Subject Name: High Performance Computing

Module Number: 03

Module Name: Graphics Processing Units

Graphics Processing Units

Syllabus

Graphics Processing Units

Graphics Processing Units : Introduction to Heterogeneous Parallel Computing. GPU architecture. Thread hierarchy. GPU Memory Hierarchy.

GPU Programming: Vector Addition, Matrix Multiplication algorithms. 1D, 2D, and 3D Stencil Operations. Image Processing algorithms, Image Blur, Gray scaling. Histogram Ming, Convolution, Scan, Reduction techniques.

Graphics Processing Units

Aim

Graphics processing unit, a specialized processor originally designed to accelerate graphics rendering. GPUs can process many pieces of data simultaneously, making them useful for machine learning, video editing, and gaming applications.

Graphics Processing Units

Objectives

The Objectives of this module are

- Identify different GPU architecture.
- Learn about GPU programming.
- Learn about 2D and 3D image processing
- Identify Image Processing technique.

Graphics Processing Units

Outcome

At the end of this module, you are expected to:

- Recognize important GPU architecture and understand working of GPU .
- Understand 2D and 3D image Processing
- Understand GPU memory hierachy.
- Recognize various image processing method

Graphics Processing Units

Contents

Introduction to Heterogeneous Parallel Computing.

GPU architecture. Thread hierarchy

GPU Memory Hierarchy.

GPU Programming: Vector Addition, Matrix Multiplication algorithms.

1D, 2D, and 3D Stencil Operations.

Image Processing algorithms – Image Blur, Gray scaling.

Histogram Ming,

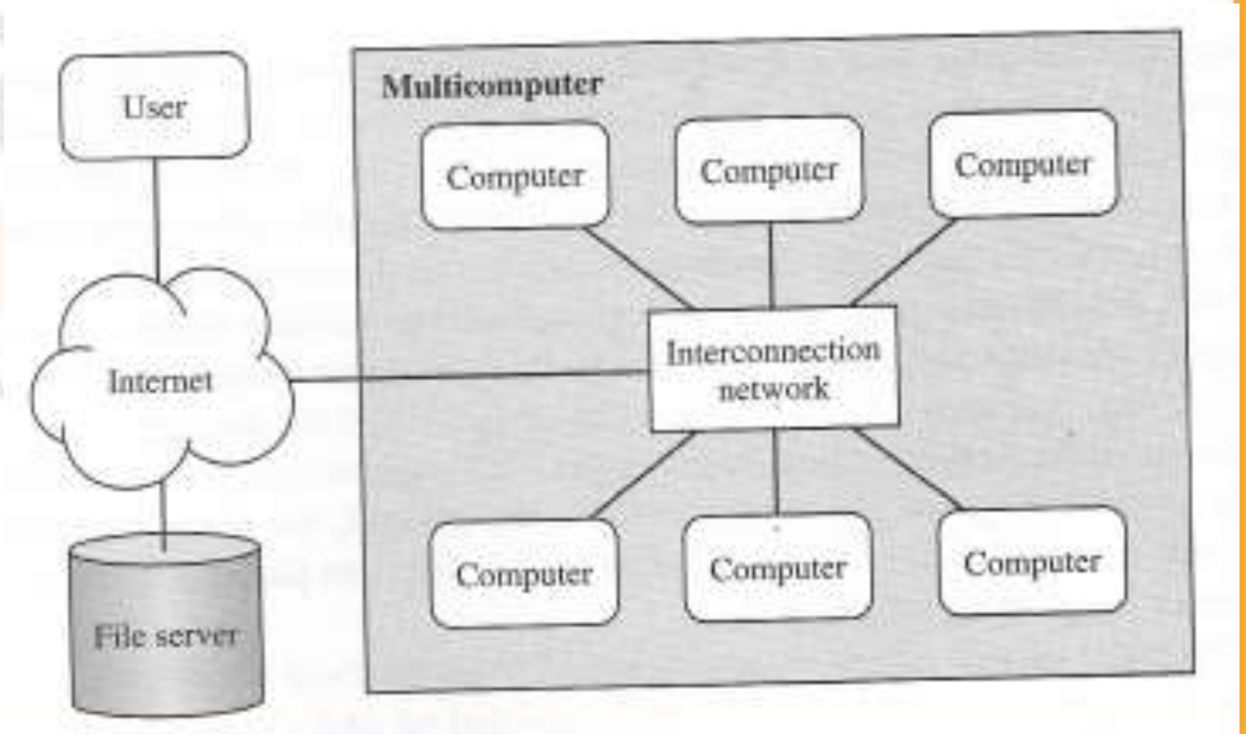
Convolution, Scan,

Reduction techniques.

