Assignment 2

Q) Compare GPU and CPU chips in terms of their strengths and weakness. In particular discuss the tradeoffs between power efficiency, programmability and performance. Also compare various MPP architecture in processor selection performance target efficiency and packaging constraints.

CPUs and GPUs have a lot in common. Both are critical computing engines. Both are silicon-based microprocessors. And both handle data. But CPUs and GPUs have different architectures and are built for different purposes.

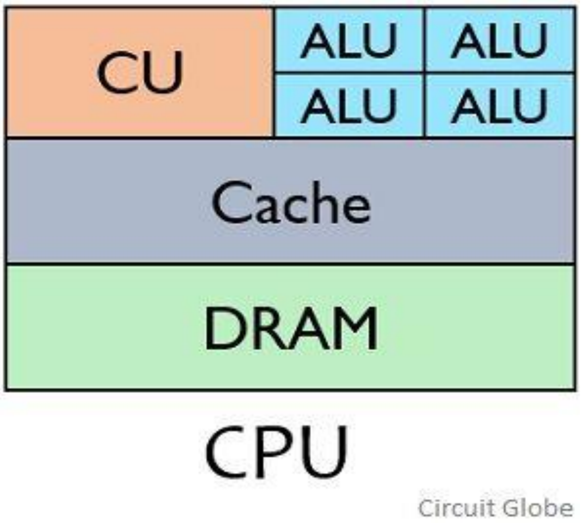
The CPU is suited to a wide variety of workloads, especially those for which latency or per-core performance are important. A powerful execution engine, the CPU focuses its smaller number of cores on individual tasks and on getting things done quickly. This makes it uniquely well equipped for jobs ranging from serial computing to running databases.

GPUs began as specialized ASICS developed to accelerate specific 3D tasks. Over time, these fixed- function engines became more programmable and more flexible. While graphics and the increasingly lifelike visuals of today's top games remain their principal function, GPUs have evolved to become more general- purpose parallel processors as well, handling a growing range of applications.

CPU:

Within every CPU, there are a few standard components, which include the following:
Core(s): The central architecture of the CPU is the "core," where all computation and logic happens. A core typically functions through what is called the "instruction cycle," where instructions are pulled from memory (fetch), decoded into processing language (decode), and executed through the logical gates of the core (execute). Initially, all CPUs were single-core, but with the proliferation of multi-core CPUs, we've seen an increase in processing power.
Cache: Cache is super-fast memory built either within the CPU or in CPU-specific motherboards to facilitate quick access to data the CPU is currently using. Since CPUs work so fast to complete millions of calculations per second, they require ultra-fast (and expensive) memory to do it—memory that is much faster than hard drive storage or even the fastest RAM..
Memory Management Unit (MMU): The MMU controls data movement between the CPU and RAM during the instruction cycle.
CPU Clock and Control Unit: Every CPU works on synchronizing processing tasks through a clock. The CPU clock determines the frequency at which the CPU can generate electrical pulses, its primary way of processing and transmitting data, and how rapidly the CPU can work. So, the higher the CPU clock rate, the faster it will run and quicker processor-intensive tasks can be completed.

