UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

High-Perfomance Computing for AI Applications in Physics

2627-2-F9103Q041

Obiettivi

Gli obiettivi formativi del corso includono:

1. **(Conoscenza e comprensione)**
   Gli studenti svilupperanno una solida conoscenza dei principi del calcolo ad alte prestazioni (HPC), incluse le architetture parallele, i modelli di programmazione e le infrastrutture necessarie per supportare carichi di lavoro di intelligenza artificiale (AI) ad alta intensità computazionale.
2. **(Capacità di applicare conoscenza e comprensione)**
   Gli studenti saranno in grado di progettare e implementare semplici applicazioni parallele utilizzando OpenMP e MPI, eseguire carichi di lavoro di AI su sistemi HPC e applicare tecniche di ottimizzazione per migliorarne le prestazioni e la scalabilità.
   3 **(Autonomia di giudizio)**
   Gli studenti valuteranno criticamente i requisiti computazionali di carichi di lavoro HPC e AI, selezionando le architetture e le strategie di parallelizzazione più appropriate, e analizzando i compromessi tra prestazioni e utilizzo delle risorse.
3. **(Abilità comunicative)**
   Gli studenti saranno in grado di comunicare in modo chiaro ed efficace – sia oralmente che per iscritto – riguardo alla progettazione, implementazione e valutazione delle prestazioni di soluzioni HPC e AI, utilizzando un linguaggio tecnico appropriato e una terminologia specifica del dominio.
4. **(Capacità di apprendimento)**
   Gli studenti acquisiranno la capacità di apprendere autonomamente nuovi strumenti HPC, framework AI e tecniche di calcolo parallelo, preparandosi a seguire l'evoluzione continua dell'HPC per l'AI.

Contenuti sintetici

Fondamenti del calcolo ad alte prestazioni (HPC). Tipi di hardware HPC. Modelli di programmazione parallela

come OpenMP e MPI. I/O e file system paralleli. Benchmaking e ottimizzazione dello scaling. Scheduling batch. Archiviazione dei dati.
Tipi di carichi di lavoro di intelligenza artificiale. Parallelismo sui dati, parallelismo sul modello e altre strategie per la distribuzione dei carichi di lavoro di intelligenza artificiale. Uso efficiente di CPU e hardware specializzato.

Programma esteso

Il programma del corso coprirà i seguenti argomenti:

- **Architetture HPC e modelli di programmazione parallela**
  Fondamenti del calcolo ad alte prestazioni (High Performance Computing). Panoramica dell'hardware HPC, inclusi CPU, GPU, gerarchie di memoria e tecnologie di interconnessione che abilitano il rapido trasferimento dei dati e il calcolo.
  Programmazione a memoria condivisa con OpenMP e programmazione a memoria distribuita con MPI. Compromessi tra overhead di comunicazione, sincronizzazione e località dei dati nelle applicazioni parallele.
  Esercitazione: scrittura di semplici programmi OpenMP e MPI.

- **Scalabilità delle applicazioni**
  Scalabilità delle applicazioni su più nodi di un cluster HPC, affrontando temi come il bilanciamento del carico e la pianificazione dei task. Fondamenti di I/O parallelo e dei file system paralleli. Principi di gestione dei dati. Strategie di checkpointing.
  Esercitazione: esecuzione di job paralleli su un cluster HPC.

- **Ottimizzazione delle prestazioni**
  Profilazione e benchmarking di applicazioni HPC. Ottimizzazione della memoria ed efficienza della cache. Vettorizzazione e parallelismo a livello di istruzioni. Panoramica su tecniche di scheduling consapevoli del consumo energetico e hardware ad alta efficienza energetica.
  Esercitazione: benchmarking dello scaling debole e forte di un'applicazione.

- **Introduzione all'AI in ambito HPC**
  Perché l'AI ha bisogno dell'HPC: sfide computazionali del deep learning. Panoramica sui carichi di lavoro AI: training vs. inferenza. Esempi di framework AI e loro integrazione con l'HPC.
  Esercitazione: esecuzione di un modello AI di base su un sistema HPC.

- **Scalabilità dei carichi di lavoro AI**
  Tecniche di containerizzazione e riproducibilità in ambito HPC. Parallelismo sui dati, parallelismo sul modello e altre strategie per il deep learning distribuito. Tecniche di ottimizzazione. Checkpointing dei carichi di lavoro AI.
  Esercitazione: training di modelli AI su più nodi HPC.

Prerequisiti

- **Competenze di programmazione in Python e conoscenze di base di C o C++**: gli studenti dovrebbero essere in grado di scrivere e fare il debug di piccoli programmi in Python. Nella prima parte del corso verranno utilizzati anche semplici programmi in C o C++.
- **Familiarità con i fondamenti dell'Intelligenza Artificiale**: si presuppone che gli studenti abbiano seguito corsi introduttivi di AI durante il primo anno del corso di laurea magistrale.

- **Esperienza pregressa con ambienti Linux/Unix e strumenti da linea di comando (facoltativa ma fortemente consigliata)**: gli studenti lavoreranno con la shell, modificheranno semplici script e si muoveranno all'interno di un file system Linux. Verranno forniti materiali pratici per i principianti.

Modalità didattica

L'insegnamento consiste in 13 lezioni da 4 ore, esclusivamente in presenza, per un totale di 52 ore.
Parte del corso (per lo più nella prima metà del corso e di ogni lezione) sarà svolta in modalità erogativa. La parte rimanente sarà in modalità interattiva.
Le lezioni saranno a una postazione informatica nel laboratorio "Marco Comi" (aula 2026, piano 2, edificio U2 Quantum, Università di Milano-Bicocca).

Materiale didattico

Il materiale didattico è disponibile al link:
https://virgilio.mib.infn.it/~marcoce/teaching/hpc4ai/

Periodo di erogazione dell'insegnamento

Primo semestre, secondo anno

Modalità di verifica del profitto e valutazione

Il corso prevede lo svolgimento di esercitazioni pratiche al computer in laboratorio. Ogni studente raccoglie i risultati delle esercitazioni in una relazione individuale da consegnare al docente prima dell'esame finale.
L'esame finale sarà un orale che valuterà sia la relazione sia la comprensione degli aspetti teorici del corso. La valutazione finale terrà conto dell'attività di laboratorio, della relazione finale e dell'esame orale.

Orario di ricevimento

Su appuntamento, scrivendo un'email (marco.ce@unimib.it) al docente del corso.

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ | IMPRESE, INNOVAZIONE E INFRASTRUTTURE