Contribution ID: **32** Type: **Oral Presentation**

Exploring Database-aided Workflows on Cloud and High-Performance Computing for Physics Simulation Data

Thursday, March 28, 2024 11:20 AM (20 minutes)

While database management systems (DBMS) are one of the most important IT concepts, scientific supercomputing makes little use of them. Reasons for this situation range from a preference towards direct file I/O without overheads to feasibility problems in reaching DBMS from High-Performance Computing (HPC) cluster nodes. However, trends such as the increasing re-usage of scientific output data or the collaboration of researchers from science, SMEs and industry give a strong motivation to revisit the topic. The Horizon Europe "Extreme Data" project EXA4MIND, for example, aims at bridging the ecosystems of DBMS, supercomputing, and European Data Spaces.

In the context of this project, the work presented here explores different approaches and systems for managing data and thus optimising data-driven workflows across Cloud-Computing (IaaS) and HPC systems. We evaluate typical workflows for physics simulations on supercomputing systems at LRZ (Garching b.M./DE) and IT4Innovations (Ostrava/CZ). The use cases we focus upon in this contribution are simulated many-body systems of molecules or elementary particles on different energy scales, either using molecular dynamics (MD, low energy) or plasma physics (high energy). As often in computational science, much work goes into postprocessing, visualising and discussing the simulated data, often several times in an iterative process. Our test datasets are, on the one hand, produced by MD simulations (Modelling for Nanotechnologies Lab, IT4Innovations, Ostrava/CZ), where the interaction of molecules is calculated via empirical force field models. On the other hand, we have outputs from the Plasma Simulation Code (Ruhl et al., Ludwig Maximilian University of Munich/DE), simulating the fields and trajectories of charged particles in a plasma. These simulations follow up to billions of particles, writing out their properties and location trajectories in each time step.

The production of such simulation outputs can take up to hundreds of thousands CPU hours, and the particle and field data can occupy Gigabytes or Terabytes. These then have to be postprocessed (e.g. aggregation of domain patches, extraction of statistical information), and evaluated, on various levels – from ensembles of simulations down to single particle trajectories or timesteps. Even for one study, the datasets are revisited typically several times, for example for weight recalculation, visualization and evaluation of the validity of force field assumptions. These workflows typically involve a lot of manual labour and attention from the researcher. We benchmark proper data backends (including DBMS) and use cross-system orchestration tools, in particular the LEXIS platform (lexis-project.eu), to make this more efficient.

Our focus includes testing the performance of typical data queries and iterative postprocessing steps with different execution methods. We strive to facilitate faster and more flexible access to the raw data by exploring the properties of different storage and database systems. These range from data access schemes facilitated by common python environments over row-based DBMS such as PostgreSQL to column DBMS like MonetDB, where techniques like SciQL can operate live queries on large array-based datasets in memory, or functionalities like Data Vaults can provide access to external repositories. We conclude our contribution stating that modern data storage concepts involving DBMS are also an optimum basis for data sharing and systematic metadata management. Thus, we aim at facilitating a research data management according to the FAIR (findable, accessible, interoperable, reusable) principles from the start.

This research received the support of the EXA4MIND project, funded by a European Union´s Horizon Europe Research and Innovation Programme, under Grant Agreement N° 101092944. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Primary author: PAUW, Viktoria (Leibniz Rechenzentrum)

Presenter: PAUW, Viktoria (Leibniz Rechenzentrum)

Session Classification: Physics & Engineering Application

Track Classification: Track 1: Physics and Engineering Applications