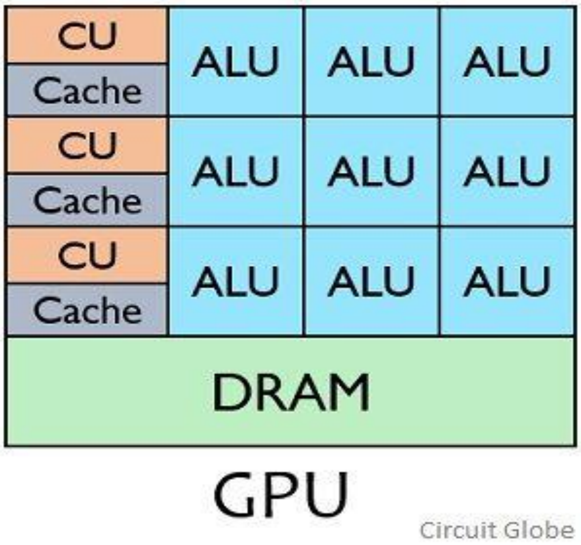
All these components work together to provide an environment where high-speed task parallelism can take place. As the CPU clock drives activities, the CPU cores switch rapidly between hundreds of different tasks per second. That's why your computer can run multiple programs, display a desktop, connect to the internet, and more all at the same time.
The CPU is responsible for all activity on a computer. When you close or open programs, the CPU must send the correct instructions to pull information from the hard drive and run executable code from RAM. When playing a game, the CPU handles processing graphical information to display on the screen. When compiling code, the CPU handles all the computation and mathematics involved.



GPU:

GPUs are similar in function to CPU: they contain cores, memory, and other components. Instead of emphasizing context switching to manage multiple tasks, GPU acceleration emphasizes parallel data processing through a large number of cores.

These cores are usually less powerful individually than the core of a CPU. GPUs also typically have less interoperability with different hardware APIs and houseless memory. Where they shine is pushing large amounts of processed data in parallel. Instead of switching through multiple tasks to process graphics, a GPU simply takes batch instructions and pushes them out at high volume to speed processing and display.



A **CPU** (central processing unit) works together with a **GPU** (graphics processing unit) to increase the throughput of data and the number of concurrent calculations within an application. A CPU can never be fully replaced by a GPU: a GPU complements CPU architecture by allowing repetitive calculations within an application to be run in parallel while the main program continues to run on the CPU. The CPU can be thought of as the taskmaster of the entire system, coordinating a wide range of general-purpose computing tasks, with the GPU performing a narrower range of more specialized tasks. Using the power of parallelism, a GPU can complete more work in the same amount of time as compared to a CPU.

Comparison of CPU and GPU:

CPU	GPU
● It stands for central processing unit.	● It stands for graphical processing unit.
● It is suitable for serial instruction processing.	● It is suitable for parallel instruction processing.
● Consumes or needs more memory.	● Consumes less memory comparatively.
● Contains minute powerful cores.	● Contains weak cores.
● Emphasis on low latency.	● Emphasis on high throughput.

Intel Core i5-8400 Vs Gigabyte RTX 3080 (an example)

Intel Core i5-8400	Gigabyte RTX 3080
● 2.8 GHz Base Clock Speed Boostable upto 4.0GHz.	● 1.26GHz Base clock speed Boostable upto 1.8 GHz.
● Contains 6 cores and 6 threads.	● Contains 8074 CUDA cores.
● Comes with 9 MB Intel smart cache (L3 cache).	● Comes with 128 kb L1 cache and 5 MB L2 cache.
● Memory capacity is 64GB.	● Memory Capacity is 10GB.
● Recommended power supply(TPD): 65W.	● Recommended power supply(TPD): 750W.
● Box Dimensions: 4.6 x 4.2 x 2.8".	● Box Dimensions: 15.8 x 9.3 x 3.7".
● Weight: 0.6 lb.	● Weight: 4.14 lb.
● Supported APIs: OpenGL: 4.5, DirectX: 12.	● Supported APIs: DirectX: 12, OpenGL: 4.6, Vulkan.
● Maximum Display Resolution: 4096 x 2304 at 60 Hz.	● Maximum Display Resolution: 7680 x 4320.
● OS Support: supports all operating systems.	● OS Support: officially supports only windows.

