Computer Architecture (Assignment - 02 (Part-2))

Sreechand R (22562) , Sindhura S (22706)

November 26, 2023

Introduction to GPU

A Graphics Processing Unit (GPU) is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device.

Key Characteristics of GPUs

Parallel Structure

GPUs are particularly efficient at handling parallel processing tasks. This makes them ideal for complex computations in computer graphics and also increasingly popular for other data-intensive tasks.

Application in General-Purpose Computing

While originally designed for computer graphics, GPUs are now also used for general-purpose computing. This practice is known as GPGPU (General-Purpose computing on Graphics Processing Units).

Use in Deep Learning and AI

GPUs are extensively used in deep learning and artificial intelligence for their ability to perform a large number of floating-point calculations quickly, accelerating neural network training and inference.

Difference from CPUs

GPUs differ from Central Processing Units (CPUs) in their architecture. CPUs are designed to handle a few software threads at a high clock speed, while GPUs are designed to handle thousands of threads simultaneously at a lower clock speed.

Architecture

CPU

- Designed for sequential processing.
- Excels in tasks requiring complex decision-making and control flow.
- Typically has fewer cores, but with higher clock speeds.
- Optimized for latency-sensitive tasks.

GPU

- Engineered for parallel processing.
- Ideal for computations that can be run concurrently.
- Contains a large number of cores suited for handling multiple tasks simultaneously.
- Optimized for throughput-intensive tasks.

Use Cases

CPU

- General-purpose tasks such as running operating systems and applications.
- Tasks that require immediate responses and complex calculations.

GPU

- Graphics rendering in video games and 3D applications.
- Data-parallel tasks common in scientific computing and machine learning.

Performance Considerations

- **CPU**: More suitable for tasks with a complex mix of operations and where single-thread performance is critical.
- **GPU**: More efficient for tasks that can be broken down into smaller, similar operations to be executed simultaneously.

Introduction to CUDA

CUDA (Compute Unified Device Architecture) is a parallel computing platform and application programming interface (API) model created by NVIDIA. It allows software developers to use a CUDA-enabled graphics processing unit (GPU) for general purpose processing (an approach known as GPGPU, General-Purpose computing on Graphics Processing Units).

Key Features of CUDA

Parallel Programming Model

CUDA extends the C programming language with a minimal set of extensions to express parallelism. It is designed to work with programming languages such as C, C++, and Fortran.

Memory Hierarchy

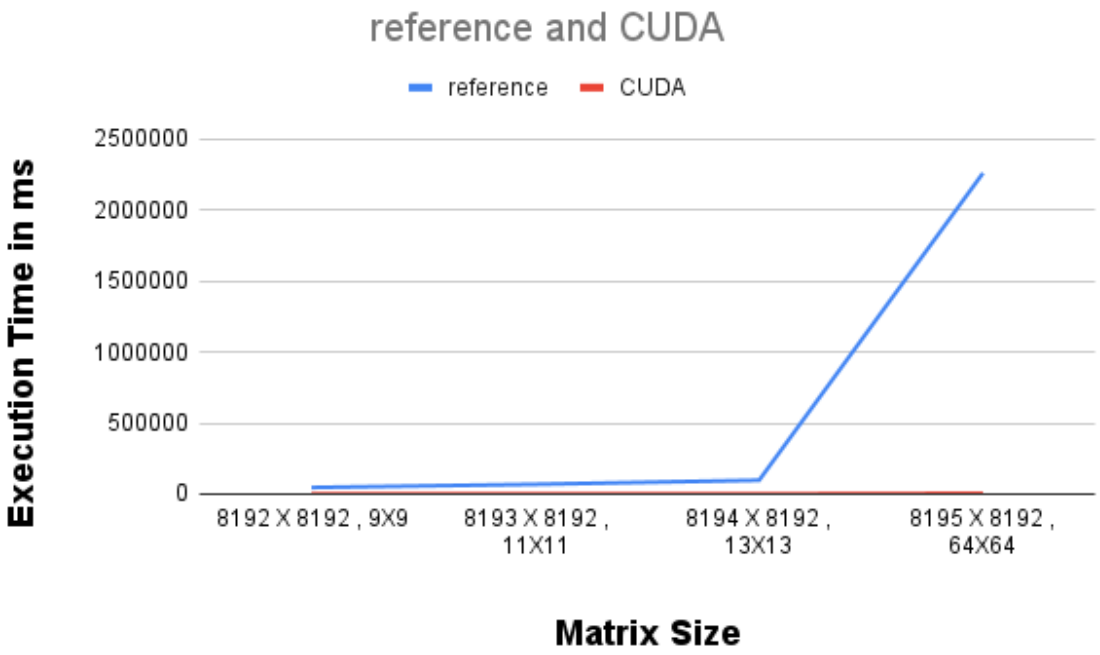
CUDA includes a hierarchical memory model that allows for efficient utilization of the GPU's memory resources. This includes various types of memory such as global, local, shared, and constant memory, each serving different purposes.

