Hardware

Moore’s Law and More

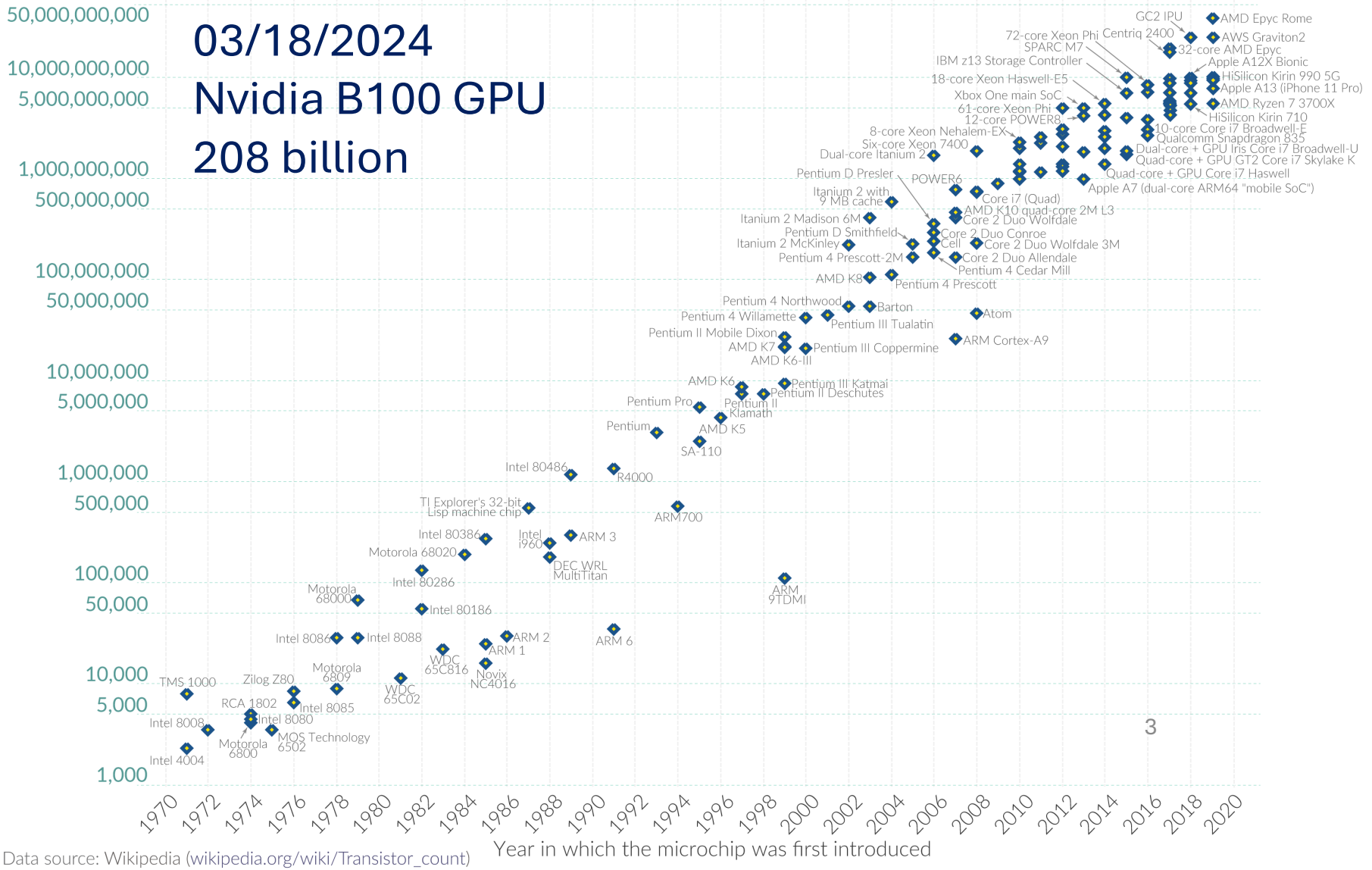
History

- Early Generations (1946-1970)
  - Vacuum tubes
  - Transistors
  - Integrated Circuits(IC)
- Microprocessor Era (1971 – present)
- Fifth generation (next 10 years?) : AI and GPU
- Future: Quantum computers

