**Writing Your First CUDA Program**

**What is CUDA?**  
CUDA (Compute Unified Device Architecture) is NVIDIA’s platform for parallel programming. It allows developers to use the GPU for general-purpose processing.

**Objective:**  
Write a basic CUDA program where a kernel runs on the GPU and prints a message from each thread.

**Step 1: Create the CUDA Program File**  
Save the following as hello.cu:

#include <stdio.h>

// This is the kernel function that runs on the GPU

\_\_global\_\_ void helloFromGPU() {

printf("Hello from GPU thread %d\n", threadIdx.x);

}

// Main function runs on the CPU

int main() {

// Launch the kernel with 5 threads in 1 block

helloFromGPU<<<1, 5>>>();

// Wait for GPU to finish before exiting

cudaDeviceSynchronize();

return 0;

}

**Step 2: Compile the Program**  
Open a terminal and run:

nvcc hello.cu -o hello

**Step 3: Run the Program**

./hello

**Expected Output:**

Hello from GPU thread 0

Hello from GPU thread 1

Hello from GPU thread 2

Hello from GPU thread 3

Hello from GPU thread 4

**Explanation:**

* \_\_global\_\_ defines a kernel function that runs on the GPU.
* helloFromGPU<<<1, 5>>>(); launches one block with 5 threads.
* threadIdx.x gives the index of the thread within the block.
* cudaDeviceSynchronize() ensures the GPU finishes execution before the program exits.