270218 - PSD - Parallelism and Distributed Systems

Coordinating unit: 270 - FIB - Barcelona School of Informatics
Teaching unit: 701 - AC - Department of Computer Architecture
Academic year: 2018
Degree: BACHELOR'S DEGREE IN DATA SCIENCE AND ENGINEERING (Syllabus 2017). (Teaching unit Compulsory)
ECTS credits: 6
Teaching languages: Catalan

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.

General:
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.

Teaching methodology
During the course there will be four types of activities:

a) Activities focused on the acquisition of theoretical knowledge. The theoretical activities include participatory lecture classes, which explain the basic contents of the course.

b) Activities focused on the acquisition of knowledge through experimentation by "learn by doing" approach in lab sessions guided by hands-on (and final report). Prior to the lab sessions the student will prepare a related piece of

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

d) Course project that will be based on technologies considered in this course

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

<table>
<thead>
<tr>
<th>Study load</th>
<th>Total learning time: 150h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory classes:</td>
<td>60h 40.00%</td>
</tr>
<tr>
<td>Guided activities:</td>
<td>6h  4.00%</td>
</tr>
<tr>
<td>Self study:</td>
<td>84h 56.00%</td>
</tr>
</tbody>
</table>
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.

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.

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.

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.

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**

**Degree competences to which the content contributes:**

**Description:**
In this subject, you will learn in a machine learning environment using the Apache Spark model, with DB key / value Cassandra and analysis TensorFlow. The most important elements of these three components that can affect in greater measure to the design of applications of machine learning with the storage of data and analysis.

**Lab sessions**

**Degree competences to which the content contributes:**

**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.
# Planning of activities

## Course introduction

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

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

| Hours: | 4h |
| Theory classes: | 2h |
| Practical classes: | 0h |
| Laboratory classes: | 0h |
| Guided activities: | 0h |
| Self study: | 2h |

| 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: |

### Specific objectives:

1. |

## Development of the theme "Parallel and Distributed Architectures"

| Hours: | 4h |
| Theory classes: | 2h |
| Practical classes: | 0h |
| Laboratory classes: | 0h |
| Guided activities: | 0h |
| Self study: | 2h |

| 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: |

### Specific objectives:

1, 2 |

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

| Hours: | 12h |
| Theory classes: | 6h |
| Practical classes: | 0h |
| Laboratory classes: | 0h |
| Guided activities: | 0h |
| Self study: | 6h |

| Description: |  |
### Development of the subject "Models of programming for supercomputers"

**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, 4

**Hours:** 12h  
- Theory classes: 6h  
- Practical classes: 0h  
- Laboratory classes: 0h  
- Guided activities: 0h  
- Self study: 6h

### Development of the subject "New software for data analysis"

**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:**
3

**Hours:** 14h  
- Theory classes: 7h  
- Practical classes: 0h  
- Laboratory classes: 0h  
- Guided activities: 0h  
- Self study: 7h

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

**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:**
4

**Hours:** 8h  
- Theory classes: 4h  
- Practical classes: 0h  
- Laboratory classes: 0h  
- Guided activities: 0h  
- Self study: 4h
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:
5

Laboratory project: Data generation in HPC environments, data storage and data analysis in context of DA (Data Analytics)

Hours: 56h
Theory classes: 0h
Practical classes: 0h
Laboratory classes: 28h
Guided activities: 0h
Self study: 28h

Description:
This project will follow the natural flow of data, from the parallelization, execution, and evaluation of codes that generate data, traditionally in an HPC environment, through storage and subsequent analysis with the new execution environments for this kind of problems.

Specific objectives:
2, 3, 4, 5

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

- Partial exam: 25%
- Final exam: 65%
- Laboratory work: 10%.

Final Note: Max(Revaluation Exam * 0.9, Partial exam * 0.25 + Final Exam * 0.65) + Laboratory Work * 0.1
Bibliography

Basic:

TORRES, Jordi. Hand-on sessions at GitHub.
TORRES, Jordi. Slides of the course.  UPC.

Complementary: