Contribution ID: **82** Type: **Short talk**

Perspectives on the Versatility of a Searchable Lineage for Scalable HPC Data Management

Thursday 23 March 2023 16:30 (10 minutes)

Checkpointing is the most widely used approach to provide resilience for HPC applications by enabling restart in case of failures. However, coupled with a searchable lineage that records the evolution of intermediate data and metadata during runtime, it can become a powerful technique in a wide range of scenarios at scale: verify and understand the results more thoroughly by sharing and analyzing intermediate results (which facilitates provenance, reproducibility, and explainability), new algorithms and ideas that reuse and revisit intermediate and historical data frequently (either fully or partially), manipulation of the application states (job pre-emption using suspend-resume, debugging), etc.

This talk advocates a new data model and associated tools (DataStates, VELOC) that facilitate such scenarios. Avoid direct use of a data service to read and write datasets; instead, during runtime, users should tag datasets with properties that express hints, constraints, and persistency semantics. Doing so will automatically generate a searchable record of intermediate data checkpoints, or data states, optimized for I/O. Such an approach brings new capabilities and enables high performance scalability, and FAIR-ness through a range of transparent optimizations. The talk will introduce DataStates and VELOC, will underline several vital technical details, and will conclude with several examples of where they were successfully applied.

JLESC topic

Primary author: NICOLAE, Bogdan (ANL)

Presenter: NICOLAE, Bogdan (ANL)

Session Classification: Short Talks on Workflows, I/O and Frameworks

Track Classification: I/O, storage and in-situ processing