Concepts and Models of Parallel and Data-centric Programming

MapReduce – Introduction

Lecture, Summer 2020

Simon Schwitanski
Dr. Christian Terboven

i12 High Performance Computing
RWTH AACHEN UNIVERSITY

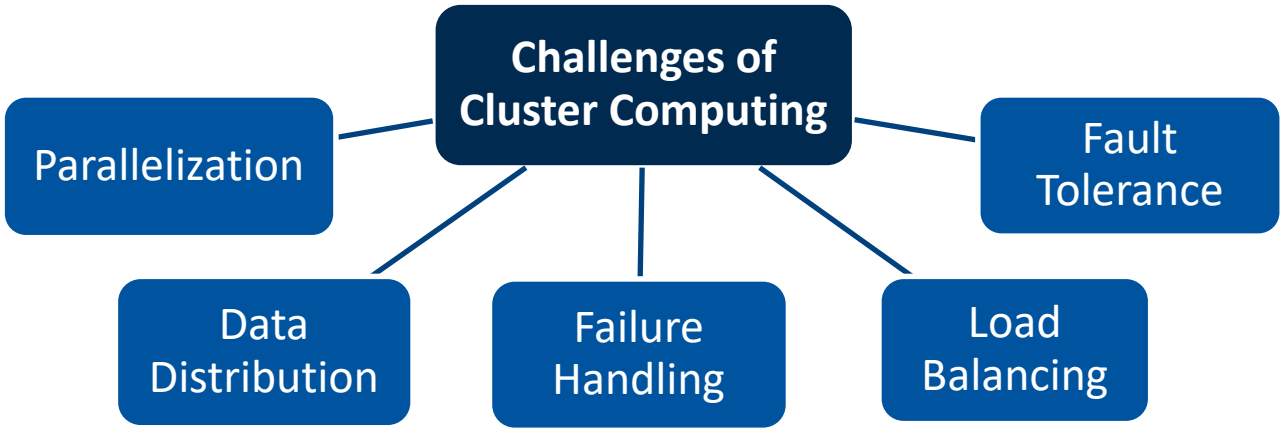
Outline

0. Organization
1. Foundations
2. Shared Memory
3. GPU Programming
4. Bulk-Synchronous Parallelism
5. Message Passing
6. Distributed Shared Memory
7. Parallel Algorithms
8. Parallel I/O
9. **MapReduce**
10. Apache Spark

a. MapReduce Programming Model
b. Parallelizing MapReduce
c. Hadoop Ecosystem
d. Hadoop Distributed File System
e. Yet Another Resource Negotiator
f. Comparison to Other Approaches
g. MapReduce Design Patterns

Motivation

- **“Big Data”**: Processing large input data (data-intensive computing)
  - Data does not fit into main memory of single machine or small cluster
  - Distribution of computation across **huge** number of machines needed
- **Problem:** Cluster computing is complex



- **Idea:** Hide those challenges from developer by using an abstraction
- **Approach:** MapReduce programming model and framework

MapReduce – Data-centric Programming

- Data-centric programming: Focus on managing and transforming data
  - Aggregation / Summarization
  - Filtering
  - …
- Processing information (“working with data”) is primary design goal of corresponding models and languages
- Data-centric programming model: MapReduce
- Data-centric programming language: SQL

MapReduce – Application Domains

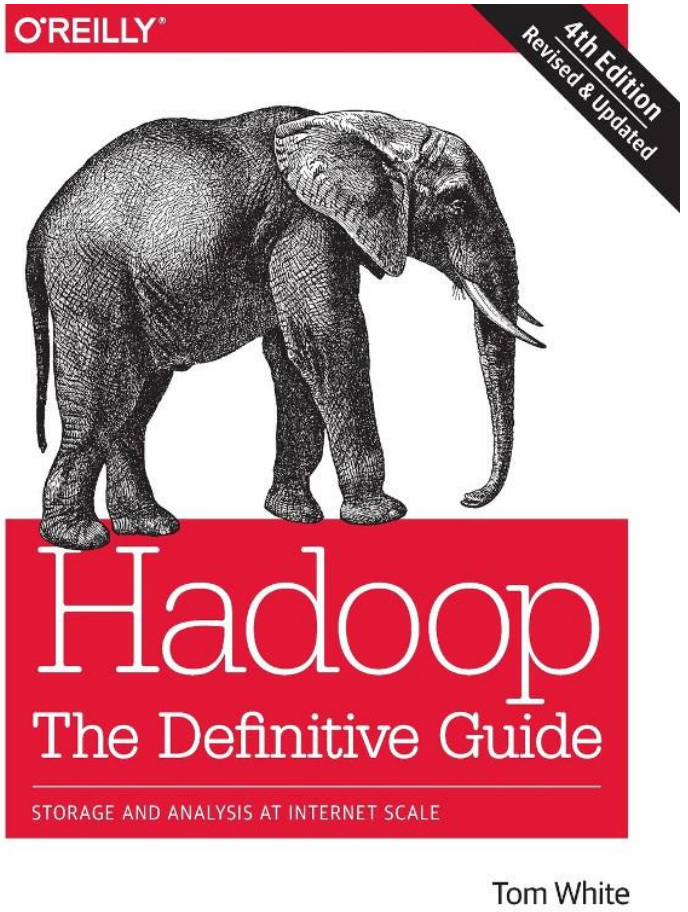
- Getting top-level view on data, summarization of data
- Large-scale indexing for search engines
- Finding popular search queries (e.g., Google Trends)
- Clustering of news articles (e.g., Google News)
- Processing satellite imagery data
- In general: Large-scale machine learning problems (e.g. Apache Mahout)
  - Clustering
  - Classification
  - …

Brief History

- **2004:** Google publishes paper introducing MapReduce
- **2005:** MapReduce implemented in Apache Nutch (web crawler)
- **2006:** Development of Hadoop MapReduce at Yahoo!
- **January 2008:** Hadoop gets top-level project at Apache
- **April 2008:** Hadoop is fastest system to sort a terabyte of data
- **Today:** Hadoop ecosystem as platform for big data in the industry
  - Companies using Hadoop: Facebook, Google, Yahoo!, New York Times

Literature

- White, Tom. “Hadoop: The Definitive Guide”. 4th Edition, O'Reilly Media, 2015
- Some slides based on this book
- Available as ebook and hard copy in the university library
- Note: Reading this book is not required, neither for the lecture and exercises nor for the exam.



Further Resources

- Google paper: Dean, Jeffrey, and Sanjay Ghemawat. “MapReduce: simplified data processing on large clusters.” *Communications of the ACM* 51.1 (2008): 107-113. https://doi.org/10.1145/1327452.1327492
- Hadoop Documentation: https://hadoop.apache.org/docs/stable