Asignatura: Computación a Gran Escala (COMP)
Código: 32497
Institución: Escuela Politécnica Superior
Grado: Master en Ingeniería Informática
Nivel: Master
Tipo: Troncal
ECTS: 6

GUIA DOCENTE: Computación a Gran Escala

Año académico: 2016-2017

Programa: Master en Ingeniería Informática
Centro: Escuela Politécnica Superior
Universidad: Universidad Autónoma de Madrid

Última modificación: 2016/09/09
Status: Aprobado 2013/05/29

1. ASIGNATURA (ID)

Cálculo intensivo y manejo de datos a gran escala

1.1. Programa

Master en Ingeniería Informática

1.2. Código de curso

32416

1.3. Áreas del curso

Ciencias de la Computación e Inteligencia Artificial

1.4. Tipo de asignatura

Obligatoria [itinerario: todos los itinerarios]

1.5. Semestral

Primer trimestre

1.6. Créditos

6 ETCS

1.7. Lenguaje

Las clases serán en español aunque algunos seminarios podrían ser en inglés.

1.8. Recomendaciones / Asignaturas relacionadas

Conocimientos de C, Bases de Datos, probabilidad y estadística a nivel básico son necesarios para seguir el curso

Asignaturas relacionadas:

- Procesamiento de información temporal
- Aprendizaje Automático: teoría y aplicaciones
- Métodos bayesianos aplicados

- Aceleración de algoritmos en sistemas heterogéneos
- Procesamiento de señales biomédicas y sus aplicaciones
- Procesamiento de audio y voz para biometría y seguridad
- Técnicas de análisis de secuencias vídeo para videovigilancia

1.9. Profesores

Añadir @uam.es a todas las direcciones de correo posteriores

Profesores:

Dr. Carlos Santa Cruz Fernández (Coordinador)
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Office: B-343
Tel.: +34 914972337
e-mail: carlos.santacruz
Web: http://www.eps.uam.es/~santacru

Dr. Miguel Ángel García García
Departamento de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Office: C-242
Tel.: +34 914976208
e-mail: miguelangel.garcia
Web: http://www.eps.uam.es/~mgarcia/

Dr. Estrella Pulido Cañabate
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Office: B-413
Tel.: +34 914972289
e-mail: estrella.pulido
Web: http://www.eps.uam.es/~epulido/

1.10. Objetivos de la asignatura

Este asignatura está divida en tres partes. La primera parte corresponde a una introducción a las técnicas de programación eficiente de las arquitecturas paralelas de memoria compartida, incluyendo los procesadores multi-núcleo y los multiprocesadores fuertemente acoplados. En concreto, se describen herramientas de análisis de rendimiento (*profilers*), se identifican los principales factores que afectan a la eficiencia de las arquitecturas paralelas, y se estudian técnicas de paralelización de bucles y programas secuenciales en arquitecturas de memoria compartida mediante OpenMP.
El objetivo de la segunda parte es introducir los algoritmos numéricos básicos y las herramientas para el manejo y manipulación de matrices, solución de ecuaciones algebraicas lineales y el problema de autovalores. Estos algoritmos se utilizan para resolver modelos generales de regresión lineal y Análisis de Componentes Principales (PCA) para la reducción de dimensiones. Octave se utiliza como herramienta para ejecutar y resolver los problemas propuestos.
Por último, la tercera parte proporciona el marco de Ingeniería, las técnicas y las herramientas necesarias para diseñar y gestionar el almacenamiento de grandes bases de datos, incluido el preprocesamiento e integración de datos, así como las herramientas OLAP para el análisis interactivo de datos multidimensionales. Además, el objetivo es entender lo que es la inteligencia de negocios, cómo funciona, dónde se usa, y por qué y cuándo utilizarla. Se describen y analizan también las principales herramientas de BI existentes en el mercado.