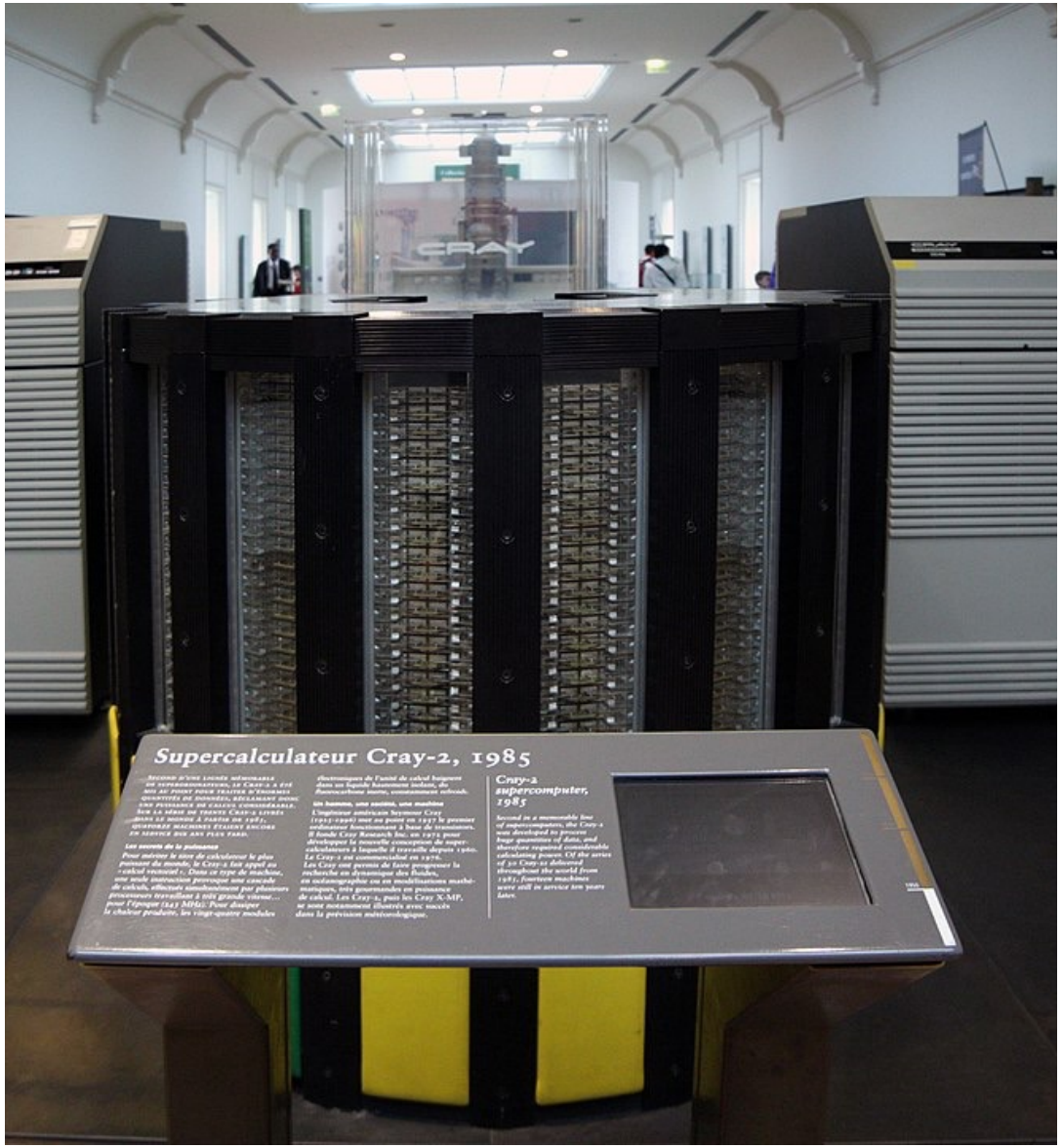
Moore's Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Our World in Data



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)


CRAY
Supercalculateur Cray-2, 1985
Cray-2
supercomputer,
1985

Cray -2	iPhone 15
Released in 1985	Released in 2023
$16 million (Today's dollar $41 million)	$1,000
1.9 gigaflops (billion operations per second)	20 Teraflops (trillion operations per second)

