Course guide
270218 - PSD - Parallelism and Distributed Systems

Last modified: 30/01/2024

Unit in charge: Barcelona School of Informatics
Teaching unit: 701 - DAC - Department of Computer Architecture.

Degree: BACHELOR'S DEGREE IN DATA SCIENCE AND ENGINEERING (Syllabus 2017). (Compulsory subject).

Academic year: 2023 **ECTS Credits:** 6.0 **Languages:** Catalan

LECTURER

Coordinating lecturer: JULITA CORBALAN GONZALEZ

Others: Segon quadrimestre:
YOLANDA BECERRA FONTAL - 11, 12, 13
JULITA CORBALAN GONZALEZ - 11, 12, 13

PRIOR SKILLS

C and Python are the programming language of choice for the labs sessions of this course. It is assumed that the student has a basic knowledge of Python and C prior to starting classes.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.

Generical:
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

Transversal:
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.
CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.

Basic:
CB1. That students have demonstrated to possess and understand knowledge in an area of ??study that starts from the base of general secondary education, and is usually found at a level that, although supported by advanced textbooks, also includes some aspects that imply Knowledge from the vanguard of their field of study.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

TEACHING METHODOLOGY

During the course there will be four types of activities:

a) Activities aimed at acquiring theoretical knowledge. Theoretical activities include participatory classes, which explain the basic contents of the course.

b) The activities focused on acquiring knowledge through experimentation using the "learn to do" approach in practice-guided laboratory sessions (and final report). Some sessions may include pre-work or post-session work depending on
the use of laboratories.

c) Few sessions during the theory classes where practical exercises will be performed to perform numerical evaluations and analysis for performance evaluation.

d) Two reports of exercises to be performed in laboratories related to HPC environments and applications and data analysis environments

This semester, as lab classes will be held in theory classrooms, students will be required to bring their own laptop. To take the exams, both theoretical and laboratory, because they will be delivered in digital format, you will also need to bring your own laptop. All the theory classes that are done online will meet a meet in the official schedule. For laboratory classes, those students who are confined will be able to meet to be able to follow the classes.

LEARNING OBJECTIVES OF THE SUBJECT

1.Conèixer els fonaments dels sistemes paral·lels i distribuïts actuals
2.Coneixer i saber usar els elements bàsics que conformen els sistemes paral·lels i distribuïts
3.Coneixer i poder triar convenientment quin els entorns d'analítica avançada que usen sistemes distribuïts i parallel
4.Us pràctic per diferents problemes plantejats dels entorns cloud, sistemes paral.lels i distribuïts disponibles actualment per a un enginyer i científic de dades
5.Familiaritzar-se amb els models de programació més habituals dels sistemes paral·lels i distribuïts

STUDY LOAD

Type	Hours	Percentage
Hours large group	30,0	20.00
Self study	90,0	60.00
Hours small group	30,0	20.00

Total learning time: 150 h

CONTENTS

Foundations of parallel and distributed supercomputing

Description:
In this topic, students will learn basic concepts of parallel computing as well as metrics that will help them evaluate both the performance of their programs and the limits derived from the application structure itself.

Parallel and distributed architectures

Description:
In this topic, students will learn the main characteristics of the parallel and distributed architectures that can most influence them when designing their data analysis programs or to understand the performance (or loss of performance) of them.

Execution environments for parallel computing and data analytics

Description:
In this topic, students will learn about the different environments that can be found mainly when executing so many applications to generate data such as those stored or analyzed. Emphasis will be placed on the differences between the three environments and their impact on the efficiency of their applications.

Programming models for supercomputers

Description:
In this topic the students will see the basic principles of the most used programming models in the HPC environments: MPI, OpenMP and hybrid MPI OpenMP models. The tools will be given to detect and manage the main details that may affect both the robustness of their programs and their efficiency.
Co-processor oriented models that offer good performance vs. efficiency will also be introduced. Energy consumption and very used in the analysis of data.

Software and execution environment specific for advanced analytics

Description:
In this topic the students will see in more detail the characteristics of the programming models and execution environments for storage and data analysis. The Apache Spark / Hadoop model will be used as a reference, as a reference for Cassandra data storage and as TensorFlow / keras analysis tools.

Powering Machine Learning with supercomputers: Case Study with Spark/Cassandra/TensorFlow

Description:
In this subject, you will learn in a machine learning environment using the Apache Spark model, with DB key / value Cassandra i com aina d'anàlisi TensorFlow. S'explicaran the elements més importants d'aquests three components that can affect in greater measure to the design of applications of machine learning with l'emmagatzematge de dades i anàlisi.

Lab sessions

Description:
The laboratory sessions will be grouped into two projects that will be carried out both in the laboratory sessions and in autonomous work. The two projects will be related to the programming, analysis and optimization of a case as realistic as possible in two environments: parallel execution environments (mpi OpenMP, queue systems, etc.) used to generate and post-process data , and specific management and data analysis environments such as Apache Stark Cassandra TensorFlow.

ACTIVITIES

Course introduction

Description:
During this activity, the objectives, contents, and operation of the subject will be explained

Full-or-part-time: 1h
Theory classes: 1h

Development of the theme "Fundamentals of parallel and distributed supercomputing"

Description:
In this topic, students will learn basic concepts of parallel computing as well as metrics that will help them assess both the performance of their programs and the limits derived from the structure of the application.

