

UNIVERSIDAD NACIONAL DE LA PLATA
FACULTAD DE INFORMÁTICA



TALLER de PROGRAMACIÓN sobre GPUs

Año 2013

Carrera/ Plan:

Licenciatura en Sistemas Plan 2003/07
Licenciatura en Informática Plan 2003/07

Año: Optativa

Regimen de Cursada: Semestral

Carácter:

Optativa

Correlativas: Programación Concurrente

Profesor: Armando De Giusti

Docentes Auxiliares: Adrian Pousa.
Victoria Sanz

Hs. semanales: 6 hs.

FUNDAMENTACIÓN

Los procesadores gráficos (GPUs) han surgido como una alternativa dentro de los procesadores con múltiples núcleos, por sus características de performance y consumo energético.

El uso de GPUs, tanto en computación de alto desempeño como en aplicaciones de propósito general comienza a ser una alternativa de bajo costo para el desarrollo de aplicaciones de muy alto rendimiento que tradicionalmente han sido exclusivas de los clusters de multicores y supercomputadoras.

En este contexto, la metodología e implementación de aplicaciones, tanto utilizando el lenguaje CUDA como OpenGL es un tema de gran interés actual.

Son objetivos de este curso: profundizar el conocimiento de las arquitecturas tipo GPU y su programación, comparar su performance con los multicores convencionales, analizar los modelos de resolución de problemas específicos e introducir conceptos de consumo y green computing a partir de la utilización de GPUs.

OBJETIVOS GENERALES:

Son objetivos de este curso:

- Profundizar el conocimiento de las arquitecturas tipo GPU y su programación.
- Comparar su performance con los multicores convencionales.
- Analizar los modelos de resolución de problemas específicos.
- Introducir conceptos de consumo y green computing a partir de la utilización de GPUs.



CONTENIDOS MINIMOS:

- GPU: Introducción a GPGPU
- Modelo de Memoria de GPU
- Modelo GPU-CPU
- Performance en GPU. Consumo en GPU.
- Programación de GPUs y resolución de aplicaciones.
- Clusters con GPUs y Clusters híbridos.

Programa detallado

GPU: Introducción a GPGPU

- GPU Programación y Arquitectura
- Características de la GPU. Arquitectura de GPU: N-vidia, Ati, Tesla, Fermi...
- GPGPU: Computación de Propósito General en GPU.
- Paradigmas de Computación Paralela en GPUs: Modelo de Memoria Compartida, Paralelismo de Datos.

Modelo de Memoria de GPU

- Modelo de Memoria de GPU.
- Jerarquías de Memoria: Registros, Memoria Compartida, Memoria constante, Texturas, Memoria Global.
- Estructuras de datos básicas.
- Creación de Estructuras de datos en la GPU.
- Ejemplos.

Programación de GPU con CUDA / OPEN GL

- Modelo de programación.
- Diseño de programas en GPU
- Threading. Bloques. Sincronización de Threads
- Estudio experimental de casos.

Modelo GPU-CPU

- Modelo CPU-GPU.
- Transferencia de datos GPU-CPU.
- Modelos de programación paralela en GPU: Reducciones, Map, Scan.
- Ejemplos

Performance y consumo en GPUs

- Análisis de performance.
- Comparación con multicores.
- Estudio de casos, incluyendo el consumo.
- Mediciones experimentales con multicores y GPUs.
- Casos de Estudio.

Arquitecturas híbridas.

- Clusters con GPUs.
- Clusters híbridos.
- Casos de Estudio. Programación de aplicaciones.



UNIVERSIDAD NACIONAL DE LA PLATA
FACULTAD DE INFORMÁTICA

METODOLOGÍA DE ENSEÑANZA Modalidad presencial

La asignatura se estructura con clases teórico-prácticas y prácticas experimentales.

- Las clases teórico-prácticas son dictadas por los Profesores de la asignatura y son obligatorias para la promoción.
- Las explicaciones de práctica son introductorias al trabajo en Laboratorio, para facilitar la utilización del equipamiento y software por los alumnos. Se desarrollan en las clases teórico-prácticas.
- El Taller propone el desarrollo de trabajos concretos con arquitecturas GPU y combinaciones de multicores y GPUs. Las actividades de Taller se hacen en máquina, en el contexto de las clases teórico-prácticas.
- Los trabajos se pueden realizar individualmente o en grupo de 2 personas. Las consultas y correcciones son realizadas en forma presencial o por medio de la plataforma de Educación a Distancia de la UNLP (WEBUNLP).
- En principio se utilizará la Sala de Cómputo de Posgrado (por la disponibilidad de placas GPU) y equipamiento especial del III-LIDI.

