UE Large scale Data Management and Distributed Systems

Level
Baccalaureate +5

ECTS
6 credits

Component
UFR IM2AG (informatique, mathématiques et mathématiques appliquées)

- **Teaching language(s):** English
- **Open to exchange students:** Yes
- **Code d'export Apogée:** GBX9MO72

Presentation

Description

The course is divided in two complementary parts: distributed systems and data management.

Part 1: Distributed systems

Summary

Distributed systems are omnipresent. They are formed by a set of computing devices, interconnected by a network, that collaborate to perform a task. Distributed systems execute on a wide range of infrastructures: from Cloud datacenters to wireless sensor networks.

The goal of this course is to study the main algorithms used at the core of Distributed systems. These algorithms must be efficient and robust. An algorithm is efficient when it sustains a high level of performance. Performance can be measured using various metrics such as throughput, latency, response time. An algorithm is robust when it is able to operate despite the occurrence of various types of (network and/or machine) and attacks.

Content

During the course, we will cover several topics that are listed below:

- Event-driven formalisms for distributed algorithms
- Basic abstractions: processes, links
- Failure detector algorithms
- Leader election algorithms
- Broadcasting algorithms
- Distributed shared memory algorithms
- Consensus algorithms
- Epidemic algorithms
- Performance models for distributed systems

Part 2: Data management

Summary

The ability to process large amounts of data is key to both industry and research today. As computing systems are getting larger, they generate more data that need to be analyzed to extract knowledge.

Data management infrastructures are growing fast, leading to the creation of large data centers and federations of data centers. Suitable software infrastructures should be used to store and process data in this context. Big Data software systems are built to take advantage of a large set of distributed resources to efficiently process massive amounts of data while being able to cope with failures that are frequent at such a scale.

In addition to the amount of data to be processed, the other main challenge that such Big Data systems need to deal with is time. For some use cases, the earlier the results of a data analysis is obtained, the more valuable the result is. Some Big Data systems especially target stream processing where data are processed in real time.

Through lectures and practical sessions, this course provides an overview of the software systems that are used to store and process data at large scale. The following topics will be covered:

- Map-Reduce programming model
- In-memory data processing
- Stream processing (data movement and processing)
- Large scale distributed data storage (distributed file systems, NoSQL databases)

Throughout the lectures, the challenges associated with performance and fault tolerance will also be discussed.

Course parts

Lectures	Lectures (CM)	30h
Practical work	Practical work (TP)	6h

Recommended prerequisites

Fundamentals of DBMS, parallel programming (threads)

Period : Semester 9

Bibliography

- Guerraoui, Rachid and Rodrigues, Luis. Introduction to Reliable Distributed Programming. Springer. ISBN 978-3-642-15260-3.
- Lynch, Nancy. Distributed Algorithms. Morgan Faufmann Publishers Inc. ISBN 978-0-08-050470-4
- Kleppmann, Martin. Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems. " O'Reilly Media, Inc.", 2017.

Useful info

Contacts

Program director

Thomas Ropars

Thomas.Ropars@univ-grenoble-alpes.fr

Campus

Grenoble - University campus