Subject: Numerical and Data-Intensive Computing (COMP)
Code: 32416
Institution: Escuela Politécnica Superior
Degree: Master's program in Research and Innovation in Information and Communications Technologies ($i^2$-ICT)
Level: Master
Type: Core
ECTS: 6

# COURSE GUIDE: Numerical and Data-Intensive Computing (COMP)

**Academic year:** 2015-2016

**Program:** Master's program in Research and Innovation in Information and Communications Technologies ($i^2$-ICT)
**Center:** Escuela Politécnica Superior
**University:** Universidad Autónoma de Madrid

**Last modified:** 2015/04/26
**Status:** Approved June 8th 2015.

# 1. ASIGNATURA / COURSE (ID)

## Cálculo intensivo y manejo de datos a gran escala
## Numerical and Data-Intensive Computing (COMP)

## 1.1. Programa / program

Máster Universitario en Investigación e Innovación en Tecnologías de la Información y las Comunicaciones ($i^2$-TIC)

Master in Research and Innovation in Information and Communications Technologies ($i^2$-ICT) [Officially certified]

## 1.2. Course code

32416

## 1.3. Course areas

Computer Science and Artificial Intelligence

## 1.4. Tipo de asignatura / Course type

Obligatoria [itinerario: todos los itinerarios]
Core [itinerary: all itineraries]

## 1.5. Semester

First semester

## 1.6. Credits

6 ETCS

## 1.7. Language of instruction

The lecture notes are in English. The lectures are mostly in Spanish. Some of the lectures and seminars can be in English.

## 1.8. Recommendations / Related subjects

Knowledge of the C programming language, Data Bases, probability and statistics at an introductory level is useful to follow the course.

Related subjects are:

- Procesamiento de información temporal [Temporal Information Processing]
- Aprendizaje Automático: teoría y aplicaciones [Machine Learning: Theory and Applications]
- Métodos bayesianos aplicados [Applied Bayesian Methods]
- Aceleración de algoritmos en sistemas heterogéneos [Algorithm Acceleration in Heterogeneous Systems]
- Procesamiento de señales biomédicas y sus aplicaciones [Biomedical Signal Processing and its Applications]
- Procesamiento de audio y voz para biometría y seguridad [Speech and Audio Processing for Biometrics and Security]
- Técnicas de análisis de secuencias vídeo para videovigilancia [Techniques of Analysis of Video Sequences for Surveillance]

## 1.9. Lecturers

Add @uam.es to all email addresses below.

**Lectures and labs:**

**Dr. Carlos Santa Cruz Fernández** (Coordinator)
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 / Course objectives

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.

This subject is divided into three parts. The first part corresponds to an introduction to efficient programming techniques for shared memory parallel architectures, including multi-core processors and tightly-coupled multiprocessors. In particular, it aims at describing performance analysis tools (profilers), identifying the main factors that affect the efficiency of parallel architectures, and studying parallelization techniques for loops and sequential programs on shared-memory parallel architectures through OpenMP.
The second part is aimed at introducing basic numerical algorithms and tools for matrix manipulation. Solution of Linear Algebraic Equations and the eigensystem problem are described. These algorithms are used to solve General Linear Regression Models and Principal Component Analysis (PCA) for dimensionality reduction. Octave is used to implement and solve the proposed problems.
Finally, the third part provides the engineering framework, the techniques and the tools needed to deliver data warehouses which involve data preprocessing, data integration, and providing on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data. An additional goal is to understand what business intelligence is, how it works, where it is used, and why and when to use it. Existing BI products and tools are also described and analyzed.

At the end of each unit, the student should be able to:

| **UNIT BY UNIT SPECIFIC OBJECTIVES** | |
|---|---|
| **PART I** | |
| **UNIT 1.- Introduction** | |
| **1.1.** | Know the different families of parallel architectures and choose the one that best suits a specific application scope. |
| **1.2.** | Measure the performance of parallel algorithms in terms of speedup and efficiency. |
| **1.3.** | Understand the basic concepts behind parallel programming, including tasks, processes and synchronization mechanisms. |
| **1.4.** | Use performance analysis tools (profilers) to analyze the efficiency of sequential algorithms and identify portions susceptible to be accelerated through parallelization. |
| **1.5.** | Implement simple parallel algorithms on shared-memory parallel architectures using C and OpenMP. |
| **UNIT 2.-** Loop parallelization on shared-memory parallel architectures | |
| **2.1.** | Understand the different types of parallel loops and choose the ones that best suit a specific problem. |
| **2.2.** | Describe the dependencies between iterations of a set of sequential loops through a dependency graph. |
| **2.3.** | Parallelize a set of sequential loops from their corresponding dependency graph. |
| **2.4.** | Apply code transformations in order to optimize the parallelization of a set of sequential loops. |
| **UNIT 3.-** General process for parallelization of sequential programs | |
| **3.1.** | Decompose sequential algorithms into tasks susceptible to be parallelized. |
| **3.2.** | Maximize load balancing in the assignment of tasks to processes. |
| **3.3.** | Minimize communication costs in the assignment of tasks to processes. |
| **3.4.** | Minimize synchronization and management overheads in the assignment of tasks to processes. |
| **3.5.** | Determine the scheduling of processes that maximizes efficiency. |
| **3.6.** | Determine the mapping of processes to compute units that maximizes efficiency. |
| **PART II** | |
| **UNIT 4.-** Introduction | |
| **4.1.** | Understand the floating-point data representation in modern processors |
| **4.2.** | Understand that the round-off error is due to the fact that arithmetic among numbers is not exact |
| **4.3.** | Understand that the errors due to the algorithm used are independent of the hardware |
| **4.4.** | Understand that errors can be successively magnified due to unstable algorithms |
| **UNIT 5.-** Solution of Linear Algebraic Equations | |
| **5.1.** | Use the LU matrix decomposition to solve linear system equations |
| **5.2.** | Use the Cholesky decomposition to solve linear system equations |

| | |
|---|---|
| **5.3.** | Write a Multiple Linear Regression problem in matrix form |
| **5.4.** | Solve some practical examples |
| **UNIT 6.- Eigensystems** | |
| **6.1.** | Obtain the main eigenvalue/eignevector |
| **6.2.** | Understand the Google's Page Rank Algorithm |
| **6.3.** | Calculate eigenvalues/Eigenvectors for a symmetric matrix |
| **6.4** | Obtain uncorrelated variables using the PCA algorithm |
| **PART III** | |
| **UNIT 7.-** Introduction to Data Warehousing | |
| **7.1.** | Explain the goals of data warehousing |
| **7.2.** | Data extraction, transformation, loading techniques for data warehousing. |
| **7.3.** | Explain accepted data warehouse terminology |
| **7.4.** | Describe methods and tools for extracting, transforming, and loading data |
| **7.5.** | Identify some of the tools for accessing and analyzing warehouse data |
| **UNIT 8.-** Business Intelligence | |
| **8.1.** | Use the terminology and concepts in business intelligence |
| **8.2.** | Explain how Business Intelligence systems work, their strength and weaknesses. |
| **8.3.** | Exploit business analytics and performance measurement tools |
| **8.4.** | Analyze emerging trends and developing BI tools |

## 1.11. Course contents

**PART I**

1. **Introduction**
   - 1.1. Parallel architectures: motivation
   - 1.2. Shared-memory parallel architectures
     - 1.2.1. Multi-core processors and tightly-coupled multiprocessors
     - 1.2.2. OpenMP
   - 1.3. Distributed-memory parallel architectures: loosely-coupled multiprocessors
   - 1.4. Grid computing and cloud computing
   - 1.5. Basic concepts
     - 1.5.1. Tasks and processes
     - 1.5.2. Semaphores and barriers
     - 1.5.3. Performance analysis tools (profilers)

2. **Loop parallelization on shared-memory parallel architectures**
   - 2.1. Parallel loops
   - 2.2. Scheduling of parallel loops
   - 2.3. Analysis of dependencies between iterations
     - 2.3.1. True dependencies
     - 2.3.2. Anti-dependencies
     - 2.3.3. Output dependencies
     - 2.3.4. Dependencies in nested loops
     - 2.3.5. Dependency graphs by levels
   - 2.4. Generation of parallel code
     - 2.4.1. Strongly connected components
     - 2.4.2. Acyclic condensation
     - 2.4.3. Barrier-free arcs
     - 2.4.4. Generation of clusters and segments
     - 2.4.5. Code generation
       - 2.4.5.1. Code generation for serial segments
       - 2.4.5.2. Code generation for parallel segments
     - 2.4.6. Example problems
   - 2.5. Transformations to support parallelization
     - 2.5.1. Loop normalization
     - 2.5.2. Scalar substitution
     - 2.5.3. Scalar expansion
     - 2.5.4. Variable copying
     - 2.5.5. Loop interchange

3. **General process for parallelization of sequential programs**
   - 3.1. Decomposition
   - 3.2. Assignment
     - 3.2.1. Load balancing
     - 3.2.2. Communication reduction
     - 3.2.3. Overhead reduction