Introduction to Heterogeneous Parallel Computing:

Concept: Heterogeneous parallel computing involves utilizing different types of processors or computing units to handle various tasks simultaneously. It aims to improve overall computational efficiency by assigning specific tasks to the most suitable processing units.

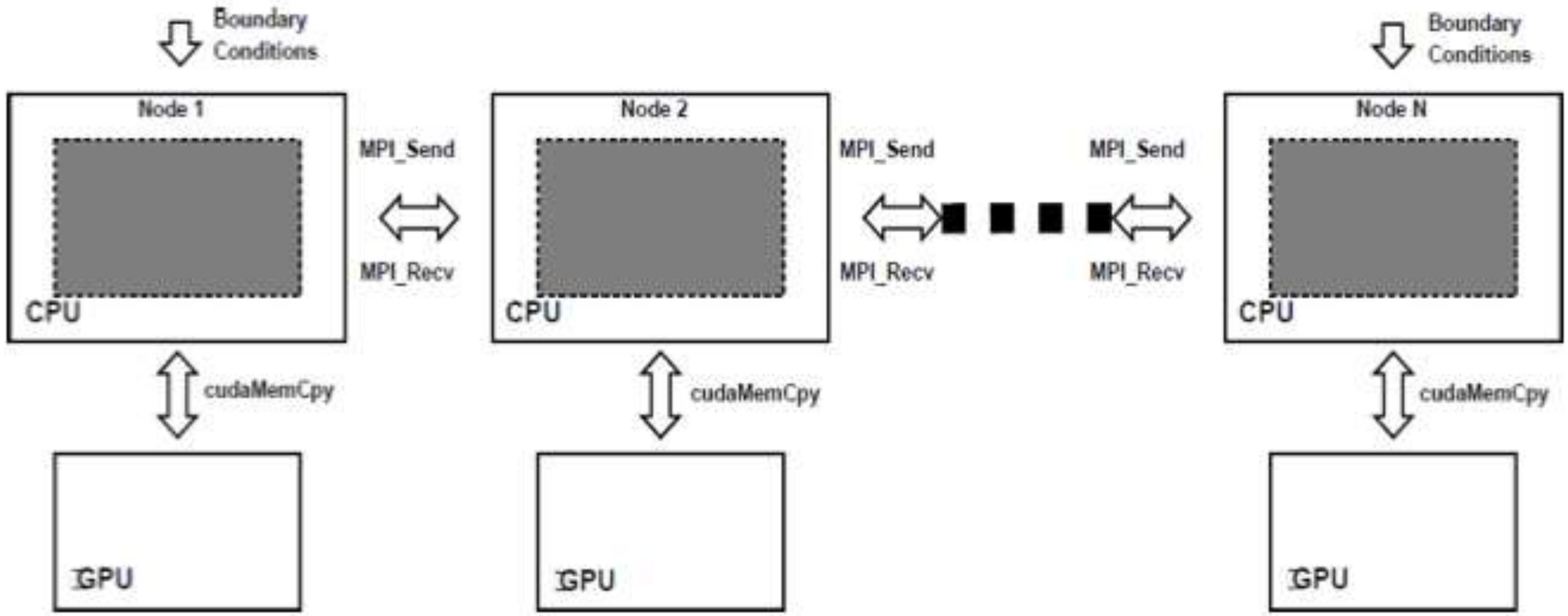
Types:
•CPUs (Central Processing Units)
•GPUs (Graphics Processing Units)
•TPUs (Tensor Processing Units)
•FPGAs (Field-Programmable Gate Arrays)



Picture taken from Quinn's book.

Graphics Processing Units

Graphics Processing Units



Graphics Processing Units

Components Heterogeneous Parallel computing

1. Processors (CPUs, GPUs, etc.)

- CPUs (Central Processing Units) are general-purpose processors handling a wide range of tasks in a computer system.

- GPUs (Graphics Processing Units) are specialized processors designed for parallel processing, primarily used in graphics rendering and parallelizable computing tasks.

- TPUs (Tensor Processing Units) are specialized accelerators optimized for machine learning tasks, particularly neural networks.

  Multi-Layer Perceptrons (MLP)
  Convolutional Neural Networks (CNN)
  Recurrent Neural Networks (RNN)

2. Memory

Memory refers to the storage space used by a computer to store data temporarily for processing.