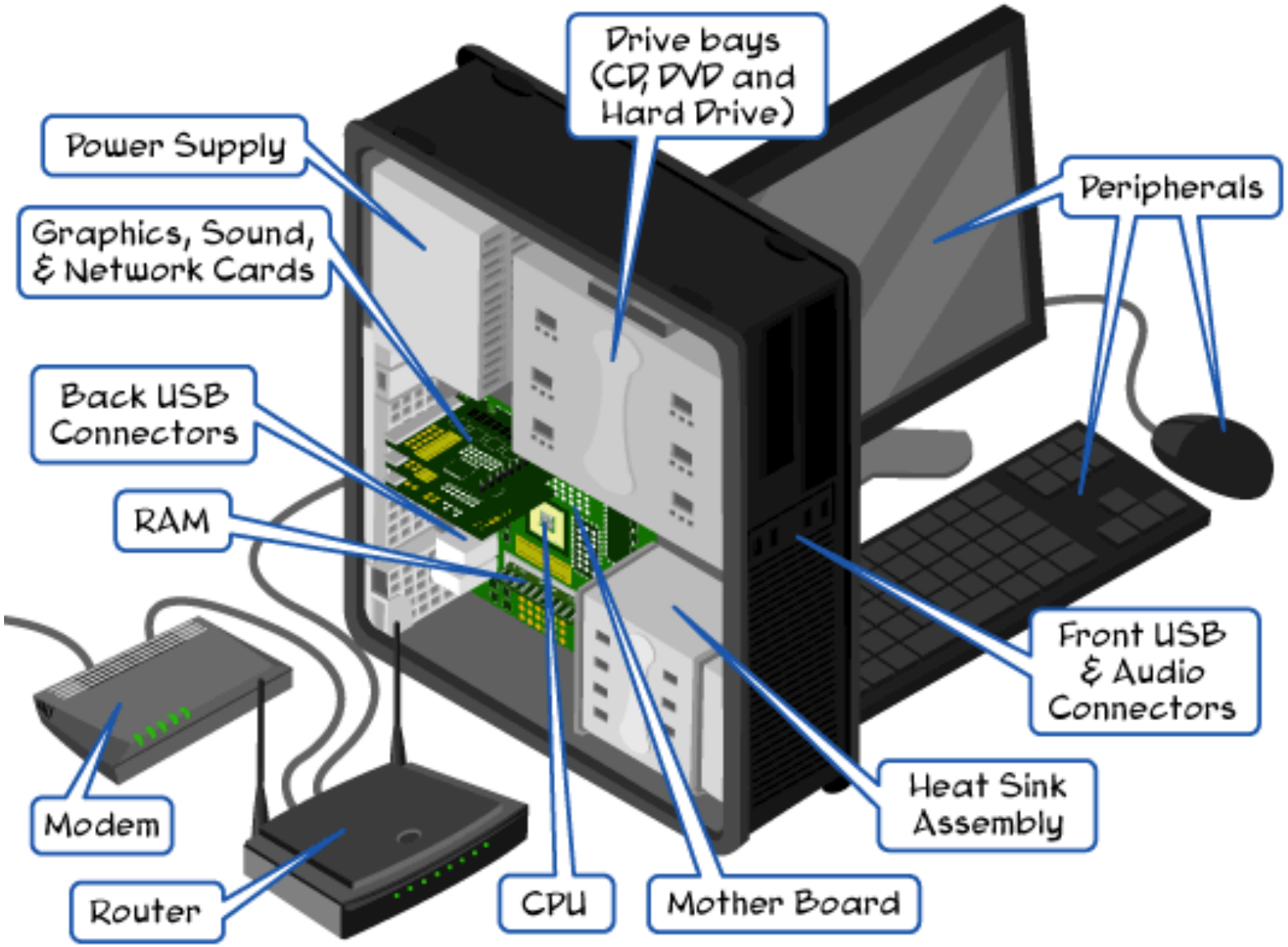
iPhone 15 is approximately 10,000 times more powerful than the Cray-2.

In 1985, the total computational power in US would be like a single modern iPhone can achieve.

