Contribution ID: **132** Type: **Oral**

Novel functional and distributed approaches to data analysis available in ROOT

Tuesday 22 August 2017 18:25 (20 minutes)

The bright future of particle physics at the Energy and Intensity frontiers poses exciting challenges to the scientific software community. The traditional strategies for processing and analysing data are evolving in order to cope with the ever increasing complexity and size of the datasets.
The traditional strategies for processing and analysing data are evolving in order to (i) offer higher-level programming model approaches and (ii) exploit parallelism to cope with the ever increasing complexity and size of the datasets.
This contribution describes how the ROOT framework, a cornerstone of software stacks dedicated to particle physics, is preparing to provide adequate solutions for the analysis of large amount of scientific data on parallel architectures.

The functional approach to parallel data analysis provided with the ROOT TDataFrame interface is then characterised. The design choices behind this new interface are described also comparing with other widely adopted tools such as Pandas and Apache Spark. Commonalities and differences with ReactiveX and Ranges v3 are highlighted.
The programming model is illustrated highlighting the reduction of boilerplate code, composability of the actions and data transformations as well as the capabilities of dealing with different data sources such as ROOT, json, csv or databases.
Details are given about how the functional approach allows transparent implicit parallelisation of the chain of operations specified by the user.

The progress done in the field of distributed analysis is examined. In particular, the power of the integration of ROOT with Apache Spark via the PyROOT interface is shown.

In addition, the building blocks for the expression of parallelism in ROOT are briefly characterised together with the structural changes applied in the building and testing infrastructure which were necessary to put them in production.

All new ROOT features are accompanied by scaling and performance measurements of real life use cases on highly parallel and distributed architectures.

Authors: TEJEDOR SAAVEDRA, Enric (CERN); PIPARO, Danilo (CERN); CANAL, Philippe (Fermi National Accelerator Lab. (US)); GUIRAUD, Enrico (CERN, University of Oldenburg (DE)); VALLS PLA, Xavier (University Jaume I (ES)); GANIS, Gerardo (CERN); MATO VILA, Pere (CERN); AMADIO, Guilherme (CERN); MONETA, Lorenzo (CERN)

Co-author: BLOMER, Jakob (CERN)

Presenter: AMADIO, Guilherme (CERN)

Session Classification: Track 2: Data Analysis - Algorithms and Tools

Track Classification: Track 2: Data Analysis - Algorithms and Tools