Contribution ID: **14** Type: **not specified**

Data Lineage in High-Performance Computing Environments

Wednesday, 10 April 2024 12:55 (10 minutes)

As the complexity of high-performance computing (HPC) continues to grow, data management becomes a critical challenge. In HPC environments, where data processing occurs on a massive scale, tracing data lineage —from its source to its utilization in analyses and computations—and understanding data provenance is a key element in ensuring data integrity, regulatory compliance, and performance optimization. Furthermore, ensuring the reproducibility of scientific results is paramount in such environments. In this presentation, we will present an analysis of data lineage, provenance, and reproducibility in the context of HPC environments, discussing techniques, tools, and best practices for data management in such complex settings. We will focus on issues related to identifying data sources, tracking the flow of data through various processing stages, ensuring data consistency and quality, establishing data provenance, and facilitating the reproducibility of scientific results.

Presenter: Dr TYKIERKO, Mateusz (Wroclaw Centre for Networking and Supercomputing)

Session Classification: Session I