Al final de cada unidad el estudiante debe ser capaz de:

OBJETIVOS ESPECÍFICOS DE CADA UNIDAD	
PARTE I	
UNIDAD 1.- Introducción	
1.1.	Conocer las diferentes familias de arquitecturas paralelas y escoger las que mejor se ajusten a un ámbito de aplicación específico.
1.2.	Medir el rendimiento de algoritmos paralelos en términos de *speedup* y eficiencia.
1.3.	Entender los conceptos básicos de programación paralela, incluyendo tareas, procesos y mecanismos de sincronización.
1.4.	Utilizar herramientas de análisis de rendimiento (*profilers*) para analizar la eficiencia de algoritmos secuenciales e identificar porciones susceptibles de ser aceleradas mediante paralelización.
1.5.	Implementar algoritmos paralelos simples sobre arquitecturas paralelas de memoria compartida usando C y OpenMP.
UNIDAD 2.- Paralelización de bucles en arquitecturas paralelas de memoria compartida	
2.1.	Entender los diferentes tipos de bucles paralelos y escoger los que mejor se ajustan a un problema específico.
2.2.	Describir las dependencias entre iteraciones de un conjunto de bucles secuenciales mediante un grafo de dependencias.
2.3.	Paralelizar un conjunto de bucles secuenciales a partir de su correspondiente grafo de dependencias.
2.4.	Aplicar transformaciones de código de cara a optimizar la paralelización de un conjunto de bucles secuenciales.
UNIDAD 3.- Proceso general de paralelización de programas secuenciales	
3.1.	Descomponer algoritmos secuenciales en tareas susceptibles de ser paralelizadas.
3.2.	Maximizar el equilibrio de carga en la asignación de tareas a procesos.
3.3.	Minimizar costes de comunicación en la asignación de tareas a procesos.
3.4.	Minimizar la sincronización y las sobrecargas de gestión en la asignación de tareas a procesos.
3.5.	Determinar la planificación de procesos que maximiza la eficiencia.
3.6.	Determinar el mapeo de procesos a unidades de cómputo que maximiza la eficiencia.
PARTE II	
UNIDAD 4.- Introducción	
4.1.	Entender la notación en como flotante de los procesadores actuales
4.2.	Entender que el error de redondeo se debe a que la aritmética entre números no es exacta.
4.3.	Entender que el Erro debido a los algoritmos son independientes del hardware
4.4.	Entender que los errores pueden ser sucesivamente magnificados debido a los algoritmos
UNIDAD 5.- Solución de sistemas de ecuaciones lineales	

5.1.	Usar la descomposición LU para resolver sistemas lineales de ecuaciones
5.2.	Usar la descomposición de Cholesky para resolver sistemas lineales de ecuaciones
5.3.	Escribir un problema de regresión lineal múltiple en notación matricial
5.4.	Resolver algunos ejemplos prácticos
UNIDAD 6.- Autovalores y Autovectores	
6.1.	Obtener los principales autovalores y autovectores de una matriz
6.2.	Entender el algoritmo Page Rank de Google
6.3.	Calcular los autovalores y autovectores de una matriz simétrica
6.4	Entender el algoritmo de Análisis de Componentes Principales (ACP)
PART III	
UNIDAD 7.- Introducción to Data Warehousing	
7.1.	Explicar los objetivos de warehousing
7.2.	Extracción de datos, transformación, técnicas de carga para data warehousing.
7.3.	Explicar la terminología aceptada del data warehouse
7.4.	Describir los métodos y herramientas de extracción, transformación y carga de datos.
7.5.	Identificar alguna de las herramientas para acceder y analizar los datos.
UNIDAD 8.- Business Intelligence	
8.1.	Usar los términos y conceptos en business intelligence
8.2.	Explicar como los sistemas funcionan, sus debilidades y fortalezas
8.3.	Explotar los sistemas analíticos y las medidas de rendimiento
8.4.	Analizar las nuevas tendencias en BI

1.11. Contenido del curso

PARTE 1

1. **Introducción**
   - 1.1. Arquitecturas paralelas: motivación.
   - 1.2. Arquitecturas paralelas de memoria compartida.
     - 1.2.1. Procesadores multi-core y multiprocesadores fuertemente acoplados.
     - 1.2.2. OpenMP.
   - 1.3. Arquitecturas paralelas de memoria distribuida: multiprocesadores débilmente acoplados.
   - 1.4. Computación en malla (*grid computing*) y en la nube (*cloud computing*).
   - 1.5. Conceptos básicos.
     - 1.5.1. Tareas y procesos.
     - 1.5.2. Semáforos y barreras.
     - 1.5.3. Herramientas de análisis de rendimiento (*profilers*).

