Cost:** More expensive per bit than DRAM due to its complexity and faster speed.
 - **Size:** Typically smaller in capacity compared to DRAM.
- **DRAM (Dynamic Random Access Memory):**
 - **Characteristics:** Slower than SRAM and needs to be refreshed periodically to maintain data. Uses capacitors to store bits of data, which can leak over time.
 - **Usage:** Used for main memory (RAM) in computers due to its higher density and lower cost.
 - **Cost:** Less expensive per bit than SRAM, allowing for larger capacities.
 - **Size:** Typically larger in capacity compared to SRAM.

SRAM vs. DRAM	
STATIC RAM	DYNAMIC RAM
Does not require refreshing	Must be continuously refreshed
Requires multiple transistors to store one bit	Requires one transistor and one capacitor to store one bit
Delivers faster access times	Delivers slower performance
Consumes less power, especially in idle	Consumes more power
Takes up more space	Requires less space
Can hold only a small amount of data	Can hold much more data
Costs more than DRAM	Costs less than SRAM
Typically used for a processor's cache	Typically used for a computer's main memory

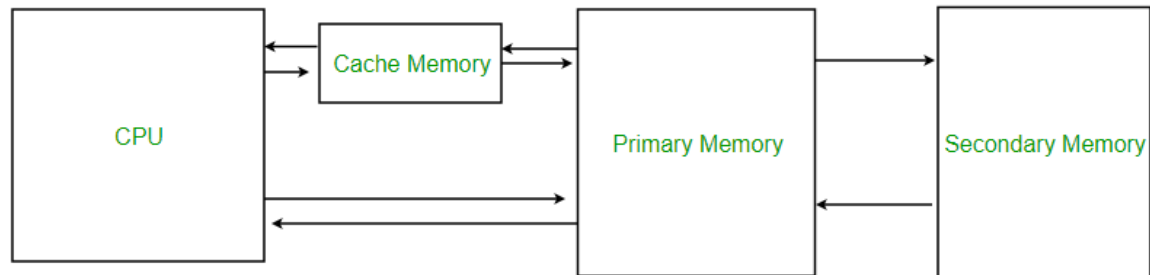
2. Locality of Reference

- **Temporal Locality:** Refers to the tendency of a program to access the same memory locations repeatedly within a short period.
- **Spatial Locality:** Refers to the tendency of a program to access memory locations that are close to each other within a short period.
- **Importance:** Locality of reference is leveraged in memory hierarchy design to improve performance by keeping frequently accessed data in faster, closer memory (e.g., caches).

3. Caching

- **Definition:** A technique to store frequently accessed data in a smaller, faster memory (cache) to reduce access time.
- **Indexing Mechanisms:**
 - **Direct-Mapped Cache:** Each memory block maps to exactly one cache line. Simple but can cause frequent cache misses if multiple blocks map to the same line.
 - **Fully Associative Cache:** Any block can be placed in any cache line. More flexible but requires complex lookup mechanisms.

- **Set-Associative Cache:** A compromise between direct-mapped and fully associative. Cache lines are divided into sets, and each block maps to a set. Example: 4-way set-associative means each set has 4 lines.



