Programación Paralela

CUDA y paralelización de algoritmos

Segundo Semestre, 2025

Fundamentación

Las CPUs han sido desarrolladas para minimizar la latencia y maximizar el rendimiento de las aplicaciones secuenciales. Sin embargo hay determinadas aplicaciones y algoritmos que sólo pueden ser optimizados si se corren de manera paralela. Para ello se requieren tantos CPUs como GPUs, pero se necesitan lenguajes de programación y técnicas de programación diferentes que las que se utilizan en las CPUs.

El objetivo de la materia Programación Paralela es introducir al alumno en el mundo de la programación paralela, basada en arquitectura de GPUs. Se busca que el alumno adquiera los conocimientos necesarios para poder programar en CUDA, y entender la forma en la que se deben abordar los problemas de programación paralela.

Objetivos

- Introducir al alumno a la programación paralela
- Comprender la arquitectura de las GPUs
- Aprender a programar en CUDA
- Entender cómo se deben abordar los problemas de programación paralela
- Conocer las técnicas de programación paralela
- Aprender a optimizar algoritmos
- Conocer las herramientas de programación paralela
- Realizar un proyecto de programación paralela

0.1 Módulo 1: Introducción a la Programación Paralela

Límites del Paralelismo: Complejidad polinomial, Clases P, Clases de Nick. **Introducción a la Programación Paralela** Evolución de los microprocesadores y la escalabilidad vertical. Importancia del paralelismo masivo y la elección de CUDA C. **Comparación entre CPUs y GPUs** Diferencias en optimización y rendimiento entre CPUs y GPUs. Casos de uso y limitaciones de cada tipo de procesador. **Modelo de Programación CUDA** Descripción del modelo de programación desarrollado por NVIDIA. Ventajas económicas y de rendimiento de utilizar GPUs con CUDA. **Paralelización de Aplicaciones** Beneficios y limitaciones de la paralelización. Ley de Amdahl y su impacto en el rendimiento de programas paralelos. **Desafíos en la Programación Paralela** Complejidad en el diseño de algoritmos paralelos. Sensibilidad a los datos de entrada y límites de acceso a memoria.