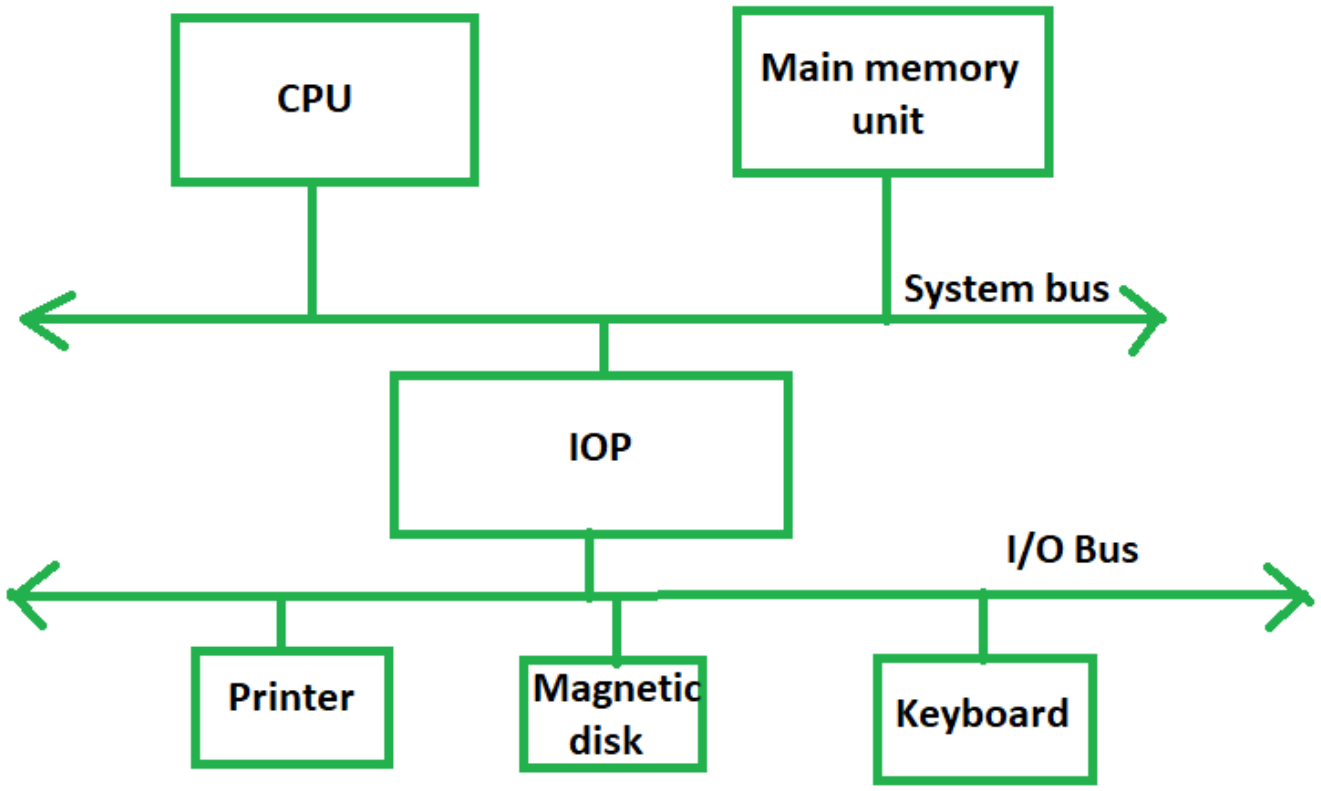
It Means Something

The Impact

- The rapid advancement and democratization of computational power is unparalleled in human history.
- This transformation underscores the critical role of information technology as a cornerstone of contemporary business strategy and economic growth.

