UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

High-Perfomance Computing for AI Applications in Physics

2627-2-F9103Q041

Aims

The educational aims of this course include:

1. **(Knowledge and understanding)**
   Students will develop a solid foundation in High Performance Computing (HPC) principles, including parallel architectures, programming models, and the infrastructure required to support computationally intensive AI workloads.
2. **(Applying knowledge and understanding)**
   Students will be able to design and implement simple parallel applications using OpenMP and MPI, deploy AI workloads on HPC systems, and apply optimization techniques to improve performance and scalability.
3. **(Making judgements)**
   Students will critically assess HPC and AI workload requirements, select appropriate architectures and parallelization strategies, and evaluate the trade-offs involved in performance and resource usage.
4. **(Communication skills)**
   Students will communicate clearly and effectively—both orally and in writing—about the design, implementation, and performance evaluation of HPC and AI solutions, using appropriate technical language and domain-specific terminology.
5. **(Learning skills)**
   Students will acquire the ability to independently learn and adapt to new HPC tools, AI frameworks, and parallel computing techniques, preparing them for ongoing developments and trends in the HPC for AI landscape.

Contents

Fundaments of High Performance Computing. Types of HPC hardware. Parallel programmiming models such as OpenMP and MPI. Parallel I/O and file systems. Scaling benchmaking and optimization. Batch scheduling. Data

storage.
Types of AI workloads. Data parallelism, model parallelism, and other strategies for the distribution of AI workloads. Efficient use of CPUs and specialized hardware.

Detailed program

The program of the course will cover the following topics:

- **HPC architectures and parallel programming models**
  The fundamentals of High Performance Computing. Overview of HPC hardware, including CPUs, GPUs, memory hierarchies, and interconnect technologies that enable fast data movement and computation.
  Shared-memory programming with OpenMP and distributed-memory programming with MPI. Trade-offs between communication overhead, synchronization, and data locality in parallel applications.
  Hands-on: Writing basic OpenMP and MPI programs.

- **Scaling applications**
  Scaling applications across multiple nodes in an HPC cluster, addressing issues like load balancing and task scheduling. Basics of parallel I/O and parallel file systems. Principles of data managment. Checkpointing strategies.
  Hands-on: Running parallel jobs on an HPC cluster.

- **Performance optimization**
  Profiling and benchmarking HPC applications. Memory optimization and cache efficiency. Vectorization and instruction-level parallelism. Overview of power-aware scheduling and energy-efficient hardware.
  Hands-on: Benchmarking weak and strong scaling of an application.

- **Introduction to AI in HPC**
  Why AI needs HPC: computational challenges in deep learning. Overview of AI workloads: training vs. inference. Examples of AI framworks and their HPC integration.
  Hands-on: Running a basic AI model on an HPC system.

- **Scaling AI workloads**
  Containerization tecniques and reproducibility in HPC. Data parallelism, model parallelism, and other strategies for distributed deep learning. Optimization techniques. Checkpointing AI workloads.
  Hands-on: Training AI models on multiple HPC nodes.

Prerequisites

- **Programming skills in Python, and basic knowledge of C or C++:** students should be comfortable writing and debugging small programs in Python. In the first part of the course, simple programs in C or C++ will also be used.
- **Familiarity with the foundations of Artificial Intelligence:** students are expected to have completed introductory AI courses from the first year of the master's programme.
- **Prior exposure to Linux/Unix environments and basic command-line tools (optional but highly recommended):** students will work with the shell, edit simple scripts, and navigate a Linux file system. Practical resources for beginners will be provided.

Teaching form

The course consists of 13 lessons of 4 hours in presence, for a total of 52 hours.
Part of the course (mostly in the first half of the course and of each lesson) will be delivered with frontal teaching. The remaining part will consist of interactive teaching.
Lessons will be at a computer station in the "Marco Comi" laboratory (room 2026, floor 2, U2 Quantum building, University of Milan-Bicocca).

Textbook and teaching resource

Teaching resources and textbook references are available at the link:
https://virgilio.mib.infn.it/~marcoce/teaching/hpc4ai/

Semester

First semester, second year

Assessment method

The course includes practical assignments to be completed in the computer lab. Each student collects the results of the assignments in an individual report to be submitted to the teacher before the exam.
The exam will consist in an oral examination that will assess both the report on the assignments and the understanding of the theoretical aspects. The final evaluation will take into account the lab activity, the final report, and the oral exam.

Office hours

By appointment, by writing an email (marco.ce@unimib.it) to the course teacher.

Sustainable Development Goals

QUALITY EDUCATION | INDUSTRY, INNOVATION AND INFRASTRUCTURE