Contribution ID: **149** Type: **Oral Presentation**

Offloading CMS data analysis on a distributed high-throughput platform with RDataFrame

Thursday, 19 March 2026 16:00 (30 minutes)

The ability to ingest, process, and analyze large datasets within minimal timeframes is a cornerstone of modern big data applications. In High Energy Physics (HEP), this need becomes increasingly critical as the upcoming High-Luminosity phase of the LHC at CERN is expected to produce data volumes approaching 100 PB per year. Recent advancements in resource management and open-source computing frameworks - such as Jupyter, Dask, and HTCondor - are driving a shift from traditional batch-oriented workflows toward interactive, high-throughput analysis environments.
Within this context, and leveraging the computing resources of the Italian "National Center for High-Performance Computing, Big Data, and Quantum Computing (ICSC)", a scalable analysis platform has been developed. Such system allows users to dynamically distribute workloads across local Kubernetes resources or offload them to remote infrastructures through interLink, a technology that extends the Virtual Kubelet concept to federate heterogeneous resources, like High-Throughput Computing (HTC), High-Performance Computing (HPC), and Cloud systems, under a unified orchestration layer.
The platform's performance has been then evaluated using a representative use case: the study of the CMS Drift Tubes (DT) muon detector performance, in phase-space regions driven by analysis needs. By exploiting the declarative model of ROOT RDataFrame (RDF) and its distributed execution via Dask, the study demonstrates significant improvements in scalability and speed-up compared to traditional serial workflows. These results confirm the effectiveness of the proposed distributed analysis approach, addressing the computational challenges posed by the High-Luminosity LHC era.

Primary authors: DIOTALEVI, Tommaso (INFN and University of Bologna); BATTILANA, Carlo (University of Bologna and INFN); FANFANI, Alessandra (University of Bologna and INFN); ROSSI, Elvira (University of Naples and INFN); SPIGA, Daniele (INFN-PG); TEDESCHI, Tommaso (University and INFN, Perugia (Italy)); CIANGOTTINI, Diego (INFN Perugia); DORIA, Alessandra (INFN); PARDI, Silvio (INFN-Napoli); SPISSO, Bernardino (INFN); STELLACCI, Simona Maria (INFN)

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

Session Classification: Infrastructure Clouds and Virtualisations - III

Track Classification: Track 8: Infrastructure Clouds and Virtualizations