Contribution ID: **495**

Type: **Talk**

® Scalable Data Management Techniques for AI workloads

Tuesday, 2 December 2025 16:30 (20 minutes)

Title: Scalable Data Management Techniques for AI workloads

Abstract: The advent of complex AI workflows that involve large learning models (training using data/pipeline/tensor parallelism, retrieval augmented generation, chaining) has prompted the need for scalable system-level building blocks that enable running them efficiently at large scale on high end machines. Of particular interest in this context are data management techniques and their implementation that bridge the gap between high-level required capabilities (fine-grain tensor access, support for transfer learning and versioning, streaming and transformation of training samples, transparent augmentation, vector databases, etc.) and the existing storage hierarchy (parallel file systems, node-local memories, etc.). This talk discusses the challenges and opportunities in the design and development of such techniques and presents several results based on VELOC and DataStates, two efforts at ANL aimed at leveraging checkpointing to capture the evolution of datasets (including AI models and their training data).

Presenting Author

Email

Student or Postdoc?

Institute

Registered for the conference?

CHPC User

CHPC Research Programme

Workshop Duration

Primary author: NICOLAE, Bogdan (Argonne National Laboratory)

Presenter: NICOLAE, Bogdan (Argonne National Laboratory)

Session Classification: HPC Technology

Track Classification: Storage and IO