

Harnessing Parallel Paradigms: A Comprehensive Guide to OpenMP, MPI, and GPU Offloading in HPC

Organization Committee

- Diego Andres Roa Perdomo (Presenter)
- Paige Kinsley
- Kevin A. Brown
- Xiaoming Li

Abstract

High-Performance Computing (HPC) provides scientists with tools that enable research of grand challenges and global-scale problems. To fully exploit the potential of these machines, users rely on heterogeneous hardware that can include arrays of CPUs and GPUs. Harnessing the full potential of modern processors requires a deep understanding of parallel programming paradigms and high-performance network interconnects. This tutorial aims to provide a detailed and practical guide to these techniques. We will explore OpenMP for shared-memory parallelism, MPI for distributed-memory systems, GPU offloading for leveraging the computational power of modern accelerators, and how to take advantage of the network interconnect. Through this comprehensive guide, participants will learn how to effectively combine these paradigms to solve complex computational problems, optimize performance, and maximize resource utilization in heterogeneous computing environments.

The tutorial is designed for HPC practitioners, researchers, and students who seek to deepen their understanding of parallel programming models. We will cover fundamental concepts, advanced techniques, and real-world applications, providing hands-on examples and performance optimization strategies. By the end of this tutorial, participants will be equipped with the knowledge and skills to implement and optimize parallel applications across diverse HPC architectures. This tutorial doesn't focus on performance, but rather on understanding concepts that can have a significant impact on how parallel programs are executed. This tutorial uses Chameleon Cloud to provide access to compute resources that contain GPUs. It is also written in Jupyter Notebooks, lowering the bar for attendees with little command-line (CLI) experience.

Topic and Relevance

- Parallel Programming Models for HPC
- Shared Memory: OpenMP
- Distributed Memory: MPI
- Accelerators: Offloading

- High-performance Networks Overview

Relevance: To use large computing systems, users are expected to have a certain level of proficiency and show scalability in their applications. This tutorial will provide users with the basic tools to map complex computational problems in these machines and support the request of additional compute resources. Using HPC resources enables science at a scale that is not possible with other machines.

Prerequisites

Attendees are expected to be familiar with:

- The tutorial will be in C and C++. People with experience in other programming languages are welcome too
- Simple understanding of how to use compilers
- Basic experience with Linux systems

Other desired, but not necessary knowledge includes:

- Understanding the differences between threads and processes
- Understanding pointers, memory, and how data is laid out in memory.
- Basic understanding of accelerator devices (e.g., GPUs)

Resources and Duration

Duration:

4 hours

Resources:

- Laptop with internet connection.
- Chameleon Cloud access (Provided)

Previous editions

- Presented at Argonne National Laboratory (July 25, 2024)