



CURSO DE POSTGRADO: COMPUTACIÓN DE PROPÓSITO GENERAL EN UNIDADES DE PROCESAMIENTO GRÁFICO (GPGPU)

Responsable Académicos

Dr. Ing. Pablo Ezzatti (Facultad de Ingeniería, Universidad de la República. Uruguay)

Equipo Docente

Dr. Ing. Pablo Ezzatti

Dr. Ing. Martín Pedemonte (Facultad de Ingeniería, Universidad de la República. Uruguay)

Dr. Ing. Ernesto Dufrechou (Facultad de Ingeniería, Universidad de la República. Uruguay)

Destinatarios

El curso está dirigido a ingenieros civiles, en computación, en bioinformática, bioingenieros, licenciados en sistemas, etc. Cualquier investigador con interés en optimizar el rendimiento de sus soluciones algorítmicas.

Alcance Profesional

Es un curso destinado a introducir los procesadores gráficos como una forma de paralelizar trabajos que requieran un alto rendimiento computacional. En esta primera experiencia con la programación paralela mediante GPU, los estudiantes adquirirán la destreza necesaria para acelerar todo tipo de procesos paralelos en campos científicos muy variados como ingeniería, física y química, entre otros.

Fundamentación

En los últimos años el área de trabajo de HPC ha cambiado de manera importante. Además de los grandes supercomputadores, se ha desarrollado de manera notable el uso de plataformas paralelas altamente potentes, pero de menor costo económico y energético. Entre éstas, se destacan aceleradores de hardware como los procesadores gráficos (GPUs, por su sigla en inglés), que ofrecen capacidades de cómputo impresionantes en comparación con los procesadores tradicionales. A modo de ejemplo, la GPU NVIDIA Tesla P100 dispone del orden de 3500 cores y presenta un impactante pico de desempeño teórico en aritmética de punto flotante de doble precisión de 4.7 teraFLOPS.



La importancia creciente del uso de GPUs para el cálculo científico fundamenta la importancia de avanzar en la formación académica de esta nueva tecnología.

Justificación

Como se mencionó anteriormente, el tema del uso de GPUs se ha difundido en la comunidad de forma vertiginosa. Teniendo en cuenta, además, la importancia del desarrollo de herramientas numéricas eficientes. Es clara la importancia de la temática.

Objetivos generales y específicos

Objetivo general: El objetivo general del curso, es introducir al estudiante en el uso de los procesadores gráficos para la resolución de problemas de propósito general.

Objetivos específicos: Los objetivos específicos apuntan al desarrollo de habilidades para la aplicación de nuevas estrategias de cómputo basadas en plataformas de hardware heterogéneas que incluyen tarjetas gráficas:

- Desarrollar programas masivamente paralelos capaces de ejecutar en GPUs NVIDIA.
- Paralelizar programas secuenciales utilizando GPUs.
- Aplicar distintos conceptos de computación paralela sobre GPUs.
- Explotar de forma eficiente las características de los procesadores gráficos.
- Evaluar y optimizar el desempeño de los programas utilizando las herramientas adecuadas.

Metodología de Trabajo

La estrategia de enseñanza se basará en clases teórico-prácticas expositivas con auxilio de medios visuales. Se complementará con prácticas en computadora de resolución de ejercicios prácticos para favorecer el análisis individual y grupal de problemas. Se realizarán también consultas del laboratorio final.



Criterios y procedimientos de evaluación

La promoción la obtienen los cursantes que, habiendo asistido con regularidad a las clases (mínimo 80% de asistencia), aprueben la evaluación final prevista mediante la realización de un trabajo tipo proyecto.

Duración

Se contempla una carga horaria total de treinta (30) horas; que incluyen clases teóricas-expositivas, prácticas en computadora y consultas de laboratorio.

Infraestructura y equipamiento

Se utilizará un aula con medios audiovisuales para el dictado de las clases teóricas. Respecto a las clases prácticas, se utilizará un salón con acceso a internet. El trabajo práctico se realizará mediante la conexión a un servidor.



UNIDADES TEMÁTICAS Y CONTENIDOS MÍNIMOS

Unidad I: Evolución histórica de las tarjetas gráficas

Breve repaso de la evolución histórica de las tarjetas gráficas. En especial diferentes la evolución en las características relacionadas a capacidad de cómputo, almacenamiento y paradigmas de ejecución y programación.

Unidad II: Aspectos de arquitectura de las GPUs

Descripción de las principales características de hardware de las GPUs actuales.

Unidad III: Introducción a CUDA

Presentación de CUDA, el framework propuesto por NVIDIA para explotar las GPUs para la resolución de problemas de propósito general. Incluye, entre otros, los conceptos de kernel, grilla, bloque, warp, jerarquía de memoria, divergencia.

Unidad IV: Conceptos básicos de paralelismo

Conceptos básicos de paralelismo necesarios para abordar el uso de GPUs. Ramificación, scheduling, división de datos y de tarea, balance de carga, overhead, condición de carrera

Unidad V: Programación CUDA BÁSICO

Desarrollo de los primeros kernels. Abordaje del problema del dimensionado de los kernels-bloques. Estudio de las condiciones de divergencia. Acceso coalesced a memoria global. Memoria compartida, conflictos de banco.

BIBLIOGRAFÍA

- Programming Massively Parallel Processors: A Hands-on Approach. Second edition, D. Kirk y W. Hwu, ISBN: 978-0-12-415992-1. Elsevier, 2012.
- CUDA by Example: An Introduction to General-Purpose GPU Programming. J. Sanders y E. Kandrot, ISBN: 0131387685. Addison-Wesley Professional, 2010.
- CUDA Programming Guide 8.0. NVIDIA. 2017.
- CUDA C Best Practices Guide Version 8.0. NVIDIA. 2017.