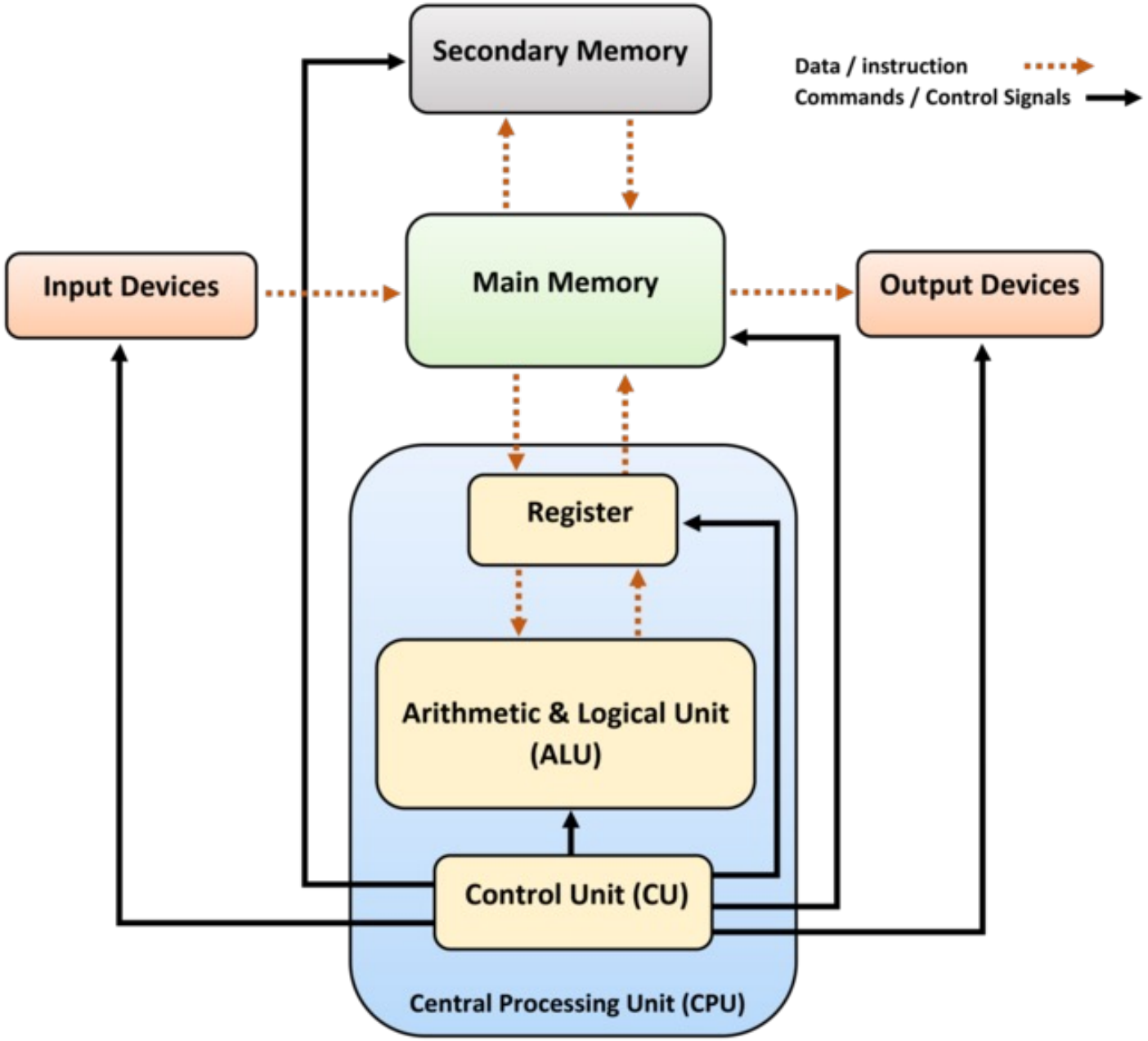

Drive bays (CD, DVD and Hard Drive)
Power Supply
Graphics, Sound, & Network Cards
Back USB Connectors
RAM
Peripherals
Front USB & Audio Connectors
Heat Sink Assembly
Modem
Router
CPU
Mother Board