Working Concept:

- RAM (Random Access Memory) provides fast, volatile storage for actively used data and programs.
- Cache memory is a smaller, faster type of memory used to store frequently accessed data for quicker retrieval by the processor.
- Storage devices like hard drives and SSDs offer non-volatile, long-term storage for data and programs.

Graphics Processing Units

3. Interconnects (for communication between processors):

❖ **Interconnects (for communication between processors):**

Interconnects are communication pathways that enable data transfer between different components in a computer system.

Working Concept:

➢ Buses, such as the system bus, connect various hardware components to the CPU, facilitating data exchange.

➢ High-speed interconnects, like PCIe (Peripheral Component Interconnect Express), enable communication between the CPU and peripheral devices.

➢ In multi-processor systems, interconnects (e.g., HyperTransport, NVLink) allow processors to share data and work collaboratively.

Graphics Processing Units

Advantages and Disadvantages Heterogeneous Parallel computing

- **Advantages**

Efficient utilization of specialized processors.
Improved performance for parallelizable tasks.
Flexibility in handling diverse workloads.

- **Disadvantages:**

Complex programming models.
Potential increased power consumption.

Graphics Processing Units

GPU Architecture:

Types of GPU

1. Integrated Graphics
2. Discrete Graphics
3. Mobile GPUs
4. Workstation and Server GPUs

Graphics Processing Units

Components of GPU Architecture

1. Streaming Multiprocessors (SMs):-
2. Processing Cores (CUDA cores or Stream Processors)
3. Thread Hierarchy
4. Memory Hierarchy

Graphics Processing Units

Advantages and Disadvantages:

Advantages:

- High parallel processing capabilities.
- Specialized units for graphics rendering.
- Suitable for scientific computing, AI, and machine learning.

Disadvantages:

- Complex programming models.
- High power consumption for high-performance GPUs.

Graphics Processing Units

Streaming Multiprocessors (SMs):-

A stream multiprocessor is a type of SIMD or a MIMD machine where the constituent processors are streaming processors (SPs) or thread processors. A stream processor is defined as a processor that deals with data streams, and its instruction set architecture (ISA) contains kernels to process these streams

- How many threads does a streaming multiprocessor have?

In general an SM can handle multiple thread blocks at the same time. An SM may contain up to 8 thread blocks in total. A thread ID is assigned to a thread by its respective SM.

Graphics Processing Units

Working of Streaming Multiprocessors (SMs):

1.Handling Multiple Thread Blocks:

1. In NVIDIA's GPU architecture, each SM can handle multiple thread blocks concurrently.
2. The ability to process multiple thread blocks simultaneously is a key feature of parallel execution on GPUs, enhancing throughput and efficiency.

2.Maximum Number of Thread Blocks:

1. The maximum number of thread blocks an SM can handle concurrently depends on the specific GPU architecture.
2. As of my last knowledge update in January 2022, the maximum number of active thread blocks per SM is typically in the range of 16 to 32.

3.Thread ID Assignment:

1. Each thread within a thread block is assigned a unique thread ID by its respective SM.
2. Thread IDs are crucial for identifying and coordinating individual threads during parallel execution.

4.Concurrency:

1. Multiple thread blocks within an SM can execute concurrently, which is essential for maximizing the parallel processing capabilities of the GPU.

Graphics Processing Units

Working of Streaming Multiprocessors (SMs):

- SMs are the heart of a GPU.
- Each GPU contains multiple SMs, which are individual processing units responsible for executing tasks in parallel.
- SMs consist of ALUs (Arithmetic Logic Units) and control units, allowing them to perform mathematical operations simultaneously

Graphics Processing Units

Graphics Processing Units

Example:

For instance, in NVIDIA's CUDA architecture, an SM might be able to handle, for example, 32 concurrent thread blocks, each comprising a certain number of threads. The exact numbers can vary based on the GPU model and architecture.