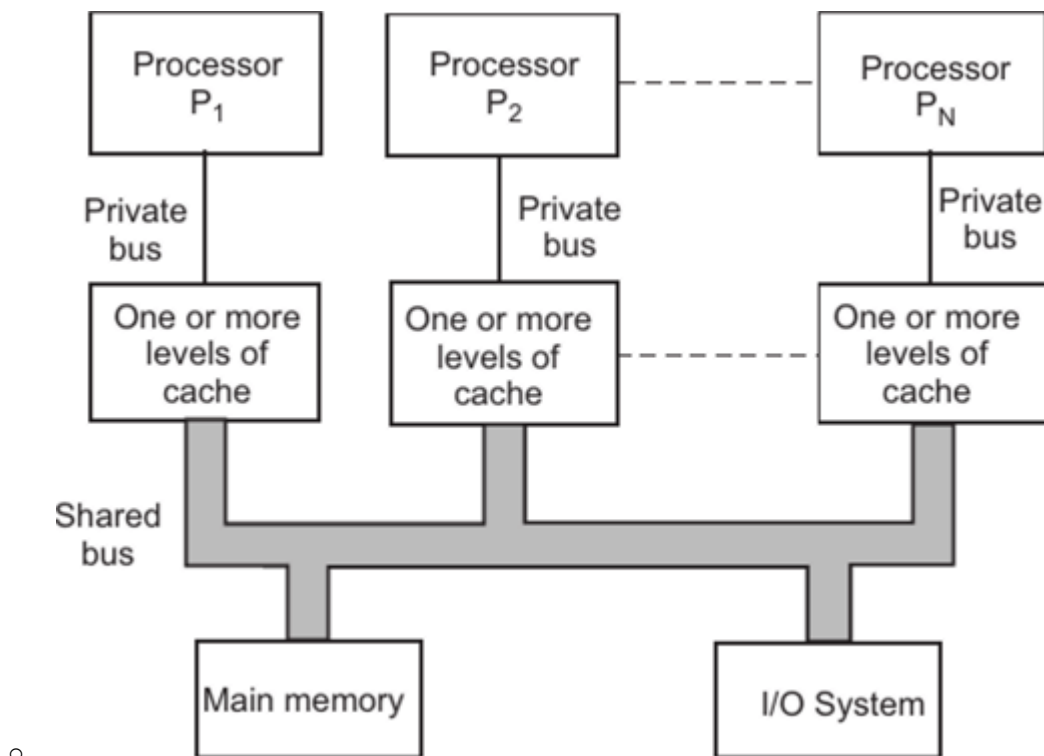
-
- **Trade-Offs:**

- **Block Size:** Larger blocks can exploit spatial locality but may increase the cache miss penalty and wastage due to overhead.
- **Associativity:** Higher associativity reduces conflict misses but increases the complexity and access time of the cache.
- **Cache Size:** Larger caches can store more data but may increase latency and power consumption.

4. Processor and Cache Interactions for Read/Write Requests

- **Read Request:**

- **Cache Hit:** Data is found in the cache, reducing access time.
- **Cache Miss:** Data is not in the cache. The cache controller fetches the data from the next level of memory (e.g., main memory) and updates the cache.



- **Write Request:**

- **Write-Through Cache:** Writes data to both the cache and the main memory. Ensures data consistency but may increase latency.
- **Write-Back Cache:** Writes data only to the cache. The cache line is marked dirty, and the data is written back to main memory only when the line is replaced. Reduces the number of writes to memory but can lead to data inconsistency if not managed correctly.

5. Basic Optimizations

- **Write-Through Cache:** Each write operation updates both the cache and the main memory. Ensures that memory is always up-to-date but can cause higher write traffic to main memory.
- **Write-Back Cache:** Writes data only to the cache and defers updates to the main memory until the cache line is evicted. Reduces the number of write operations to main memory but requires mechanisms to manage consistency.

6. Average Memory Access Time

- **Definition:** The average time it takes to access memory, considering cache hits and misses.

- **Formula:**

Average Memory Access Time=(Hit Time×Hit Rate)+(Miss Time×Miss Rate)

$$\text{Average Memory Access Time} = (\text{Hit Time} \times \text{Hit Rate}) + (\text{Miss Time} \times \text{Miss Rate})$$

- **Components:**

- **Hit Time:** Time to access data from the cache.
- **Miss Time:** Time to fetch data from the next level of memory, including cache access and memory access time.

7. Cache Replacement Policies

- **Least Recently Used (LRU):** Replaces the cache line that has not been used for the longest time. Balances between recency and frequency of access.
- **First-In, First-Out (FIFO):** Replaces the oldest cache line. Simple but may not always select the optimal line to replace.
- **Random Replacement:** Randomly selects a cache line to replace. Can be simple to implement but may perform poorly in some cases.

8. Memory Interleaving

- **Definition:** A technique to distribute memory addresses across multiple memory banks to improve access performance.
- **Method:**
 - **Interleaving:** Memory addresses are spread across multiple banks, allowing simultaneous access to different banks. Reduces contention and improves parallelism.
- **Benefits:** Enhances bandwidth and reduces access time by enabling concurrent accesses to different memory banks.

These concepts highlight the importance of memory hierarchy in optimizing access times and improving overall system performance. Each level of the hierarchy, from registers to main memory, plays a crucial role in managing and accessing data efficiently.