Hardware Components



Input-Output Processor

Computer Architecture



A Hierarchical Architecture

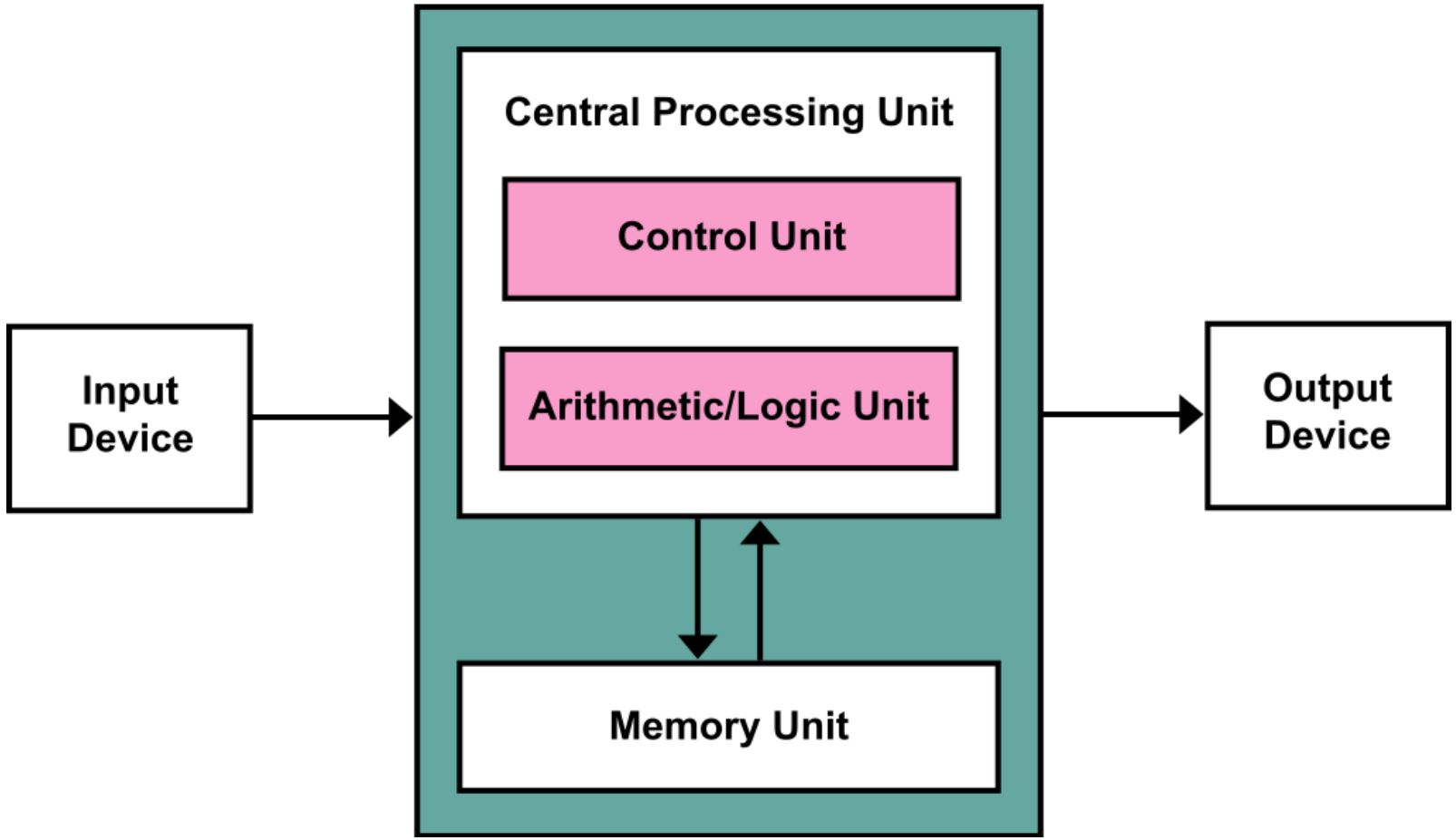
- CPU: 1-2 clock cycles (approximately 0.25-0.5 nanoseconds)
- CPU cache, which is a small, high-speed memory located close to the CPU. With access times of 2-10 clock cycles (approximately 0.5-2.5 nanoseconds)
- Memory (RAM) is the next level in the hierarchy, with access times of 50-200 clock cycles (approximately 12.5-50 nanoseconds).
- Disk storage is the slowest component, with access times measured in milliseconds (approximately 1,000,000-10,000,000 nanoseconds)
- Latency Numbers: https://colin-scott.github.io/personal_website/research/interactive_latency.html