EVALUACIÓN Modalidad presencial:

Los alumnos deben aprobar las entregas de los diferentes trabajos experimentales, para obtener la aprobación de los TP de la asignatura.

Posteriormente se les propondrá un trabajo final experimental a defender en un coloquio como Trabajo Final de Promoción, en una fecha de examen final..

METODOLOGÍA DE ENSEÑANZA Modalidad no presencial

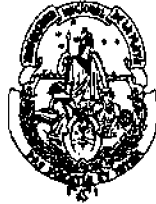
Se hace notar que por la característica de las tareas experimentales, el alumno deberá tener acceso a algún modelo de arquitectura con GPU para poder realizar los trabajos que se solicitan en el curso.

El alumno puede seguir los temas por el entorno WEB-UNLP y asistir a las consultas que se fijen para los alumnos presenciales.

EVALUACIÓN Modalidad no presencial

Deben cumplir las entregas de los mismos trabajos experimentales para obtener la aprobación de los TP de la asignatura.

Deben rendir examen final, que puede incluir un trabajo final experimental, pero necesariamente contendrá preguntas de la Teoría presencial.



BIBLIOGRAFÍA BASICA

M. F. Piccoli, "Computación de Alto Desempeño utilizando GPU". XV Escuela Internacional de Informática. Editorial Edufp, 2011.

Guil N. y Ujaldón M. "La GPU como arquitectura emergente para supercomputación". In *XIX Jornadas de Paralelismo de Castellon*. 2008.

Kirk, D.,Hwu, W.. "Programming Massively Parallel Processors: A Hands-on Approach". ISBN: 978-0-12-381472-2. Elsevier. 2010.

Luebke D. H.G. "How GPUs work". *EEE Computer*, 40(2), 2007.

Sanders, J., Kandrot, E.. "Cuda by Example: An Introduction to General- Purpose Gpu Programming". ISBN: 0131387685. Addison-Wesley Professional. 2010.

General-Purpose Computation on Graphics Processing Units. <http://ggpu.org>.

Kerr A.and Damos G. y Yalamanchili S. "Modeling GPU-CPU workloads and systems". In *3rd Workshop on GP Computation on Graphics Processing Units*. ACM, 2010.

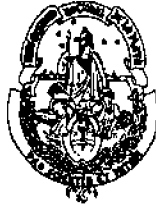
Grama A, Gupta A, Karypis G, Kumar V. "Introduction to parallel computing". Second Edition. Pearson Addison Wesley, 2003.

Kindratenko, V.V el al "GPU clusters for high-performance computing," Cluster Computing and Workshops, 2009. CLUSTER '09. IEEE International Conference on , vol., no., pp.1,8, Aug. 31 2009-Sept. 4 2009

<http://www.cs.caltech.edu/courses/cs101gpu/>

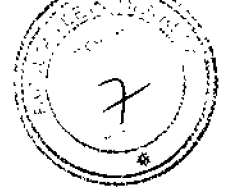
Montes de Oca E., De Giusti L., De Giusti A., Naiouf M. "Comparación del uso de GPU y cluster de multicore en problemas con alta demanda computacional". XII Workshop de Procesamiento Distribuido y Paralelo. CACIC2012. ISBN: 978987-1648-34-4. Pág. 267-275. Bahía Blanca, Buenos Aires, Argentina, Octubre 2012.

Montes de Oca E., Naiouf M., De Giusti L., Chichizola F., Giacomantone J., De Giusti A. "Una implementación paralela de las Transformadas DCT y DST en GPU. Análisis de performance". XII Workshop de Procesamiento Distribuido y Paralelo. CACIC2012. ISBN: 978987-1648-34-4. Pág. 276-285. Bahía Blanca, Buenos Aires, Argentina, Octubre 2012.