Advantages:

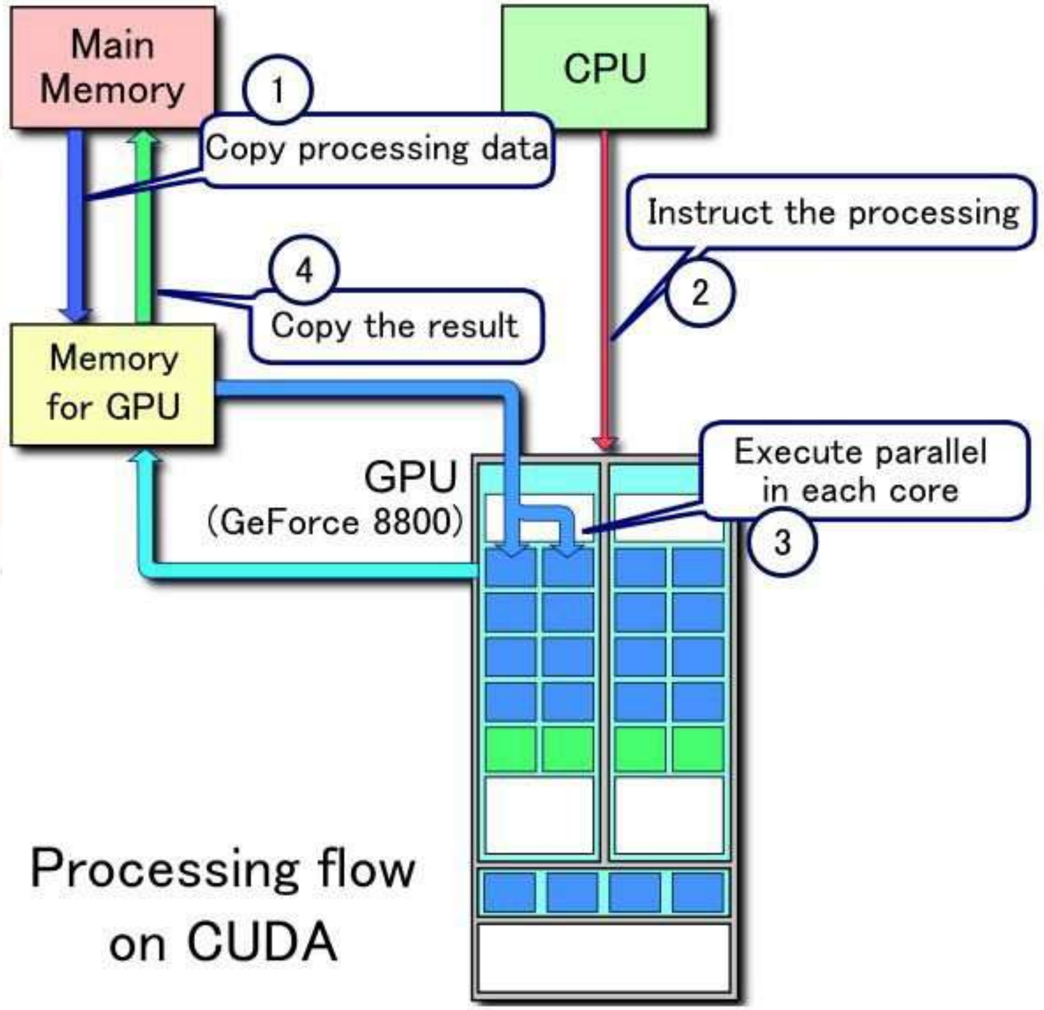
- Increased parallelism: Multiple thread blocks can be processed simultaneously, enhancing overall throughput.
- Efficient resource utilization: Allows the GPU to efficiently manage and execute diverse workloads concurrently.

Disadvantages:

- Increased complexity: Coordinating the execution of multiple thread blocks requires careful programming to avoid race conditions and ensure proper synchronization.
- Resource constraints: The total number of thread blocks that can be accommodated is limited by the specific GPU architecture.

Graphics Processing Units

Processing Cores (CUDA cores or Stream Processors)



