Contribution ID: **50** Type: **Oral**

CMS Analysis and Data Reduction with Apache Spark

Tuesday, 22 August 2017 16:45 (20 minutes)

Experimental Particle Physics has been at the forefront of analyzing the world's largest datasets for decades. The HEP community was among the first to develop suitable software and computing tools for this task. In recent times, new toolkits and systems for distributed data processing, collectively called "Big Data"technologies have emerged from industry and open source projects to support the analysis of Petabyte and Exabyte datasets in industry. While the principles of data analysis in HEP have not changed (filtering and transforming experiment-specific data formats), these new technologies use different approaches and tools, promising a fresh look at analysis of very large datasets that could potentially reduce the time-to-physics with increased interactivity. Moreover these new tools are typically actively developed by large communities, often profiting of industry resources, and under open source licensing. These factors result in a boost for adoption and maturity of the tools and for the communities supporting them, at the same time helping in reducing the cost of ownership for the end-users. In this talk, we are presenting studies of using Apache Spark for end user data analysis. We are studying the HEP analysis workflow separated into two thrusts: the reduction of centrally produced experiment datasets and the end-analysis up to the publication plot. Studying the first thrust, CMS is working together with CERN openlab and Intel on the CMS Big Data Reduction Facility. The goal is to reduce 1 PB of official CMS data to 1 TB of ntuple output for analysis. We are presenting the progress of this 2-year project with first results of scaling up Spark-based HEP analysis. Studying the second thrust, we are presenting studies on using Apache Spark for a CMS Dark Matter physics search, investigating Spark's feasibility, usability and performance compared to the traditional ROOT-based analysis.

Primary authors: CMS, Collaboration; GUTSCHE, Oliver (Fermi National Accelerator Lab. (US))

Presenter: GUTSCHE, Oliver (Fermi National Accelerator Lab. (US))

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

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