Contribution ID: **152** Type: **Oral**

A quantitative review of data formats for HEP analyses

Thursday 24 August 2017 15:40 (20 minutes)

The analysis of High-Energy Physics (HEP) data sets often take place outside the realm of experiment frameworks and central computing workflows, using carefully selected "n-tuples" or Analysis Object Data (AOD) as a data source. Such n-tuples or AODs may comprise data from tens of millions of events and grow to hundred gigabytes or a few terabytes in size. They are typically small enough to be processed by an institute's cluster or even by a single workstation. N-tuples and AODs are often stored in the ROOT file format, in an array of serialized C++ objects in columnar storage layout. In recent years, several new data formats emerged from the data analytics industry. We provide a quantitative comparison of ROOT and other popular data formats, such as Apache Parquet, Apache Avro, Google Protobuf, and HDF5. We compare speed, read patterns, and usage aspects for the use case of a typical LHC end-user n-tuple analysis. The performance characteristics of the relatively simple n-tuple data layout also provides a basis for understanding performance of more complex and nested data layouts. From the benchmarks, we derive performance tuning suggestions both for the use of the data formats and for the ROOT (de-)serialization code.

Author: BLOMER, Jakob (CERN)

Presenter: BLOMER, Jakob (CERN)

Session Classification: Track 1: Computing Technology for Physics Research

Track Classification: Track 1: Computing Technology for Physics Research