MODULO C

MPI avanzato, GPU avanzate, algoritmi e deployment (40h)

Prerequisiti

Contenuti dei Moduli A e B o competenze equivalenti (ottimizzazione seriale, OpenMP, MPI, nozioni di architettura del nodo e della rete di interconnessione).

Contenuti didattici

- GPU I: architettura GPU e confronto con la CPU (lato HW e SW); modello di programmazione; CUDA di base; offloading direttivo (OpenMP/OpenACC); memory model - coalescing, shared memory, latency hiding.
- GPU II (avanzata e distribuita): esecuzione asincrona e programmazione multi-stream; programmazione multi-GPU; HIP e performance portability; introduzione alle GPU AMD e differenze con NVIDIA; nuove architetture (riduzione FP32/FP64, crescita FP16/FP8, mixed-precision); memoria distribuita su GPU - GPU-aware MPI, NCCL/RCCL, NVSHMEM; approccio ibrido MPI+GPU su cluster multi-nodo.
- Profiling su larga scala: NVIDIA Nsight family (Systems e Compute), compute-sanitizer; profiling di applicazioni massicciamente parallele e analisi delle prestazioni alla scala di produzione.

Durata

40h — GPU I + profiling ~16h - GPU II (multi-GPU/multi-nodo) + profiling ~18h - Ottimizzazione e profiling su larga scala ~6h.

Obiettivi formativi specifici

Comprendere l'architettura GPU e saper scrivere e ottimizzare kernel di base sfruttando il memory model; saper gestire scenari multi-GPU e multi-nodo con communication libraries dedicate (GPU-aware MPI, NCCL/RCCL, NVSHMEM) adottando un approccio misto MPI+GPU; saper profilare applicazioni alla scala di produzione con strumenti di profiling avanzato.

Competenze specifiche in uscita

Il partecipante sa scrivere e ottimizzare kernel di base e portarli su GPU ottimizzandone gli accessi in memoria, misurandone le prestazioni con strumenti di profiling avanzato. Sa utilizzare cluster multi-nodo e multi-GPU programmando i propri codici secondo un approccio misto MPI+GPU e