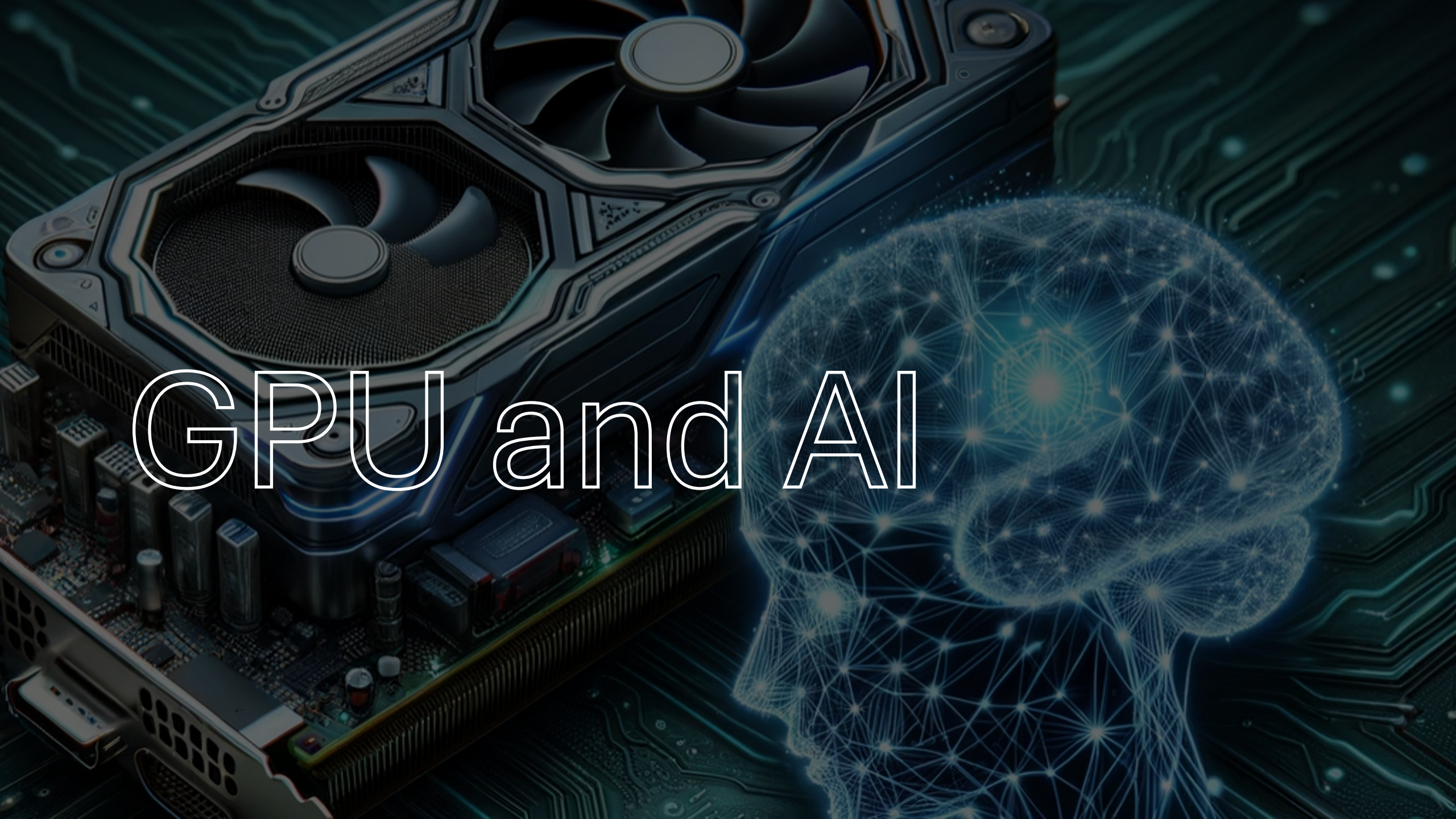
Big RAM and Fast SSD

- This hierarchy of speeds leads to a hierarchical architecture design, where each component plays a crucial role in optimizing performance.
- A balanced computer system ensures that the various components – cache, RAM, and SSD – are proportionally allocated to prevent bottlenecks and maximize efficiency.
- Different Applications
  - Video editing software like Adobe Premiere Pro requires large amounts of RAM.
  - 3D modeling and rendering in software such as Autodesk Maya benefit significantly from a fast CPU.

Von Neumann Architecture

- Program and data stored in memory.
- Fetch-Execute Cycle.
- Program Counter
- Memory Addressing