2. **Paralelización de bucles en arquitecturas paralelas de memoria compartida.**
   - 2.1. Bucles paralelos.
   - 2.2. Planificación (*scheduling*) de bucles paralelos.
   - 2.3. Análisis de dependencias entre iteraciones.
     - 2.3.1. Dependencias verdaderas.
     - 2.3.2. Antidependencias.
     - 2.3.3. Dependencias de salida.
     - 2.3.4. Dependencias en bucles imbricados.
     - 2.3.5. Grafos de dependencias por niveles.
   - 2.4. Generación de código paralelo.
     - 2.4.1. Componentes fuertemente conexas.
     - 2.4.2. Condensación acíclica.
     - 2.4.3. Arcos libres de barreras.
     - 2.4.4. Generación de *clusters* y segmentos.
     - 2.4.5. Generación de código.
       - 2.4.5.1. Generación de código para segmentos serie.
       - 2.4.5.2. Generación de código para segmentos paralelos.
     - 2.4.6. Problemas de ejemplo.
   - 2.5. Transformaciones para soportar paralelización.
     - 2.5.1. Normalización de bucles.
     - 2.5.2. Substitución escalar.
     - 2.5.3. Expansión escalar.
     - 2.5.4. Copiado de variables.
     - 2.5.5. Intercambio de bucles.

3. **Proceso general de paralelización de programas secuenciales.**
   - 3.1. Descomposición.
   - 3.2. Asignación.
     - 3.2.1. Equilibrado de carga.
     - 3.2.2. Reducción de comunicación.

3.2.3. Reducción de sobrecargas.
3.3. Orquestación.
3.4. Mapeo.

PARTE II

4. **Introducción**
   4.1. Representación en coma flotante
   4.2. Error de redondeo
   4.3. Error de truncado
   4.4. Estabilidad

5. **Solución de sistemas de ecuaciones lineales**
   5.1. Descomposición LU
   5.2. Descomposición de Cholesky
   5.3. Regresión Lineal múltiple en notación matricial
   5.4. Aplicaciones prácticas

6. **Autovalores**
   6.1. Método de Potencias
   6.2. Aplicación al algoritmo Page Rank
   6.3. Transformación de Jacobi de una matriz simétrica
   6.4. Análisis de Componentes Principales y sus aplicaciones

PARTE III

7. **Introducción al Data Warehousing**
   7.1. Sistemas para la toma de decisiones
   7.2. Data warehousing
   7.3. Arquitectura para el Data warehouse
   7.4. ETL (extraction, cleansing, transformation and loading)
   7.5. Modelo multidimensional
   7.6. Meta-data
   7.7. Acceder al data warehouses
   7.8. Temas adicionales: seguridad, calidad …

8. **Business Intelligence**
   8.1. Introducción y conceptos
   8.2. Modelos de usuario para Business Intelligence
   8.3. BI Productos y proveedores
      8.3.1. Enterprise Business Intelligence products
      8.3.2. Database and Packaged Products
      8.3.3. Data Discovery & Visualization

1.12. Bibliografía

1. "Parallel Computer Architecture: A Hardware/Software Approach" D. Culler, J.P. Singh, A. Gupta. *Ed. Morgan Kaufmann, 1998.*
2. "Supercompilers for parallel and vector computers" H. Zima, B. Chapman *Ed. ACM Press, 1991.*
3. "Optimizing Compilers for Modern Architectures: A Dependence-based Approach". Allen, K. Kennedy *Ed. Morgan Kaufmann, 2001.*
4. "Computer Architecture: A Quantitative Approach" (5a. ed.) J.L. Hennessy, D.A. Patterson *Ed. Morgan Kaufmann, 2011.*
5. "Advanced Computer Architecture: Parallelism, Scalability, Programmability" K. Hwang. *Ed. McGraw-Hill, 1992.*
6. "Numerical Recipes in C: The Art of Scientific Computing", W. H. Teulosky, A. A. Vetterling, W. T. Flannery, B. P., Cambridge University Press, 1992
7. "Numerical Mathematics. Theory and Computer Applications" C. E. Froberg.. Addison-Wesley, Reading, Massachusetts, 1985.
8. "Scientific Computing: An Introductory Survey" M. T. Heath., 2nd. ed. McGraw-Hill, New York, 2001.
9. "Análisis numérico con aplicaciones" C. F. Gerald and P.O. Wheatley., 6a ed.Prentice Hall, México, 2000.
10. "Data Warehouse Design: Modern Principles and methodologies" M. Golfarelli, S. Rizzi. *McGraw-Hill, 2009.*
11. "Data Warehousing Fundamentals for IT Professionals" P. Ponniah. J*ohn Wiley & Sons. 2010.*
12. "The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing" and Business Intelligences" R. Kimball, M. Ross. *John Wiley & Sons. 2010.*
13. "Business Intelligence" R. Sabherwal, I. Becerra-Fernandez. *John Wiley & Sons. 2010.*
14. "Business Intelligence: Data Mining and Optimization for Decision Making" C.Vercellis. *John Wiley & Sons. 2009.*
15. "Decision Support and Business Intelligence Systems" E. Turban, R. Sharda, D. Delen *Prentice Hall. 2010.*
16. "Business Intelligence: A Managerial Approach" E. Turban; R. Sharda; D. Delen; D. King; J. Aronson. *Prentice Hall, 2011.*
17. "Business Intelligence" S. Misner, E. Vitt *Microsoft Press, 2008*

