COPERNICUS MASTER IN DIGITAL EARTH

GeoData Science Track

HPC FOR BIG DATA

Frédéric RAIMBAULT, Nicolas COURTY

University of South Brittany, France

IRISA laboratory, OBELIX team

Overview

- **The problem: Big Data**
  - Data is growing exponentially
    - Hardware performance doesn’t scale as quickly
  - How to get business value out of this data
- **The (computer scientist) solution: HPC**
  - High Performance Computing
    - Parallel computers (hardware)
      - Cluster of machines with multiple processors containing multiple cores, GPU farms...
    - Distributed systems (software)
      - For storage (distributed FS) and processing (programming frameworks)

Course Objectives

1) **Understand some issues**
   - Complexity of managing datasets due to their Volume, Velocity and Variety (3V)
   - Distributed infrastructure is difficult to use
2) **Become familiar with existing tools**
   - Hadoop technology stack
   - GPGPU Cuda framework
3) **Practice on real (big) data**
   - AIS signals dataset from AIS-UBS
   - The Copernicus datasets

Programming Languages and Tools

- **Python**
  - Version > 3.
  - Libraries : snakebite, mrjob, PySpark, Numba,...
  - Programming Environment : Anaconda, Spyder, Jupyter Notebook
- **Shell**
  - Command Line Interface (text) to read commands and run programs (and connect to remote machines)
- **AWS EMR**
  - MapReduce tools from the Amazon Cloud computing resources

Resources

- **Computing environment**
  - Locally host server of 40 cores + 1 Titan-X GPU: `dmis`
  - Virtual/cloud servers hosted in Amazon EC2
  - Remote access using “ssh” (a remote CLI)
  - OS (Operating system) : Linux (a Unix alternative)
- **Datasets**
  - AIS-UBS: https://ais.univ-ubs.fr
  - Copernicus: www.copernicus.eu/en/access-data/conventional-data-access-hubs
- **Online material**
  - ENT “Big GeoData”, enrollment key: CDE
    https://moodle.univ-ubs.fr/course/view.php?id=5920

Evaluation

- **Several tests on theoretical and practical exercises during the semester.**

Lecture Plan

1) Introduction (1 session)

- Big Data issues and HPC principles

2) GPU-based processing (3 sessions)

3) Hadoop Stack (6 sessions)

- HDFS
- MapReduce
- Spark