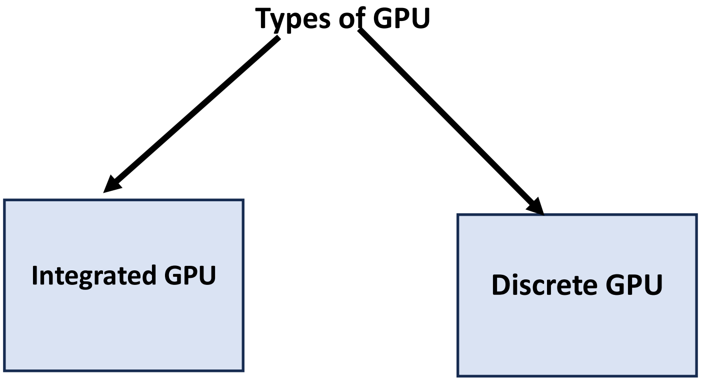
Unit-3

Topic 1

The CPU, GPU system as an accelerated computational platform

A CPU, or Central Processing Unit, is the primary component of a computer that performs most of the processing inside the computer. It interprets instructions from the computer's memory, processes them, and performs arithmetic and logical operations.

A GPU, or Graphics Processing Unit, is a specialized electronic circuit designed to accelerate the processing of images and videos in a computer. Originally developed for rendering graphics in video games and multimedia applications. GPUs consist of thousands of smaller cores that can handle multiple tasks simultaneously. This parallel architecture makes them highly efficient for tasks that can be parallelized.



Integrated GPUs are built into the same chip as the central processing unit (CPU). They share system memory (RAM) with the CPU and are commonly found in laptops, Ultrabook’s, and budget desktop computers. The AMD(Advanced Micro Devices) integrated GPUs are called Accelerated Processing Units (APUs). These are a tightly coupled combination of the CPU and a GPU. In the AMD APU, the CPU and GPU share the same processor memory. Integrated GPUs are suitable for basic tasks like web browsing, office applications, and multimedia playback.

A discrete GPU/Dedicated GPU (Graphics Processing Unit) is a separate graphics card that is installed on a computer's motherboard as an additional hardware component. Unlike integrated GPUs, which are integrated into the same chip as the CPU, discrete GPUs have their own dedicated video memory (VRAM) and are designed to handle graphics-related tasks independently. Discrete GPUs offer significantly higher performance and are suitable for demanding applications such as gaming, video editing, 3D rendering, and professional graphics work.

Communication Between CPU and GPU

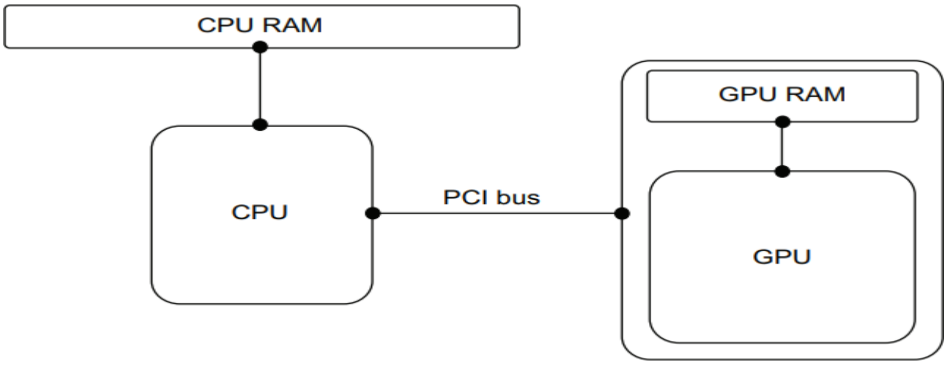


Figure : Block diagram of GPU-accelerated system using a dedicated GPU. The CPU and GPU each have their own memory. The CPU and GPU communicate over a PCI bus.

Components of GPU Accelerated System

CPU—The main processor that is installed in the socket of the motherboard.

CPU RAM—The "memory sticks" or dual in-line memory modules (DIMMs) containing Dynamic Random- Access Memory (DRAM) is a type of computer memory module that is used in desktop computers, servers, and workstations that are inserted into the memory slots in the motherboard.

GPU—A large peripheral card installed in a Peripheral Component Interconnect Express (PCIe) slot on the motherboard.

GPU RAM—Memory modules on the GPU peripheral card for exclusive use of the GPU.

