Programa de contenidos

Programación de GPUs con CUDA
(18 horas)

Profesor:

Manuel Ujaldón Martínez (ujaldon@uma.es). Catedrático de Arquitectura de Computadores en la UMA y DLI Ambassador en Nvidia Corporation.

Distribución horaria:

- Miércoles, 29 de Junio
  - Clase 1: Arquitectura de la GPU e innovaciones para computación de altas prestaciones. [90']
  - Clase 2: Aceleración con GPUs dotadas de miles de cores. Las 8 generaciones de GPUs: Evolución desde Tesla a Ampere. [75']
- Jueves, 30 de Junio
  - Clase 1: Escribir, compilar y ejecutar código en la GPU. Sincronización con la CPU. [90']
  - Clase 2: Jerarquía de hilos: Mallas, bloques y warps. [75']
- Viernes, 1 de Julio
  - Clase 1: Paralelización de aplicaciones en CUDA. Aplicación del paralelismo masivo de datos. [90']
  - Clase 2: Primeros ejemplos prácticos: Suma de vectores, producto de matrices y simulación de conducción térmica. [75']
- Miércoles, 6 de Julio
  - Clase 1: Conociendo el profiler de la línea de comandos (nvprof) y los multiprocesadores de CUDA. [90']
  - Clase 2: Alojamiento de memoria en CUDA y memoria unificada. Optimizaciones. [75']
- Jueves, 7 de Julio

    - Clase 1: Manejo del Visual Profiler (nvvp). Optimización de las transferencias de datos entre CPU y GPU. [90']
    - Clase 2: Uso de flujos concurrentes. CUDA Occupancy Calculator [75']

- Viernes, 8 de Julio
    - Clase 1: Casos estudio: Operadores de reducción, filtros patrón (stencils) e inversión de los elementos de un vector. [90']
    - Clase 2: Evaluación final. Realización del proyecto para lograr la certificación del DLI. [75']

Objetivos:

Aprender las principales técnicas y herramientas para acelerar aplicaciones escritas en lenguaje C de forma que se ejecuten sobre millones de hilos en la GPU utilizando CUDA.
A la conclusión del curso, el alumno adquirirá las siguientes competencias:

- Escribir código para que sea ejecutado en la GPU.
- Exponer y expresar paralelismo de datos con CUDA en aplicaciones escritas en C.
- Analizar cuándo merece la pena portar un código a la GPU.
- Gestionar la memoria en la GPU y optimizar la transferencia de datos utilizando prebúsqueda asíncrona.
- Utilizar los *profilers* para optimizar las aplicaciones paralelas.
- Emplear *streams* para combinar paralelismo de tareas y paralelismo de datos.

A la conclusión del curso, el alumno tendrá la oportunidad de evaluarse dentro del DLI para lograr la certificación en computación acelerada con CUDA avalada por la empresa Nvidia, lo que acreditará sus conocimientos para afianzar su carrera profesional y méritos curriculares en este área.

Tecnologías:

Utilizaremos `nvcc`, `nvprof`, `nvpp`, GPUs de última generación de Nvidia, Jupyter notebooks, Amazon Web Services (AWS). Todos estos recursos se encuentran ya instalados, configurados y accesibles a través de la plataforma del DLI de Nvidia, que constituye el eje central de desarrollo para las tareas planificadas a lo largo de este curso.

Prerrequisitos:

Estar familiarizado con los fundamentos de la programación básica, como manejo de arrays, uso de bucles y llamadas a procedimientos. Se utilizarán *Jupyter notebooks* para la resolución en

tiempo real de ejercicios dentro del DLI de Nvidia, cuya plataforma y recursos de computación en la nube se describirán en una breve introducción del módulo. El DLI proporcionará al alumno toda la infraestructura necesaria para completar su formación a través de una sesión personalizada e interactiva que establecerá desde su propio navegador Web. En este sentido, para seguir el curso sólo es necesario disponer de un computador conectado a Internet.