BIBLIOGRAFÍA ADICIONAL

- Joselli M., Zamith M., Clua E., Montenegro A., Conci A., Leal-Toledo R., Valente L., Feijo B., Dórnellas M., y Pozzer C. "Automatic dynamic task distribution between CPU and GPU for real-time systems". In *11th IEEE International Conference on Computational Science and Engineering*. 2008.
- OpenGL Red Book – General resource for OpenGL/graphics programming
- OpenGL Orange Book GPU/GLSL version of the Red Book
- The CUDA Zone: <http://www.nvidia.com/cuda> Examples, documentation, drivers, etc.
- NVIDIA. "Nvidia cuda compute unified device architecture, programming guide version 2.0". In *NVIDIA*. 2008a.
- NVIDIA. "Nvidia geforce 8800 gpu architecture overview". In *NVIDIA*. 2006.
- NVIDIA. Nvidia geforce gtx 200 gpu architectural overview. In *NVIDIA*. 2008b.
- W. Hwu)Buck I. "Gpu computing with Nvidia Cuda". *ACM SIGGRAPH 2007 courses ACM*, 2007. New York, NY, USA.
- Chen W. y Hang H. "H.264/avc motion estimation implementation on compute unified device architecture (cuda)". In *IEEE*, editor, *IEEE International Conference on Multimedia*. 2008.
- Goyal N., Ormont J., Smith R., Sankaralingam K., y Estan C. "Signature matching in network processing using simd-gpu architectures". In *University of Wisconsin*. 2008.
- Lieberman M., Sankaranarayanan J., y Samet H. "A fast similarity join algorithm using graphics processing units". In *ICDE 2008. IEEE 24th International Conference on Data Engineering 2008*. 2008.
- Lloyd D., Boyd C., y Govindaraju N. "Fast computation of general fourier transforms on gpuS". In *IEEE International Conference on Multimedia and Expo*. 2008.
- Luebke D. "Cuda: Scalable parallel programming for high-performance scientific computing". In *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2008*. 2008.
- Nottingham A. y Irwin B. "Gpu packet classification using opencl: a consideration of viable classification methods". In *Research Conf. of the South African Inst. of Comp. Sc. and Inf. Technologists*. ACM, 2009.
- Ryoo S., Rodríguez C., Baghsorkhi S., Stone S., Kirk D., y Hwu W. Optimization principles and application performance evaluation of a multithreaded GPU using CUDA. In *ACM*. ACM, 2008.
- Mc. Cool M. "Programming models for scalable multicore programming". 2007. <http://www.hpcwire.com/features/17902939.html>
- Rucci E., De Giusti A., Chichizola F., Naiouf M., De Giusti L. "DNA Sequence Alignment: hybrid parallel programming on multicore cluster". Proceedings of the International Conference on Computers, Digital Communications and Computing (ICDCCC '11), Vol. 1, Nikos Mastorakis, Valeri Mladenov, Badea Lepadatescu, Hamid Reza Karimi, Costas G. Helmis (Editors), WSEAS Press, September 15-17, 2011, Barcelona, ISBN: 978-1-61804-030-5, pp. 183-190.
- Feng, W.C., "The importance of being low power in high-performance computing". *Cyberinfrastructure Technology Watch Quarterly (CTWatch Quarterly)*. 2005.
- Muresano Cáceres R. "Metodología para la aplicación eficiente de aplicaciones SPMD en clústers con procesadores multicore" Ph.D. Thesis, Universidad Autónoma de Barcelona, Barcelona, España, Julio 2011.
- Sinha, R.; Prakash, A.; Patel, H.D., "Parallel simulation of mixed-abstraction SystemC models on GPUs and multicore CPUs," *Design Automation Conference (ASP-DAC), 2012 17th Asia and South Pacific*, vol., no., pp.455,460, Jan. 30 2012-Feb. 2 2012.
- Lingyan Wang, Miaqing Huang, and Tarek El-Ghazawi, "Towards efficient GPU sharing on multicore processors". In *Proceedings of the second international workshop on Performance modeling, benchmarking and simulation of high performance computing systems (PMBS '11)*. ACM, New York, NY, USA, 23-24.
- Chao-Tung Yang, Chih-Lin Huang, Cheng-Fang Lin. "Hybrid CUDA, OpenMP, and MPI parallel programming on multicore GPU Clusters", *Computer Physics Communications* 182 (2011) 266–269, Elsevier.
- Alexandra Fedorova, Juan Carlos Saez, Daniel Shelepov and Manuel Prieto. Maximizing Power Efficiency with Asymmetric Multicore Systems. *Communications of the ACM*, Vol. 52 (12), pp 48-57. December 2009.
- Nottingham A. y Irwin B. "Gpu packet classification using opencl: a consideration of viable classification methods". In *Research Conf. of the South African Inst. of Comp. Sc. and Inf. Technologists*. ACM,



CRONOGRAMA TENTATIVO DE CLASES Y EVALUACIONES

El cronograma detallado se pone en conocimiento de los alumnos al inicio del curso.

