Contribution ID: **28** Type: **Oral Presentation**

Quasi interactive analysis of High Energy Physics big data with high throughput

Tuesday, 18 March 2025 15:00 (20 minutes)

The ability to ingest, process, and analyze large datasets within minimal timeframes is a milestone of big data applications. In the realm of High Energy Physics (HEP) at CERN, this capability is especially critical as the upcoming high-luminosity phase of the LHC will generate vast amounts of data, reaching scales of approximately 100 PB/year. Recent advancements in resource management and software development have enabled more flexible and dynamic data access, alongside the integration with open-source tools like Jupyter, Dask, and HTCondor. These advancements facilitate a shift from a traditional "batch-like" processing to an interactive, high-throughput platform that utilizes a distributed, parallel back-end architecture. This approach is further supported by the DataLake model developed by the Italian National Center for "High-Performance Computing, Big Data, and Quantum Computing Research Centre" (ICSC).

This contribution highlights the transition of various data analysis applications, from legacy batch processing to a more interactive, declarative paradigm using tools like ROOT RDataFrame. These applications are executed on the aforementioned cloud-based infrastructure, with workflows distributed across multiple worker nodes and results consolidated into a unified interface. Additionally, the performance of this approach will be evaluated through speed-up benchmarks and scalability tests using distributed resources. The analysis aims to identify potential bottlenecks or limitations of the high-throughput interactive model, providing insights that will guide its further development and implementation within the Italian National Center.

Primary authors: Dr GRAVILI, Francesco Giuseppe (Università del Salento e INFN); DIOTALEVI, Tommaso (INFN and University of Bologna)

Presenter: DIOTALEVI, Tommaso (INFN and University of Bologna)

Session Classification: Infrastructure Clouds & Virtualisation

Track Classification: Track 8: Infrastructure Clouds and Virtualizations