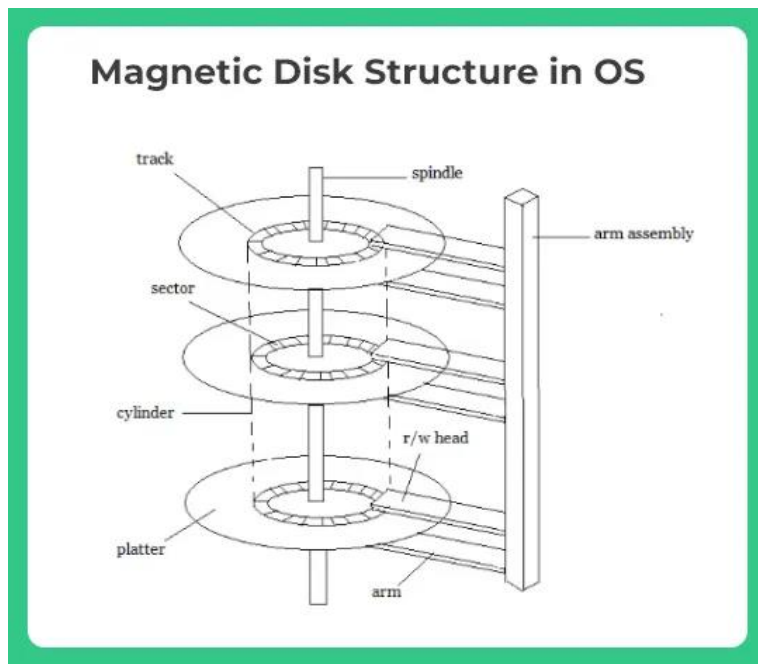
UNIT-V

Storage and I/O Overview

1. Introduction to Magnetic Disks

- **Magnetic Disks:**

- **Definition:** A type of non-volatile storage device that uses magnetic storage to record data. It includes hard disk drives (HDDs) and floppy disks.
- **Components:**
 - **Platters:** Circular disks coated with a magnetic material. Data is stored in concentric tracks.
 - **Read/Write Heads:** Devices that read from and write to the platters.
 - **Spindle:** Rotates the platters at high speeds.



- **Characteristics:**

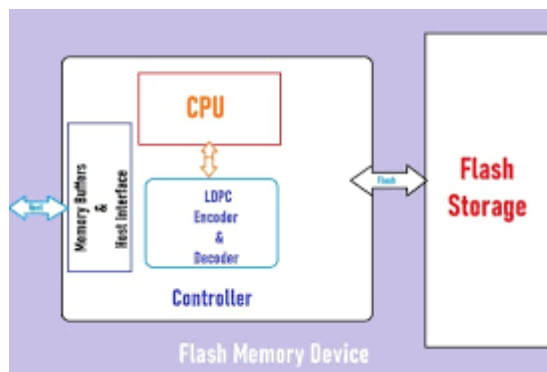
- **Capacity:** Typically large, suitable for long-term storage.
- **Speed:** Slower compared to SSDs due to mechanical movement.

- **Durability:** More prone to physical damage compared to solid-state storage.

2. Flash Memory

- **Flash Memory:**

- **Definition:** A type of non-volatile memory that retains data without power, used in solid-state drives (SSDs), USB drives, and memory cards.
- **Types:**
 - **NAND Flash:** Used for high-capacity storage, such as in SSDs. Data is stored in memory cells organized in a grid.
 - **NOR Flash:** Used for code storage and execution, such as in firmware. Provides faster read speeds but is more expensive and less dense than NAND.



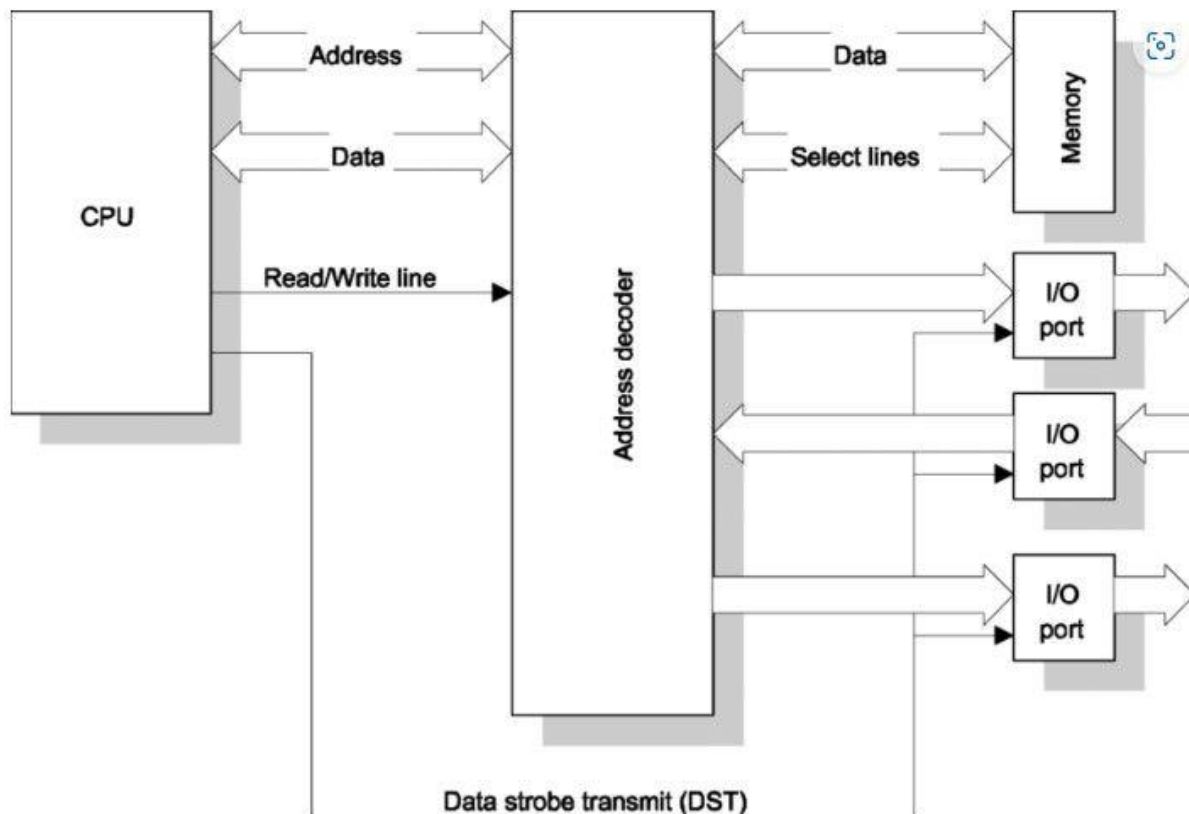
- **Characteristics:**
 - **Capacity:** Varies from gigabytes to terabytes.
 - **Speed:** Faster read/write speeds compared to magnetic disks.
 - **Durability:** More resistant to physical shock compared to magnetic disks.

