

Undergraduate Topics in Computer Science

Frank Nielsen

Introduction to HPC with MPI for Data Science



This gentle introduction to High Performance Computing (HPC) for Data Science using the Message Passing Interface (MPI) standard has been designed as a first course for undergraduates on parallel programming on distributed memory models, and requires only basic programming notions.

Divided into two parts the first part covers high performance computing using C++ with the Message Passing Interface (MPI) standard followed by a second part providing high-performance data analytics on computer clusters.

In the first part, the fundamental notions of blocking versus non-blocking point-to-point communications, global communications (like broadcast or scatter) and collaborative computations (reduce), with Amdahl and Gustafson speed-up laws are described before addressing parallel sorting and parallel linear algebra on computer clusters. The common ring, torus and hypercube topologies of clusters are then explained and global communication procedures on these topologies are studied. This first part closes with the MapReduce (MR) model of computation well-suited to processing big data using the MPI framework.

In the second part, the book focuses on high-performance data analytics. Flat and hierarchical clustering algorithms are introduced for data exploration along with how to program these algorithms on computer clusters, followed by machine learning classification, and an introduction to graph analytics. This part closes with a concise introduction to data core-sets that let big data problems be amenable to tiny data problems.

Exercises are included at the end of each chapter in order for students to practice the concepts learned, and a final section contains an overall exam which allows them to evaluate how well they have assimilated the material covered in the book.

Computer Science

ISBN 978-3-319-21902-8



► springer.com

UTiCS
Nielsen



Introduction to HPC with MPI for
Data Science

Undergraduate Topics in Computer Science

Frank Nielsen

Introduction to HPC with MPI for Data Science



 Springer