3.3. Orchestration
3.4. Mapping

**PART II**

4. **Introduction**
   4.1. Floating-Point Representation
   4.2. Roundoff Error
   4.3. Truncation Error
   4.4. Stability

5. **Solution of Linear Algebraic Equations**
   5.1. LU Decomposition
   5.2. Cholesky decomposition
   5.3. General Linear Regression Model in matrix term
   5.4. Practical applications

6. **Eigensystems**
   6.1. Power method
   6.2. Application to the Page Rank algorithm
   6.3. Jacobi transformation for a symmetric matrix
   6.4. Principal Component Analysis and its applications to dimensionality reduction and lineal models

**PART III**

7. **Introduction to Data Warehousing**
   7.1. Decision support systems
   7.2. Data warehousing
   7.3. Data warehouse architecture
   7.4. ETL (extraction, cleansing, transformation and loading)
   7.5. Multidimensional model
   7.6. Meta-data
   7.7. Accessing data warehouses
   7.8. Additional issues: security, quality …

8. **Business Intelligence**
   8.1. Introduction and basics
   8.2. Business Intelligence user models
   8.3. BI Products and Vendors
      8.3.1. Enterprise Business Intelligence products
      8.3.2. Database and Packaged Products
      8.3.3. Data Discovery & Visualization

## 1.12. Course bibliography

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. Coursework and evaluation

The course involves lectures, weekly assignments, lab assignments, a seminar presentation and one exam.

In the ordinary exam period it is necessary to have a pass grade (≥ 5) in the exam to pass the course. In the extraordinary exam period it is necessary to have a pass grade (≥ 5) in the report on a related research topic to pass the course.

- In the ordinary exam period, the evaluation will be made according to the following scheme
    - 50 % Lab assignments
    - 50 % Exam [end of term]

  The grades of the individual parts are kept for the extraordinary exam period.

- In case of a fail grade in the ordinary exam period, in the extraordinary exam period, the student has the opportunity to
    - Turn in all the lab assignments with corrections
    - Turn in a report on a research topic about numerical and data-intensive computing.

  The grade will be determined by
    - 50 % Lab assignments [only if the lab assignments are turned in]
    - 50 % Report on a related research topic [only if the report is turned in]

## 1.14. Timeline

| Week | Content |
|---|---|
| 1 | **H1:** Course presentation.<br>**H2:** Units 1.1 to 1.3.<br>**H3:** Units 1.4 and 1.5. |
| 2 | **H1:** Laboratory 1: Profilers (perf and gprof).<br>**H2:** Laboratory 1: OpenMP (compilation and basic directives).<br>**H3:** Units 2.1 and 2.2. |
| 3 | **H1:** Units 2.3.1 to 2.3.3.<br>**H2:** Units 2.3.4 and 2.3.5.<br>**H3:** Units 2.4.1 to 2.4.3. |
| 4 | **H1:** Laboratory 2: *OpenMP* (Parallel loops)<br>**H2:** Laboratory 2: *OpenMP* (Parallel loops)<br>**H3:** Unit 2.4.4. |
| 5 | **H1:** Unit 2.4.5.<br>**H2:** Unit 2.4.6 (I).<br>**H3:** Unit 2.4.6 (II). |
| 6 | **H1:** Laboratory 3: Loop parallelization with *OpenMP*.<br>**H2:** Laboratory 3: Loop parallelization with *OpenMP*.<br>**H3:** Unit 2.5 (I). |
| 7 | **H1:** Units 2.5 (II) and 3.1.<br>**H2:** Unit 3.2 (I).<br>**H3:** Units 3.2 (II) to 3.4. |
| 8 | **H1:** Unit 4.<br>**H2:** Unit 5.1<br>**H3**: Laboratory: introduction to Octave |
| 9 | **H1:** Units 5.2 and 5.3<br>**H2:** Laboratory: Linear regression models<br>**H3:** Laboratory |
| 10 | **H1:** Unit 6.1 |

| Week | Content |
|---|---|
| | **H2:** Laboratory: Page Rank algorithm<br>**H3:** Laboratory |
| 11 | **H1:** Unit 6.2 and 6.3<br>**H2:** Laboratory<br>**H3:** Laboratory |
| 12 | **H1:** Unit 6.4<br>**H2:** Laboratory<br>**H3:** Laboratory |
| 13 | **H1:** Unit 7<br>**H3:** Laboratory. Definition of a cube in an Analysis Services project within SQL Server 2012. |
| 14 | **H1:** Unit 8.1 and 8.2<br>**H2:** Laboratory. Business Intelligence with PowerPivot<br>**H3:** Unit 8.3 |