Introduction to Big Data Processing Infrastructures

Introduction

- This course is not about Big Data, it is about infrastructures
- We should have visited a datacenter, but… Corona
- We need SSH, a little bit of C/C++, vim, Python
- We will use AWS and Google Cloud computing!
- Of course we will make a project and an oral exam
- Prof of course is a physicist of the IIFN
- The IIFN-CNAF hosts the Tier-1 datacenter in Bologna

Setting up AWS

- It is a pay for use model
- You have at the beginning a 50$ credit

Big data

- They tend to be non structured
- They are characterized by 4 Vs: Volume, Variety, Veracity (can I trust them?), Velocity
- Another important point is Value, what I want to extract from the data

Computational challenge

- Find a substring in a string
- Doing it brute force is really slow
- I can create an index once and then I can search every time much faster
  - It is blast essentially
  - The BWA algorithm is another possibility
    - It is used for more similar sequences
    - It is faster when I have many reads to be aligned
- Using the right approach is much more effective than increasing computational power
- If possible use open source code
  - You avoid vendor lockin and the approach can be scaled more easily
- Creating a new approach from scratch is usually wrong
  - You will not create a good system from scratch
- Our challenge: align 554k sequences per patient against the human genome
  - The aim of this course is creating a cloud based model for computing this
  - I also want to estimate the time required

Some notes

- A checksum is essential when we are moving data
  - It is a small string used to detect error in data transmission or storage
  - It is possible to set an extended file attribute with the checksum (if the file system allows it)
- Files are moved compressed, moving uncompressed files is a crime