Specific objectives:
1

Related competencies :
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CB1. That students have demonstrated to possess and understand knowledge in an area of ??study that starts from the base of general secondary education, and is usually found at a level that, although supported by advanced textbooks, also includes some aspects that imply Knowledge from the vanguard of their field of study.

Full-or-part-time: 4h
Theory classes: 2h
Self study: 2h

Development of the theme "Parallel and Distributed Architectures"

Description:
In this topic, students learn the main features of parallel and distributed architectures that can influence the design of their data analysis programs and understand the performance (or loss of performance) of these: They will be seen , for example features of systems with multi-core architecture, hyperthreading, shared-distributed memory, local time-space data, type of storage (local, remote), typology networks, etc.

Specific objectives:
1, 2

Related competencies :
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB1. That students have demonstrated to possess and understand knowledge in an area of ??study that starts from the base of general secondary education, and is usually found at a level that, although supported by advanced textbooks, also includes some aspects that imply Knowledge from the vanguard of their field of study.

Full-or-part-time: 4h
Theory classes: 2h
Self study: 2h

Development of the theme "Execution environments for parallel computation and data analysis"

Description:
In this topic, students will learn about the different environments that can be found mainly when executing so many applications to generate data such as those stored or analyzed. Emphasis will be placed on the differences between the three environments and their impact on the efficiency of their applications. Running environment with queues for HPC, cloud computing for DA. During this topic, it will be divided into HPC environments and data analysis environments (DAs). Problems will also be exercised during theory classes.

Specific objectives:
2, 3

Related competencies :
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

Full-or-part-time: 12h
Theory classes: 6h
Self study: 6h

Development of the subject "Models of programming for supercomputers"

Description:
In this topic, students will see the basic principles of the most used programming models in the HPC environments: MPI, OpenMP and MPI + OpenMP hybrid models. The tools will be provided to detect and manage the main details that can affect both the robustness of their programs and their efficiency. Coprocessor-oriented models that offer good performance vs. efficiency Energy consumption and much used in the analysis of data.

Specific objectives:
5

Related competencies :
CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.
CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 12h
Theory classes: 6h
Self study: 6h

Development of the subject "New software for data analysis"

Description:
In this topic, the students will see in more detail the characteristics of the programming models and execution environments for the storage and the analysis of data. The Apache Spark / Hadoop model will be used as a reference, as a reference for the Cassandra data storage and as TensorFlow / keras analysis tools.

Specific objectives:
3

Related competencies :
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.

Full-or-part-time: 14h
Theory classes: 7h
Self study: 7h

Development of the subject "Machine Learning in Supercomputers: Case Based on Spark / Cassandra / TensorFlow"

Description:
In this topic we will study in a Machine Learning environment using the Apache Spark model, such as DB key / value Cassandra and TensorFlow analysis tool. The most important elements of these three components will be explained, which can affect, in greater measure, the design of machine learning applications as well as the storage of data and analysis.

Specific objectives:
4

Related competencies :
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

Full-or-part-time: 8h
Theory classes: 4h
Self study: 4h

Laboratory sessions and deliverables: Application execution in HPC environments, Data generation in HPC environments, data storage and data analysis in context of DA (Data Analytics)

Description:
During the lab exercises will be proposed that will be done most during the classes. Some of these exercises will aim at practicing specific aspects of both more traditional HPC and data analytics environments. Others will be part of a larger exercise over the course of sessions. There will be two exercises: one for the most HPC and one more specific for data analysis environments. It will first be delivered just after the end of sessions dedicated to HPC environments and applications. The second just after the data analysis environment sessions are over.

Specific objectives:
2, 3, 4, 5

Related competencies :
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.
CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.

Full-or-part-time: 56h
Laboratory classes: 28h
Self study: 28h

GRADING SYSTEM

The evaluation of the subject will come out of three components:

- Partial exam: 35% (First half of the course)
- Final exam: 35% (Second half of the course)
- Laboratory: 30%.

The lab grade will come from the evaluation of the lab deliverables
The final grade is compted: 0.3*lab+0.35*final exam + 0.35*Partial exam

In case the grade is less than 5.0, student will be allowed to do the reevaluation exam. In that case, the grade will be computed as:

Final Note: Max(Revaluation Exam * 0.7, Partial exam * 0.35 + Final Exam * 0.35) + Laboratory * 0.3

BIBLIOGRAPHY

Basic:
- TORRES, Jordi. Hand-on sessions at GitHub.
- Torres, J. Slides of the course. UPC,
- Torres, J. Understanding supercomputing: with Marenostrum Supercomputer in Barcelona. Universitat Politècnica de Catalunya, Barcelona Supercomputing Center, 2016. ISBN 9781365376825.
- Torres, J. Hello world en TensorFlow. Universitat Politècnica de Catalunya, Barcelona Supercomputing Centrer, 2016. ISBN 9781326532383.
- Macias, M.; Gómez, M.; Tous, R.; Torres, J. Introducción a Apache Spark: para empezar a programar el big data. UOC, 2015. ISBN 9788491160373.
- Articles from Technical Journals in the area.

Complementary:
- Torres, J. Empresas en la nube: ventajas y retos del cloud computing. Libros de Cabecera, 2011. ISBN 9788493908225.