



SAPIENZA
UNIVERSITÀ DI ROMA

Faculty of Information Engineering, Informatics and
Statistics
Department of Computer Science

Big Data Computing

Author:
Simone Lidonnici

26 september 2025

Contents

1	Introduction	1
1.1	Data center	1
1.1.1	Reliability and Programmability problems	2

1

Introduction

With **Big Data** we refer to an actual phenomenon, defined by five properties called **5V**:

- **Value**: extracting knowledge from data is valuable.
- **Volume**: large amount of data.
- **Variety**: different format of data.
- **Velocity**: the speed at which the data are generated is very high.
- **Veracity**: reliability of the data used.

To do a computation on a lot of data at a high speed (like a google search), there are different problems that occurs:

- Disks are not large enough.
- Disks are not fast enough.
- CPUs are not fast enough.

1.1 Data center

To do this type of computations **data center** are used, which can upgrade their performances in two way:

- **Scale up**: buying new and more powerful components. This is a problem because the increase in the velocity required is faster then the performance increase.
- **Scale out**: buying more components and interconnect them to work in parallel.

A data center is composed by a series of rack interconnected between them with a private network. Every rack has servers inside them that contain CPUs, GPUs and memory. A server is considered a node in the network.

The major problem with the data centers is the bottleneck caused by the network. The network can't send all the data required for the computation in a fast way and this slow down the computation. Also, the increase in GPU power (that does the major part of computation in a data center) is much faster than the increase in network speed during the years.

1.1.1 Reliability and Programmability problems

Another problem is related to **reliability**, so the probability that a server, disk or network component fails. Even if the probability of a single component is very low (one fail every 10 years) a large data center with thousand of components will have failures very frequently. This mean that checkpoint must be used, and every failure reset to the last one.

Last problem is the **programmability**, so how to write an efficient parallel program being aware of all the components. Things that have to be taken in account are how to store data, how to send data efficiently, what to execute on CPU or GPU and many other.