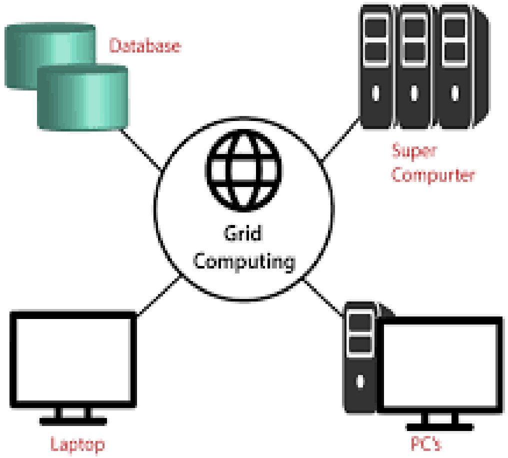
MPP:

Massively Parallel Processing (MPP) is a processing paradigm where hundreds or thousands of processing nodes work on parts of a computational task in parallel. Each of these nodes run individual instances of an operating system. They have their own input and output devices, and do not share memory. They achieve a common computational task by communicating with each other over a high-speed interconnect.
Types of MPP database architecture:

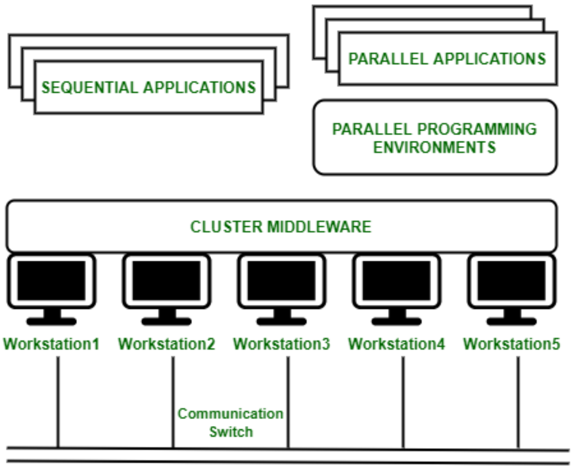
- Grid computing
- Computer clustering

Grid Computing:

Grid computing is the use of widely distributed computer resources to reach a common goal. A computing grid can be thought of as a distributed system with non-interactive workloads that involve many files. Grid computing is distinguished from conventional high-performance computing systems such as cluster computing in that grid computers have each node set to perform a different task/application. Grid computers also tend to be more heterogeneous and geographically dispersed (thus not physically coupled) than cluster computers.



(Grid computing)



(Computer clustering)

Computer Clustering:

A computer cluster is a set of computers that work together so that they can be viewed as a single system. Unlike grid computers, computer clusters have each node set to perform the same task, controlled, and scheduled by software. The components of a cluster are usually connected to each other through fast local area networks, with each node (computer used as a server) running its own instance of an operating system. In most circumstances, all of the nodes use the same hardware and the same operating system, although in some setups, different operating systems can be used on each computer, or different hardware.

Comparison of cluster computing and grid computing:

Cluster computing	Grid computing
● A Computer Cluster is a local network of two or more homogeneous computers. A computation process on such a computer network i.e. cluster is called Cluster Computing.	● Grid Computing can be defined as a network of homogeneous or heterogeneous computers working together over a long distance to perform a task that would rather be difficult for a single machine.
● Components of a Cluster Computer consists of Cluster Nodes, Cluster Operating System, The switch or node interconnect and Network switching hardware.	● Grid computing consists of control node, provider and user.
● Computers are located close to each other.	● Computers may be located at a huge distance from one another.
● Whole system has a centralized resource manager.	● It is not centralized, every node manages it's resources independently.
● Computers in a cluster are dedicated to the same work and perform no other task.	● Computers in a grid contribute their unused processing resources to the grid computing network.
● It is a homogenous network whose devices have the same hardware components and the same OS connected together in a cluster.	● It is a heterogenous network whose devices have different hardware components and the different OS connected together in a grid. Machines can be both homogenous and heterogenous.
● Cluster nodes are located in a single location.	● Cluster nodes are located in a different location.
● Devices are connected through fast local area network.	● Devices are connected through low speed network or internet.
● It is used to solve issues in databases or web logic application servers.	● It is used to solve predictive modeling, simulation, engineering design, automation, etc.