Especialización en Cómputo de Altas Prestaciones – Modalidad a distancia

Arquitecturas para Cómputo de Altas Prestaciones

Año 2021

Duración: 70 hs. Totales.

Cantidad de horas presenciales/VC: 25 hs.

Cantidad de horas de actividades en línea y de trabajo final: 45 hs.

OBJETIVOS GENERALES:

- Revisar las técnicas actuales de diseño de procesadores, dando una visión integrada de las interdependencias entre la evolución de la tecnología y la arquitectura de estos procesadores integrados.
- Analizar las arquitecturas actuales multicore (simétricos y asimétricos) y multithreading con énfasis en la gestión de recursos compartidos y su impacto en el consumo y la programación.
- Exponer las características de arquitecturas orientadas, tales como GPUs, FPGAs, MIC y TPUs. Estudiar las configuraciones híbridas que vinculan servidores multicore con estas nuevas arquitecturas.
- Configuraciones para cómputo de altas prestaciones: clusters, supercomputadores y cloud.

COMPETENCIAS A DESARROLLAR EN RELACION CON EL OBJETIVO DE LA CARRERA

C.1- Analizar problemas del mundo real que por su complejidad y/o volumen de datos requieran cómputo paralelo y diseñar soluciones desde el punto de vista del hardware necesario, lo que requiere un conocimiento de las arquitecturas paralelas actuales.

C.2- Conocer los fundamentos para el desarrollo de Sistemas Paralelos (incluyendo la relación entre hardware y software).

C.4- Conocer y analizar arquitecturas dedicadas para procesamiento paralelo. Tener capacidad de configurar arquitecturas y desarrollar programación en la nube (Cloud Computing).

C.5- Conocer las tecnologías y el manejo de sistemas de Fog Computing y Edge Computing, su middleware y aplicaciones.

CONTENIDOS MINIMOS:

- Conceptos de microarquitecturas paralelas.
- Multithreading. Multicores simétricos y asimétricos.
- Otras arquitecturas multiprocesador actuales: GPUs / FPGAs / MIC / TPUs
- Configuración de clusters / supercomputadores y cloud.

PROGRAMA

- Microarquitectura y paralelismo a nivel de instrucción
  - Tendencias tecnológicas en la arquitectura de Procesadores. Coste, rendimiento, consumo.
  - Paralelismo a nivel de instrucción: planificación dinámica. Tratamiento de dependencias de control: Predicción de saltos. Especulación.
  - Ejecución de múltiples instrucciones por ciclo. Límites del paralelismo a nivel de instrucción.
  - Acceso a Memoria: Prebúsqueda SW, Prebúsqueda HW, Caches sin bloqueo.
  - Especulación de Load. Manejo del flujo de datos. Localidad de datos.; Técnicas no especulativas y Técnicas especulativas

- Paralelismo a nivel de thread. Motivación.
  - Procesadores multithreading, Formas de multithreading. ejemplos
  - Multiprocesadores en un chip (Multi/Many cores).
  - Ley de Amdahl en multicores
  - Modelos HW
  - Multicores simétricos. Multicores asimétricos. Multicores dinámicos

- Arquitecturas orientadas a procesamiento paralelo en Cómputo de Altas Prestaciones
  - GPUs. Clusters de GPUs. Lenguajes y Programación paralela sobre GPUs.
  - FPGAs (Field Programmable Arrays). Características. Programación sobre FPGAs.
  - Placas MIC. (many integrated core architectures). Incorporación de las placas MIC a servidores para cómputo de altas prestaciones.
  - TPUs (Unidades de procesamiento tensorial). Características. Incorporación de arquitecturas TPU para procesamiento paralelo. Lenguajes de programación sobre TPUs.
  - Estudios de integración de servidores multicore con placas orientadas de GPUs / MIC / TPU / FPGA.

