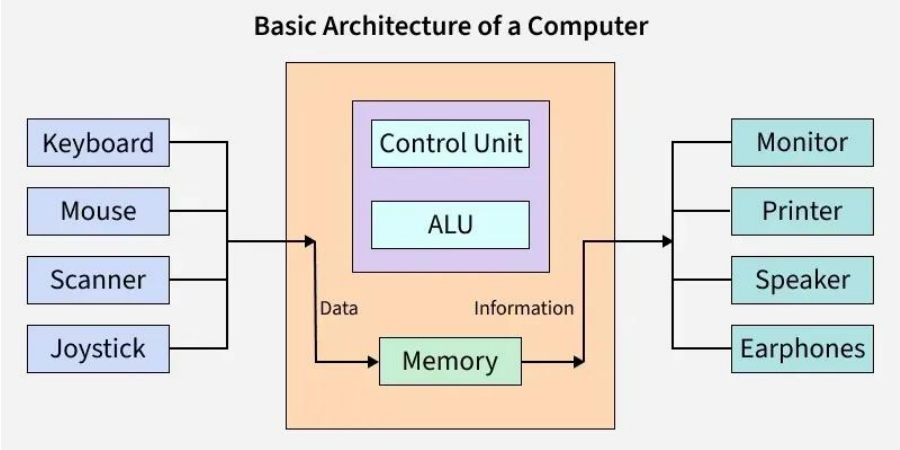
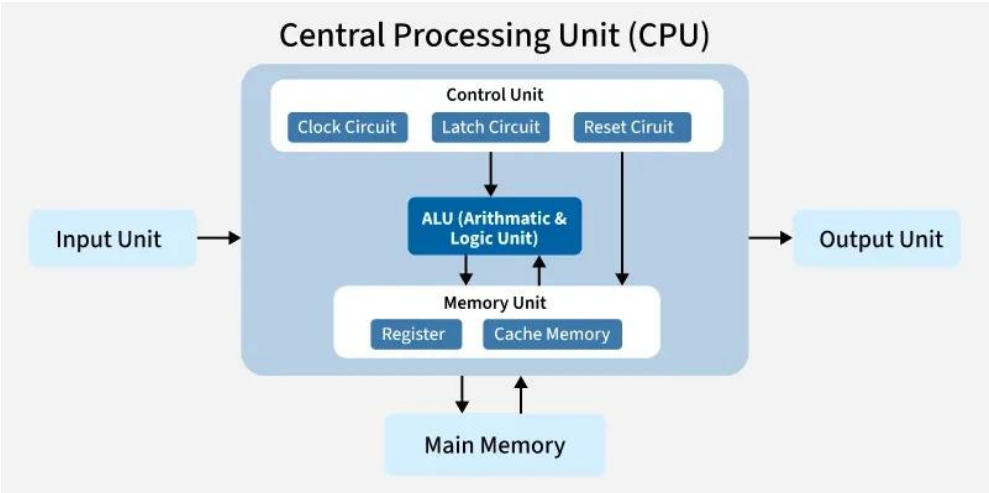
Computer Architecture, CPU, Cores, RAM, and Cache

Computer Architecture



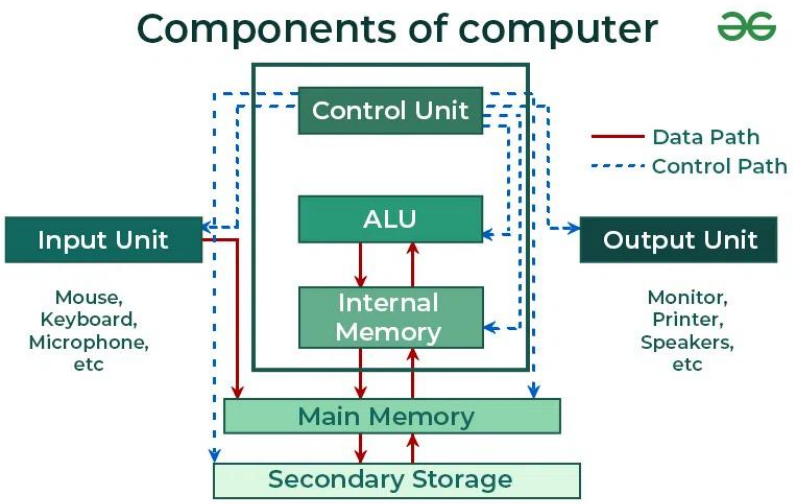
Computer Architecture defies how a computer's components communicate through electronic signals to perform input, processing, and output operations.
It is typically broken down into three main categories:
1. Instruction Set Architecture (ISA): The abstract reference between the hardware and the software.
2. Microarchitecture (Computer Organization):The detailed physical implementation of the ISA, including the internal structure of the CPU and the design techniques used to achieve performance.
3. System Design: All other hardware components within the system, such as the memory hierarchy, bus structure, and I/O mechanisms.

CPU



The Central Processing Unit (CPU), often called the "brain" of the computer, is the primary component that performs arithmetic, logical, and input/output (I/O) operations by executing program instructions. Its core function is a continuous Fetch-Decode-Execute-Store cycle (also known as the Instruction Cycle).