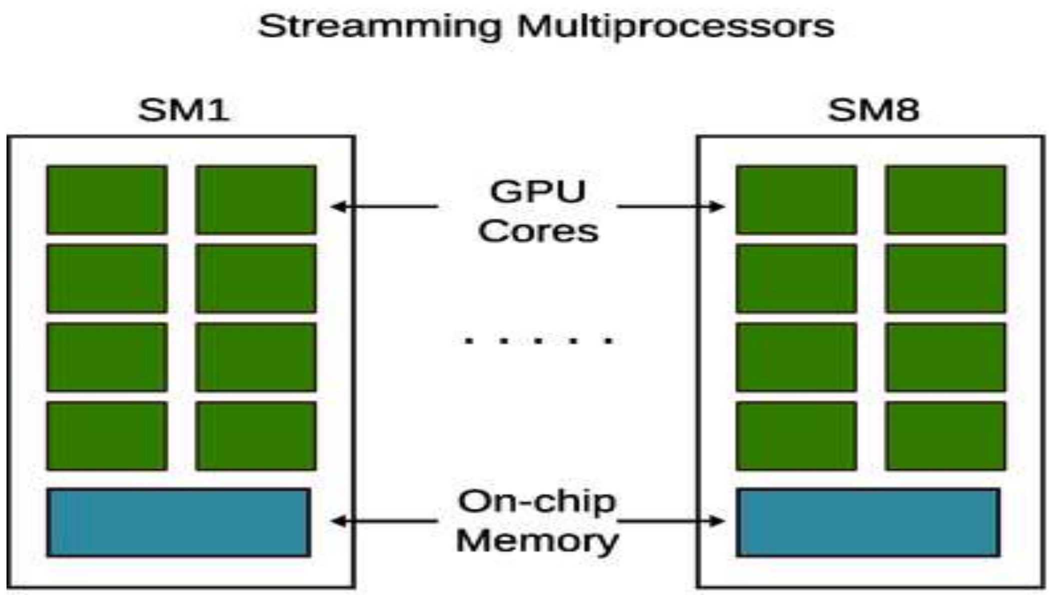
Graphics Processing Units

What are CUDA processing cores?

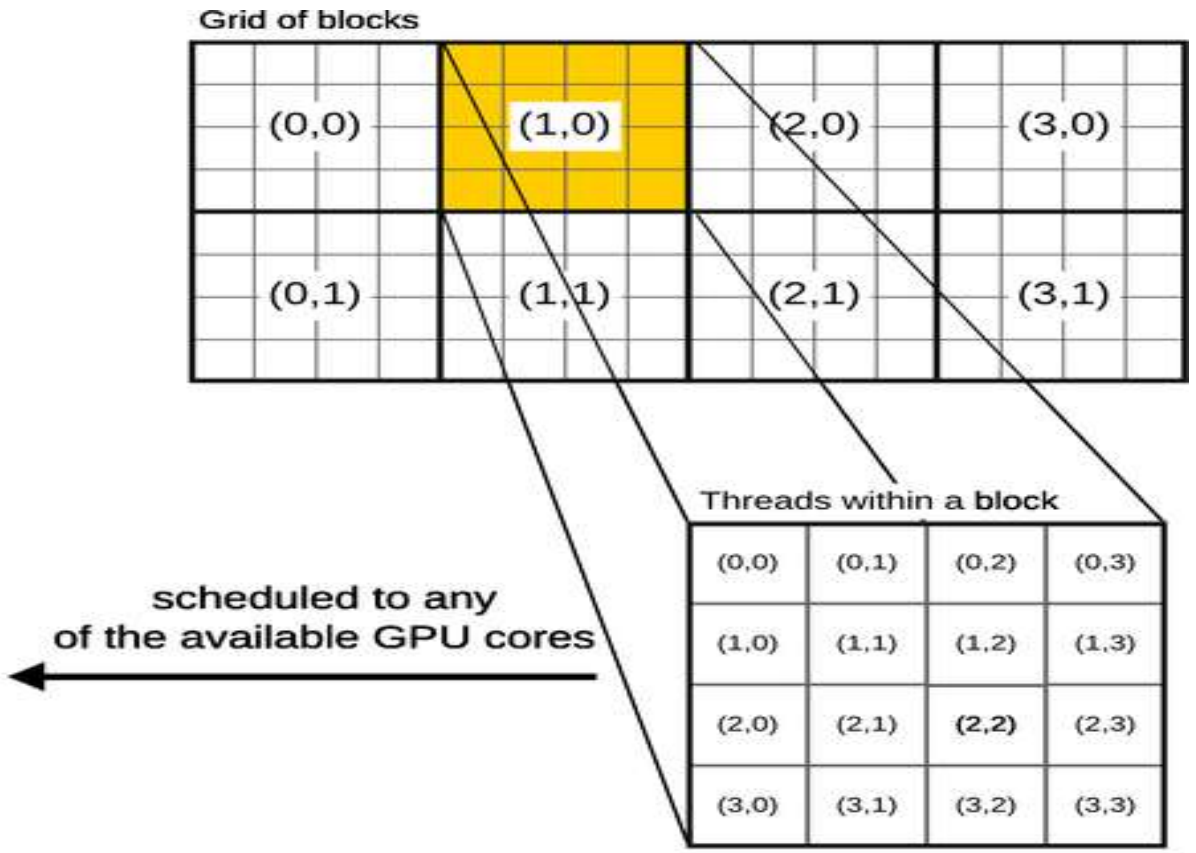
CUDA cores are parallel processors in Nvidia's GPUs that perform computations. They are designed to handle multiple tasks simultaneously, making them ideal for applications that require high computational power, such as gaming, machine learning, and AI

Graphics Processing Units

CUDA Core Working Flow



(a)



(b)

Graphics Processing Units

Graphics Processing Units

Definition:

CUDA cores, also known as streaming processors or CUDA processing units, are the individual processing units within a GPU that perform parallel computations. These cores are responsible for executing parallel threads concurrently, enabling the GPU to process massive amounts of data in parallel.