PCI bus—The wiring that connects the peripheral cards to the other components on the motherboard.

Figure : conceptually illustrated a CPU-GPU system with a dedicated GPU. A CPU has access to its own memory space (CPU RAM) and is connected to a GPU via a PCI bus. It is able to send data and instructions over the PCI bus for the GPU to work with. The GPU has its own memory space, separate from the CPU memory space. In order for work to be executed on the GPU, at some point, data must be transferred from the CPU to the GPU. When the work is complete, and the results are going to be written to file, the GPU must send data back to the CPU. The instructions the GPU must execute are also sent from CPU to GPU. Each one of these transactions is mediated by the PCI bus.

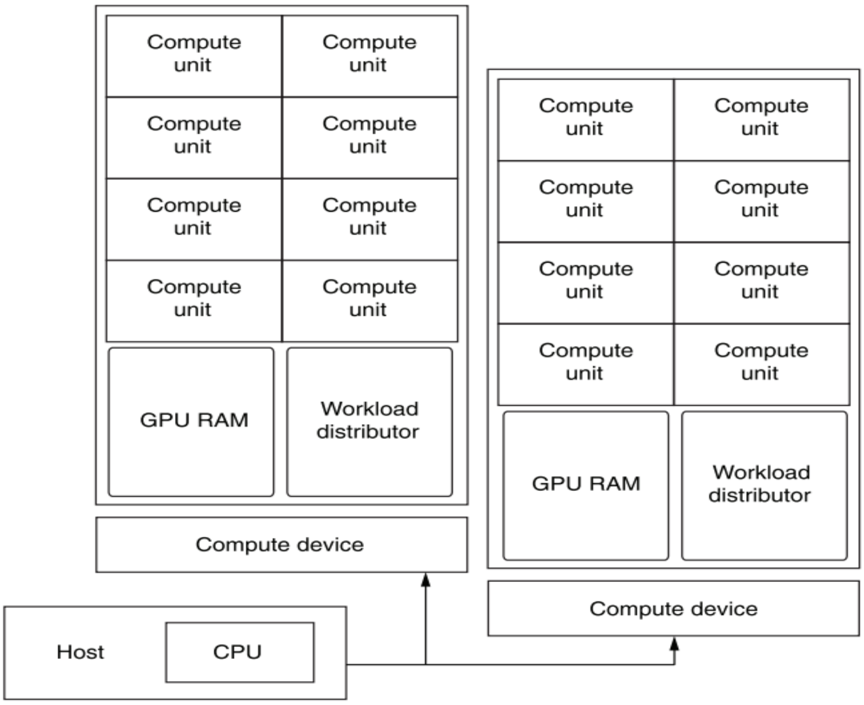
The GPU and the thread engine

The thread engine within a CPU manages the execution of threads, schedules tasks for processing, and ensures efficient utilization of available resources. The graphics processor is like the ideal thread engine.

- **The components of this thread engine are**

- A seemingly infinite number of threads
- Zero-time cost for switching or starting threads: refers to the ideal scenario where the process of initiating or switching between threads occurs instantaneously, without any additional computational overhead.
- Latency hiding of memory accesses through automatic switching between work groups: Memory latency occurs because accessing data from the main memory (RAM) is significantly slower compared to accessing data from the CPU's cache or registers.

- **For Example here we will go through a single node system with a single multiprocessor CPU and two GPUs**



- **Fig 9.2 Simplified Block Diagram of a GPU System consisting of two compute devices each having multiple compute units and separate GPU Memory.**

A GPU is composed of

- **Compute Device :** A compute device in a GPU is a subset of the GPU that is dedicated to general-purpose parallel processing tasks. It consists of multiple compute units
- **GPU RAM** : (also known as global memory) refers to the dedicated memory that is integrated into a graphics processing unit (GPU) or graphics card.
- **Workload distributor**: Instructions and data received from the CPU are processed by the workload distributor. The distributor coordinates instruction execution and data movement onto and off of the Compute Units.
- **Compute units (CU):** Compute units are the fundamental processing units within a compute device. Each compute unit typically consists of multiple ALUs and multiple graphics processors called processing elements (PEs). CUs have their own internal architecture, often referred to as the microarchitecture.