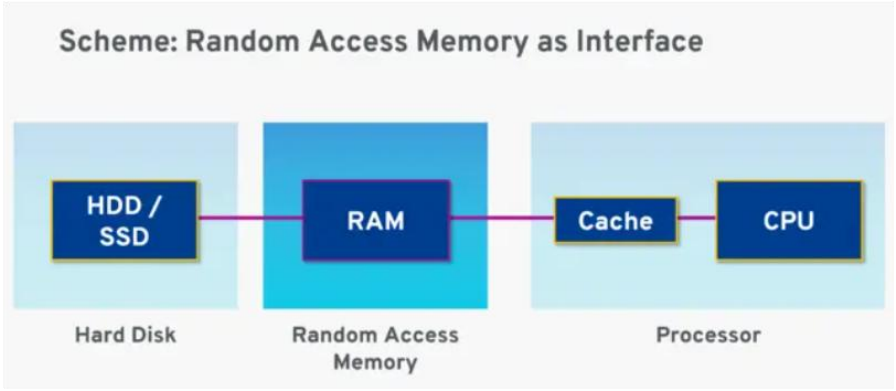
CPU Cores



A core is an individual processing unit within the CPU. Modern CPU s are typically muti-core processors, meaning they contain two or more independent cores.

1. Function: Each core operates independently, allowing the CPU to handle multiple tasks or threads simultaneously (parallel processing), which significantly improves multitasking and overall system performance.
2. Performance Impact: A CPU with more cores can efficiently run more demanding software or handle more background processes than a single-core CPU, all else being equal.

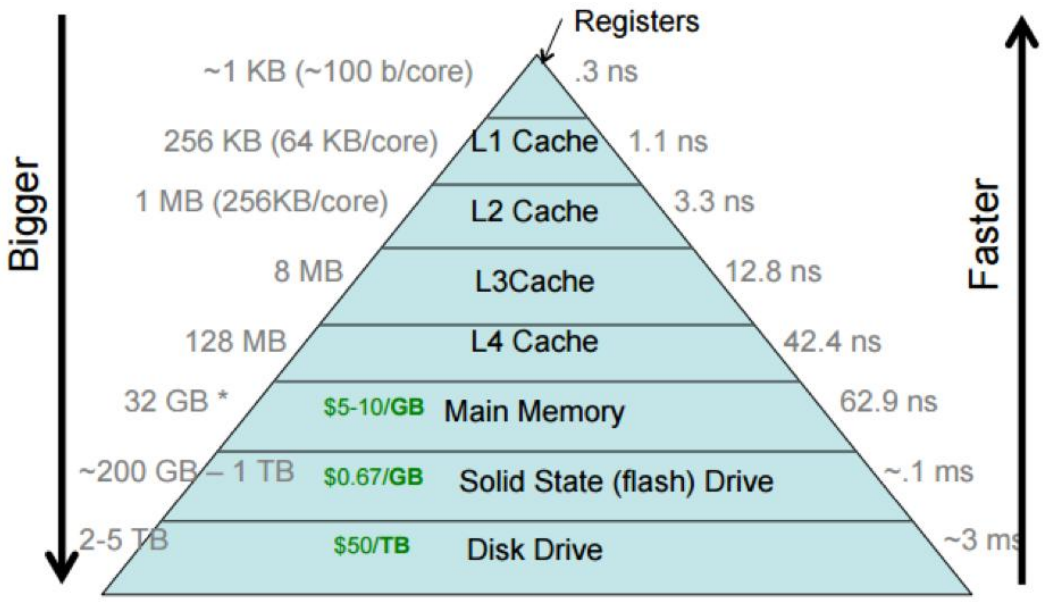
RAM



Random Access Memory (RAM) is the computer's primary memory (also called main memory) that serves as temporary, high-speed storage for data and program instructions currently being used by the CPU.

1. Volatile: RAM is volatile, meaning all data stored in it is lost when the computer's power is turned off.
2. Role in System: When you open an application or file, it is loaded from the slower storage (like an SSD or HDD) into RAM so the CPU can access it quickly for processing.

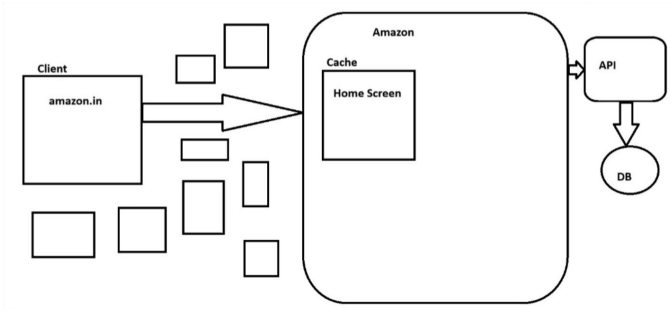
Cache Memory



Cache is a small, very fast memory buffer between the Central Processing Unit (CPU) and slower Random Access Memory (RAM). It stores frequently used data and instructions to minimize the time the CPU spends waiting for information, which is crucial for performance.

The memory is organized in a hierarchy of speed and size:

1. L1 Cache: The fastest and smallest level, typically split into instruction and data cache, and dedicated to a single CPU core.
2. L2 Cache: Larger and slightly slower than L1. It may be dedicated to a core or shared between a small group of cores.
3. L3 Cache: The largest and slowest cache level, but still faster than RAM. It is typically shared by all cores on the processor.



This cache diagram illustrates how frequently accessed data, such as the Home Screen, is served directly from the Cache layer within the Amazon server, bypassing the slower API and DB (Database) layers. Historically, client visits took longer because the system had to access the database for every request. By implementing caching, the most visited screens are delivered almost instantly, drastically improving response time and making the overall client experience much faster.