3. I/O Mapped I/O and Memory Mapped I/O

- **I/O Mapped I/O:**

- **Definition:** A technique where I/O devices are assigned specific addresses in a separate I/O address space.

- **Access:** I/O instructions (such as IN and OUT in x86 assembly) are used to interact with these devices.
- **Pros:** Simplifies I/O device address management.
- **Cons:** Limited address space for I/O operations and requires special instructions.

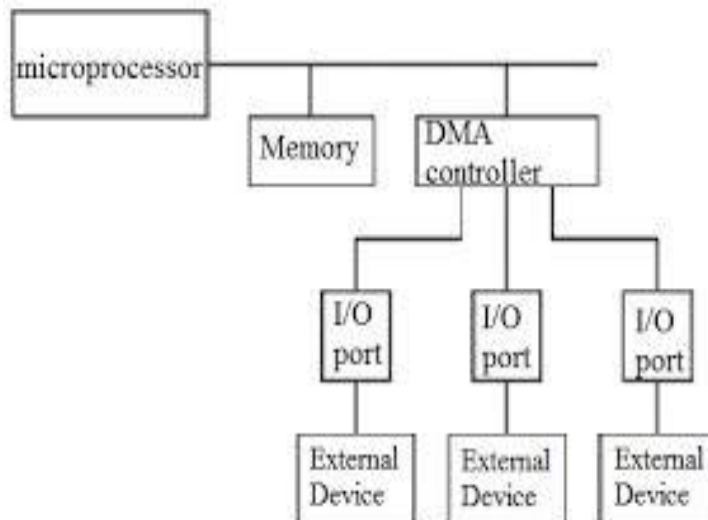


- **Memory Mapped I/O:**

- **Definition:** A technique where I/O devices are mapped to the same address space as regular memory.
- **Access:** Standard memory instructions (such as LOAD and STORE) are used to interact with I/O devices.
- **Pros:** Uses the same instructions as for memory, allowing for easier programming and addressing.
- **Cons:** Reduces the available address space for actual memory if I/O devices occupy part of the address space.

4. I/O Data Transfer Techniques

- **Programmed I/O:**
 - **Definition:** A technique where the CPU actively manages I/O operations by polling the I/O device to check its status and transferring data manually.
 - **Characteristics:**
 - **CPU Involvement:** High CPU overhead as the CPU waits and checks for I/O operations.
 - **Simplicity:** Simple to implement but inefficient for large data transfers.
- **Interrupt-Driven I/O:**
 - **Definition:** A technique where the I/O device interrupts the CPU when it is ready to transfer data, allowing the CPU to perform other tasks in the meantime.
 - **Characteristics:**
 - **Efficiency:** Reduces CPU idle time and improves performance compared to programmed I/O.
 - **Complexity:** Requires handling of interrupt requests and associated context switching.
- **Direct Memory Access (DMA):**
 - **Definition:** A technique where an external DMA controller manages data transfers between memory and I/O devices, bypassing the CPU.



○

○ **Characteristics:**

- **Efficiency:** Offloads data transfer work from the CPU, allowing it to focus on other tasks.
- **Speed:** Provides high-speed data transfers with minimal CPU intervention.
- **Complexity:** Requires setup of DMA channels and control registers but provides significant performance improvements for large data transfers.

These concepts cover various aspects of storage and I/O systems, highlighting the different technologies and methods used to manage data and interact with peripheral devices in computing systems.