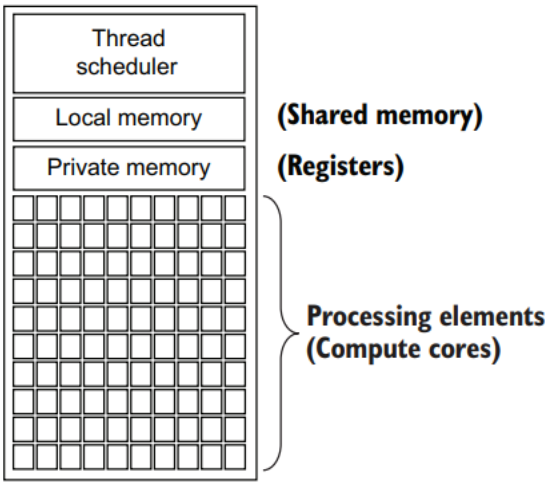


Figure 9.3 Simplified block diagram of a compute unit (CU) with a large number of processing elements (PEs).

- **Processing Element :** PEs within a GPU are designed to execute instructions in parallel. They can handle multiple threads and data elements simultaneously, allowing for massive parallelism.

➢ **Hardware Terminology**

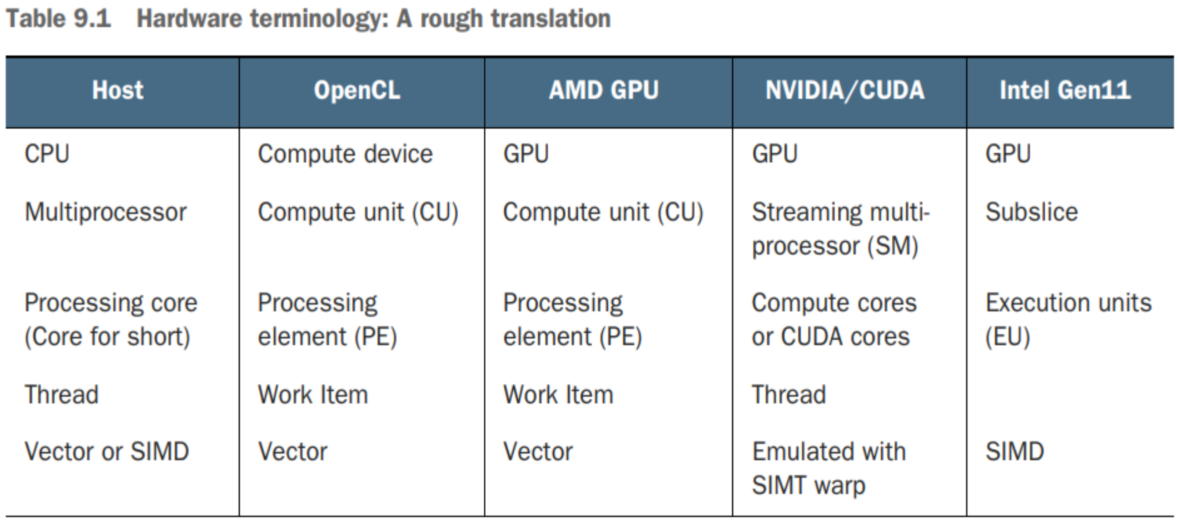


Table 9.1 Hardware terminology: A rough translation

Host	OpenCL	AMD GPU	NVIDIA/CUDA	Intel Gen11
CPU	Compute device	GPU	GPU	GPU
Multiprocessor	Compute unit (CU)	Compute unit (CU)	Streaming multiprocessor (SM)	Subslice
Processing core (Core for short)	Processing element (PE)	Processing element (PE)	Compute cores or CUDA cores	Execution units (EU)
Thread	Work Item	Work Item	Thread	
Vector or SIMD	Vector	Vector	Emulated with SIMT warp	SIMD

Table : summarizes the rough equivalence of terminology, in different hardware architectures. Example CPU in Host is termed as Compute Device (OpenCL), GPU in (AMD GPU) ,GPU (NVIDIA/CUDA) and GPU in Intel Gen11.

- **Calculating the peak theoretical flops for some leading GPUs**

FLOPS provide a measure of a computer system's processing speed, especially when dealing with numerical and scientific computations. It allows researchers and developers to compare the performance of different hardware architectures and configurations.

The peak theoretical flops can be calculated by taking the clock rate times the number of processors times the number of floating-point operations per cycle. The flops per cycle accounts for the fused-multiply add (FMA), which does two operations in one cycle.

Peak Theoretical Flops (GFlops/s) = Clock rate MHZ × Compute Units × Processing units × Flops/cycle.