OR

CUDA cores can be defined as a parallel computing platform and application programming interface (API) model created by Nvidia, to handle multiple tasks simultaneously, making them highly efficient for jobs that can be broken down into parallel processes

Graphics Processing Units

Functionality: and Architecture:

Architecture:

•**Parallel Processing Units:**
- CUDA cores are organized into parallel processing units, each capable of executing a thread independently.

•**SIMT Architecture:**
- NVIDIA GPUs use the Single Instruction, Multiple Threads (SIMT) architecture, where each CUDA core processes multiple threads simultaneously.

Functionality:

•**Parallel Execution:**
- CUDA cores excel at parallel execution, allowing them to perform multiple calculations simultaneously

Task Division:

•CUDA cores divide the workload into smaller tasks that can be executed concurrently, optimizing throughput.

Graphics Processing Units

CUDA Cores vs. CPU Cores:

Specialization:

- While CPU cores are designed for general-purpose computing, CUDA cores are specialized for parallel computing tasks, particularly in graphics and scientific computing.

High Throughput:

- CUDA cores are optimized for high-throughput parallel processing, making GPUs suitable for tasks that involve processing large datasets concurrently.

Graphics Processing Units

Advantage and Disadvantage of CUDA Core

Advantages:

- **Parallel Processing Power:**
  - CUDA cores provide significant parallel processing power, making GPUs suitable for tasks like graphics rendering, scientific simulations, and machine learning.
- **Optimized for Graphics:**
  - Originally designed for graphics rendering, CUDA cores are well-suited for parallelizable tasks in graphics applications.

Disadvantages:

- **Specialized Use Cases:**
  - CUDA cores are optimized for specific parallel workloads, and their efficiency may vary for general-purpose computing tasks.
- **Programming Complexity:**
  - Writing efficient CUDA code requires an understanding of parallel programming concepts, which can be complex.

Graphics Processing Units

Advantage and Disadvantage of CUDA Core

Example:

Suppose you are performing a matrix multiplication using CUDA. Each CUDA core could be responsible for calculating a specific element in the resulting matrix. As the matrix is split into smaller tasks, CUDA cores handle these tasks concurrently, speeding up the overall computation compared to a sequential approach.

Graphics Processing Units

Assignment

1. Explain Heterogeneous Parallel Computing with diagram.
2. Describe GPU architecture
3. Define GPU Memory Hierarchy.
4. Why Image Processing algorithms is requred?
5. Explain Image Blur in image processing .
6. Define Convolution .
7. Explain Reduction techniques .

Graphics Processing Units

Document Links

Topic	URL	Notes
Complete Introduction to GPU Programming With Practical Examples in CUDA and Python	https://www.cherryservers.com/blog/introduction-to-gpu-programming-with-cuda-and-python#:~:text=GPU%20Programming%20is%20a%20method,(GPGPU%20computing)%20as%20well.	**this link explain Complete Introduction to GPU Programming With Practical Examples in CUDA and Python**
vector Add with CUDA	http://selkie.macalester.edu/csinparallel/modules/ConceptDataDecomposition/build/html/Decomposition/CUDA_VecAdd.html	This link explains about cuda vector code

Graphics Processing Units

Video Links

Topic	URL	Notes
Graphics Processing Units	https://www.youtube.com/watch?v=wFlejBXX9Gk&list=PL3xCBlatwrsXCGW4SfEoLzKiMSUCE7S_X	Explains about GPU PROGRAMMING
CUDA	https://www.youtube.com/watch?v=2NgpYFdsduY&list=PLxNPSjHT5qvtYRVdNN1yDcdSl39uHV_sU	This video talks about CUDA architure

Graphics Processing Units

E-Book Links

Topic	URL	Notes
Handbook of High-Performance Computing	https://www.routledge.com/rsc/downloads/High_Performance_Computing_ChapterSampler.pdf	Ebook gives an comprehensive knowledge on High-Performance Computing