Clase	Contenidos/Actividades
1- Semana del 5/08	<u>GPU: Introducción a GPGPU</u> <ul style="list-style-type: none">GPU Programación y ArquitecturaCaracterísticas de la GPU. Arquitectura de GPU: N-vidia, Ati, Tesla, Fermi.
2- Semana del 12/08	<u>GPU: Introducción a GPGPU</u> <ul style="list-style-type: none">GPGPU: Computación de Propósito General en GPU.Paradigma de Computación Paralela en GPUs: Modelo de Memoria Compartida, Paralelismo de Datos.
3- Semana del 19/08	<u>Modelo de Memoria de GPU</u> <ul style="list-style-type: none">Modelo de Memoria de GPU.Jerarquías de Memoria: Registros, Memoria Compartida, Memoria constante, Texturas, Memoria Global.Estructuras de datos básicas.
4- Semana del 26/08	<u>Modelo de Memoria de GPU</u> <ul style="list-style-type: none">Creación de Estructuras de datos en la GPU.Ejemplos.Programación de casos.
5- Semana del 29/4	<u>Programación de GPU con CUDA</u> <ul style="list-style-type: none">Modelo de programación.Diseño de programas en GPUThreading. Bloques. Sincronización de ThreadsEstudio experimental de casos.
6- Semana del 02/09	<u>Programación de GPU con OPEN GL</u> <ul style="list-style-type: none">Modelo de programación.Diseño de programas en GPUDiferencias con CUDAEstudio experimental de casos.
7- Semana del 09/09	<u>Modelo GPU-CPU</u> <ul style="list-style-type: none">Modelo CPU-GPU.Transferencia de datos GPU-CPU.Programación paralela en GPU: Reducciones, Map, Scan.Ejemplos
8- Semana del 16/09	<u>Performance y consumo en GPUs</u> <ul style="list-style-type: none">Análisis de performance.Comparación con multicores.Estudio de casos, incluyendo el consumo.Mediciones experimentales con multicores y GPUs.Casos de Estudio.
9- Semana del 23/09	<u>Performance y consumo en GPUs</u> <ul style="list-style-type: none">Comparación de casos en clusters con multicores.Consumo comparativo de GPUs y multicores.Casos de Estudio.



UNIVERSIDAD NACIONAL DE LA PLATA
FACULTAD DE INFORMÁTICA

10- Semana del 30/09	<u>Arquitecturas de Cluster</u> <ul style="list-style-type: none">▪ Clusters con GPUs.▪ Comunicación con el servidor del cluster.▪ E/S▪ Casos de Estudio. Programación de aplicaciones.
11- Semana del 7-10	<u>Arquitecturas de Cluster híbridos</u> <ul style="list-style-type: none">▪ Clusters con GPUs y Multicores▪ Comunicaciones.▪ Soportes de S.O. y de Lenguajes.▪ Aplicaciones.
12- Semana del 14-10	Entrega y Evaluación de Trabajos Prácticos.
13- Semana del 21-10	Entrega y Evaluación de Trabajos Prácticos
14- Semana del 28-10	CONSULTAS
15- Semana del 4-11	Definición de Trabajos Finales
16- Semana del 11-11	Recuperación de Entregas de TP y Definición Trabajos Finales

Contacto de la cátedra (mail, página, plataforma virtual de gestión de cursos):

Plataforma virtual: webunlp.unlp.edu.ar

Web: <http://weblidi.info.unlp.edu.ar/catedras/tallerGPU/>

Mail: tallerGPU@lidi.info.unlp.edu.ar

Firmas del profesor responsable:

Ing. ARMANDO DE GIUSTI
PROF. TITULAR DED. EXCL.
INV. PRINCIPAL CONICET