- Configuraciones de Servidores para Cómputo de Altas Prestaciones.
  - Clusters de multicores, GPUs e híbridos.
  - Configuraciones de supercomputadores actuales.
  - Cloud. Configuraciones en la nube.
  - Métricas de rendimiento y eficiencia, en diferentes modelos de servidores.

ACTIVIDADES EXPERIMENTALES y DE INVESTIGACION

Tareas en Laboratorio (presencial o remoto)

Tal como se explica en el item relacionado con la metodología, ésta se basa en clases sincrónicas (presenciales o remotas) combinadas con actividades demostrativas en el laboratorio para aplicar los conceptos teóricos y que así el alumno adquiera las competencias y habilidades sobre cada uno de los temas que forman parte del contenido de la asignatura.
Además el alumno debe analizar un proyecto/desarrollo relacionado con los temas dictados en la teoría, cuya implementación concreta se realiza sobre placas / máquinas específicas / clusters y el vínculo con un Cloud privado en la misma Facultad y/o con los Clouds de acceso académico de los proveedores de este servicio (Amazon / Microsoft, etc).

Investigación/ Estudios adicionales:
Los alumnos analizarán papers relacionados con la evolución de las arquitecturas de procesamiento paralelo, desde el nivel de microarquitectura hasta la configuración de supercomputadores o cloud.
Se pondrá énfasis en el análisis de las herramientas de desarrollo de software y programación paralela en cada caso, de modo de comparar alternativas.
Se les propondrán temas de I+D orientados al estudio comparativo de evaluación de rendimiento en diferentes tipos de arquitecturas multiprocesador, de modo de potenciar el conocimiento trasmitido en la teoría.

METODOLOGIA Y MODALIDAD DE EVALUACION

La metodología se basa en clases sincrónicas a través del sistema de videoconferencias adoptado por el Postgrado de Informática combinadas con sesiones en el laboratorio remoto para aplicar los conceptos teóricos y que así el alumno adquiera las competencias y habilidades sobre cada uno de los temas que forman parte del contenido de la asignatura.
Es de hacer notar que en el laboratorio remoto, además de las máquinas convencionales, se dispone de placas con aceleradores (GPUs, MIC, TPU y FPGAs) sobre los cuales se pueden plantar y resolver problemas que hacen al objetivo de la asignatura.
Se requiere un 80% de asistencia a los encuentros sincrónicos, incluyendo el encuentro inicial de presentación de la materia, y el encuentro final de integración, ambos de asistencia obligatoria.
El trabajo se complementa con un proyecto experimental que debe desarrollar el alumno para cumplimentar las horas asignadas con soporte tutorizado por el profesor (*on-line*) y seguimiento a través del Entorno Virtual IDEAS contemplado en el SIED de la Facultad de Informática de la UNLP.

La evaluación se realizará mediante un examen escrito al final de las sesiones sincrónicas para evaluar el grado de conocimientos del alumno (20%), el proyecto/desarrollo experimental que deberá entregar el alumno al final de las horas programadas (70%) y la participación y aportaciones de calidad/excelencia a las soluciones propuestas (10%).

RECURSOS Y MATERIALES DE ESTUDIO

Como materiales de estudio, se dispone de:

- Presentaciones multimedia desarrolladas ad-hoc para introducir cada uno de los diferentes ejes temáticos.
- Píldoras formativas con la explicación de algunos temas
- Ejemplos donde se aplican los conceptos teóricos
- Ejercicios prácticos que son desarrollados en clase
- Material de lectura para estudiar y profundizar conceptos abordados en las clases
- Enlaces a artículos de actualidad de repositorios reconocidos en el área
- Acceso a equipamiento remoto situado en la Facultad de Informática de la UNLP y también en la nube (Cloud).

- Software específico para determinadas actividades de laboratorio que se detallan en este programa.

ACTIVIDADES EXPERIMENTALES Y APROPIACIÓN DE SABERES

Los trabajos experimentales pueden desarrollarse en cada clase o continuarse en más de una clase. Parten de una especificación/consigna del docente (explicada en la clase) y un trabajo individual o en grupos que interactúan en el que los alumnos resuelven un problema experimental concreto relacionado con la temática.