GPU and AI

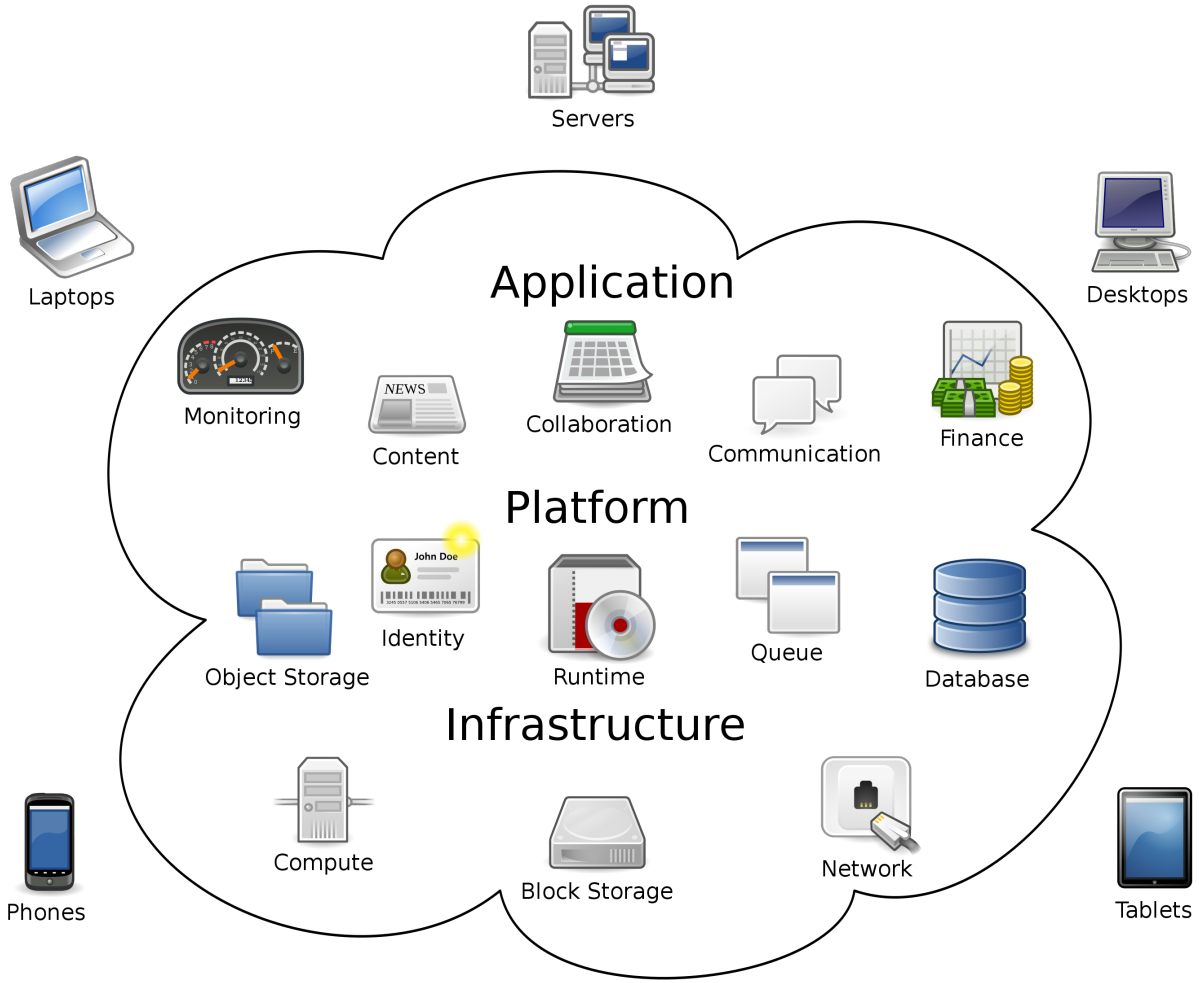
GPU

- GPUs are designed to perform many simple, repetitive operations simultaneously.
- This makes them ideal for the matrix and vector computations that are fundamental to neural networks.

GPU Demo: https://www.youtube.com/watch?v=-P28LKWTzrI

Cloud Computing

- Cloud computing represents a paradigm shift in how computing resources are managed and utilized.
- It involves the virtualization of physical components such as CPUs, RAM, and storage disks to provide scalable and efficient solutions.
- This virtualization allows users to dynamically scale their resources, optimizing performance and cost efficiency.
- Amazon AWS, Microsoft Azure, and Google Cloud Platform (GCP) are top three cloud computing service providers.


Servers
Laptops
Desktops
Application
Monitoring
NEWS
Content
Collaboration
Communication
Finance
Platform
John Doe
Identity
Object Storage
Runtime
Queue
Database
Infrastructure
Compute
Block Storage
Network
Phones
Tablets

Cloud Features

- **On-Demand Self-Service**: Users can provision and manage computing resources as needed, without requiring human intervention from the service provider.
- **Broad Network Access**: Services are available over the network and can be accessed through standard mechanisms by a variety of client devices, such as laptops, tablets, and smartphones.
- **Rapid Elasticity**: Resources can be quickly and elastically provisioned, in some cases automatically, to scale out and scale in rapidly, commensurate with demand.
- **Measured Service**: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth).

Cloud Tradeoffs

- **Downtime**: Since cloud computing systems are internet-based, service outages are always an unfortunate possibility and can occur for any reason.
- **Security and Privacy**: While cloud providers often have robust security measures, storing data and critical files on external servers always opens up risks. Sensitive data must be carefully monitored and managed to ensure compliance with relevant regulations and standards.
- **Cost Management and Unpredictability**: Although cloud computing can be cost-effective, without proper management, costs can quickly escalate due to pay-as-you-go pricing models and overuse of resources.
- **Limited Control and Flexibility**: Cloud users have limited control over the backend infrastructure. While it is easy to use and manage, it may not offer the same level of customization as a traditional in-house setup.
- **Vendor Lock-In**: Moving services from one cloud provider to another can be complicated. Organizations may face challenges related to data migration, compatibility issues, and different APIs, which can limit flexibility and increase costs.

Case Study: ATM