1.13. Trabajos y evaluación

El curso consta de clases presenciales, tareas semanales, tareas de laboratorio y un examen final.

En el periodo ordinario de examen es necesario aprobar (>=5) el examen para pasar el curso. En el periodo extraordinario es solo necesario aprobar (>=5) el trabajo de investigación para aprobar el curso.

- En el periodo ordinario la evaluación constará de las siguientes partes:
  - 50 % Trabajos de laboratorio
  - 50 % Examen

  Las notas de las partes se guardarán para el periodo extraordinario

- En caso de suspender en el periodo ordinario, en el periodo extraordinario el alumno tiene la oportunidad de:
  - Presentar todas las practicas después de haberlas corregido de nuevo
  - Presentar un trabajo de investigación sobre un tema acordado con el profesor de la asignatura.

  La nota final será determinada por
  - 50 % Prácticas de laboratorio [solo si han sido presentadas de Nuevo]
  - 50 % Trabajo de investigación [solo si el trabajo se presenta]

1.14. Horario

Semana	Contenido
1	**H1:** Presentación del curso. **H2:** Unidades 1.1 a 1.3. **H3:** Unidades 1.4 y 1.5.
2	**H1:** Laboratorio 1: Profilers (perf y gprof). **H2:** Laboratorio 1: OpenMP (compilación y directivas básicas). **H3:** Unidades 2.1 y 2.2.
3	**H1:** Unidades 2.3.1 a 2.3.3. **H2:** Unidades 2.3.4 y 2.3.5. **H3:** Unidades 2.4.1 a 2.4.3.
4	**H1:** Laboratorio 2: *OpenMP* (Bucles paralelos) **H2:** Laboratorio 2: *OpenMP* (Bucles paralelos) **H3:** Unidad 2.4.4.
5	**H1:** Unidad 2.4.5. **H2:** Unidad 2.4.6 (I). **H3:** Unidad 2.4.6 (II).
6	**H1:** Laboratorio 3: Paralelización de bucles con *OpenMP*. **H2:** Laboratorio 3: Paralelización de bucles con *OpenMP*. **H3:** Unidad 2.5 (I).
7	**H1:** Unidades 2.5 (II) y 3.1. **H2:** Unidad 3.2 (I). **H3:** Unidades 3.2 (II) a 3.4.
8	**H1:** Unidad 4. **H2:** Unidad 5.1 **H3**: Laboratorio: introducción a Octave
9	**H1:** Unidad 5.2 y 5.3 **H2:** Laboratorio: Modelos de regresión lineal **H3:** Laboratorio
10	**H1:** Unidad 6.1

Semana	Contenido
	H2: Laboratorio: Algoritmo Page Rank **H3:** Laboratorio
11	**H1:** Unidad 6.2 y 6.3 **H2:** Laboratorio **H3:** Laboratorio
12	**H1:** Unidad 6.4 **H2:** Laboratorio **H3:** Laboratorio
13	**H1:** Unit 7 **H3:** Laboratory. Definition of a cube in an Analysis Services project within SQL Server 2012.
14	**H1:** Unit 8.1 and 8.2 **H2:** Laboratory. Business Intelligence with PowerPivot **H3:** Unit 8.3