Los trabajos podrán ser individuales o grupales. Para esto último se configura el entorno virtual para que los alumnos del mismo grupo se encuentren en un espacio virtual diferente del resto. Durante el desarrollo del trabajo, el docente estará conectado respondiendo dudas y consultas.

Estos trabajos pretenden desarrollar y/o fortalecer las aptitudes de opinión crítica en los temas relativos del curso. Los alumnos deberán sintetizar su comprensión de los temas, al realizar correctamente la tarea experimental propuesta.

También se pretende desarrollar la capacidad de poder comunicar y transmitir los resultados, en presentaciones pautadas a lo largo del curso.

En general, finalizada una actividad, hay una sesión de discusión conjunta donde los participantes comunicarán sus opiniones e intercambiarán los distintos puntos de vista.

BIBLIOGRAFÍA BASICA

"Parallel programming for Multicores and Cluster systems"
T. Rauber, G. Rünger. Springer. 2013.

"Fundamentals of Parallel Multicore Architectures "
Yan Solihin. CRC. 2015

"Computer Architecture: A Quantitative Approach" (5 edition)
J. Hennessy, D. Patterson, Morgan Kaufmann Publishers, Inc. 2012.

"Processor Microarchitecture.*An Implementation Perspective"*
A Gonzalez, F Latorre, G Magklis, Synthesis Lecture on Computer Science, Morgan&Claypool, 2011.

"**Multithreading Architecture**"
Mario Nemirovsky, Dean M. Tullsen, Synthesis Lecture on Computer Science, Morgan&Claypool, 2013.

"High-performance Dynamic Programming on FPGAs with OpenCL".
Sean Settle:In: IEEE High Performance Extreme Computing Conference. 2013.

"Cloud Tensor Processing Unit (TPU)"
Google Inc. Disponible en https://cloud.google.com/tpu/docs/tpus?hl=es-419

BIBLIOGRAFÍA COMPLEMENTARIA

"ACFS: A Completely Fair Scheduler for Asymmetric Single-ISA Multicore Systems".
Juan Carlos Saez, Adrian Pousa, Daniel Chaver, Fernando Castro, Manuel Prieto Matias:In: ACM SAC 2015 (The 30TH ACM/SIGAPP Symposium on applied computing). 2015.

"Accelerating Pattern Matching on Intel Xeon Phi Processors",
V. Sanz, A. Pousa, M. Naiouf, and A. De Giusti Algorithms and Architectures for Parallel Processing. ICA3PP 2020., ISBN: 978-3-030-60245-1, págs. 262-274, doi. 10.1007/978-3-030-60245-1_18, 2020.

"Towards a Malleable Tensorflow Implementation",
L. A. Libutti, F. D. Igual, L. Piñuel, L. D. Giusti, and M. Naiouf, Cloud Computing, Big Data & Emerging Topics. 8th Conference, JCC-BD&ET 2020, La Plata, Argentina, September 8-10, 2020, Proceedings, págs. 30-40, doi. 10.1007/978-3-030- 61218-4_3, 2020.

"Comparación de Arquitecturas HPC para Computar Caminos Mínimos en Grafos. Intel Xeon Phi KNL vs NVIDIA Pascal"
Costanzo, M., Rucci, E., Costi, U., Chichizola, F., Naiouf, M.In: Actas del XXVI Congreso Argentino de Ciencias de la Computación (CACIC 2020).

"Modern Processor Design"
J.P. Shen, M. H. Lipasti, McGraw Hill, 2005.

"Microprocessor Architecture"
J-L Baer, Crambridge University Press, 2010.

"Parallel and distributed computing. Architectures and Algorithms"
S.K Basu. Phi Learning 2016.

"Chip Multiprocesor Architecture"
K Olukotun, L Hammond, J Laudon, Synthesis Lecture on Computer Science, Morgan&Claypool, 2007.