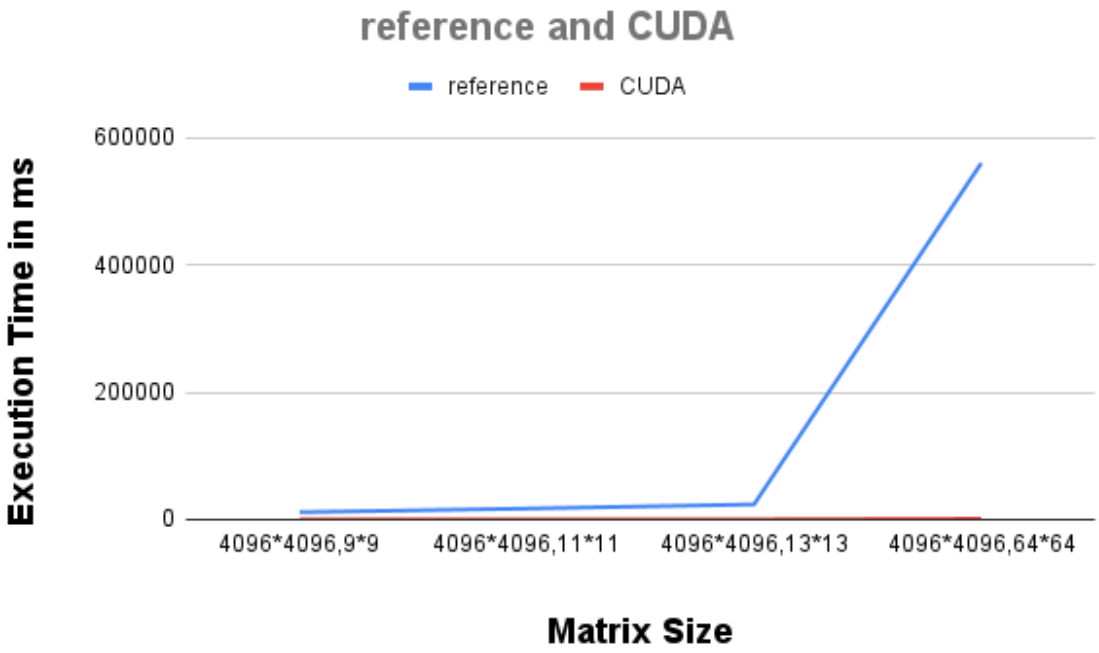
Scalability

The CUDA architecture is scalable and can efficiently handle increases in the number of processor cores. This makes it suitable for a wide range of GPU architectures.

Applications of CUDA

- **High-Performance Computing**: CUDA is widely used in scientific computing and high-performance computing applications.
- **Machine Learning and Deep Learning**: CUDA accelerates neural network training and inference, making it a cornerstone in the field of AI.
- **Computer Graphics**: Used in 3D rendering and other graphics-intensive applications.
- **Data Science**: Employed in data-intensive tasks for faster processing and analysis.





Conclusion

- speed up = reference / CUDA execution
- The speed up for the following matrix in CUDA execution are
  4096*4096, 9*9 = 42.30
  4096*4096, 11*11 = 57.07
  4096*4096, 13*13 = 70.79
  4096*4096, 64*64 = 431.90
  8192*8129, 9*9 = 57.58
  8192*8192, 11*11 = 83.32
  8192*8192, 13*13 = 103.67
  8192*8192, 64*64 = 499.624.
- The average speed up = 168.28