Contribution ID: **509** Type: **not specified**

Optimization and Placement Patterns for training and inference workloads

Monday, 1 December 2025 14:10 (20 minutes)

The current paradigm in HPC for AI has shifted from simple “data parallelism” to multi-dimensional (3D) parallelism and heterogeneous co-execution. In this talk, we will discuss optimization and placement patterns for training and inference workloads. For Training, we will focus on topology-aware placement that minimizes inter-node communication latency using 3D parallelism. For inference, We will also focus on maximizing throughput per watt and utilizing “stranded” capacity via hybrid CPU-GPU pipelining and dynamic model partitioning (e.g., Multi-Instance GPU or MIG). We will then demonstrate how these placement strategies can be used to harness the power of HPC in AI workloads by applying 3D parallelism and heterogeneous co-execution.

Presenting Author

Email

Student or Postdoc?

Institute

Registered for the conference?

CHPC User

CHPC Research Programme

Workshop Duration

Primary author: Dr MOHAPI, Lerato (Eclipse Holdings)

Presenter: Dr MOHAPI, Lerato (Eclipse Holdings)

Session Classification: HPC Technology