Contribution ID: **513** Type: **Oral**

SciDAC-Data, A Project to Enabling Data Driven Modeling of Exascale Computing

Tuesday 11 October 2016 11:45 (15 minutes)

The SciDAC-Data project is a DOE funded initiative to analyze and exploit two decades of information and analytics that have been collected, by the Fermilab Data Center, on the organization, movement, and consumption of High Energy Physics data. The project is designed to analyze the analysis patterns and data organization that have been used by the CDF, DØ, NO⊠A, Minos, Minerva and other experiments, to develop realistic models of HEP analysis workflows and data processing. The SciDAC-Data projects aims to provide both realistic input vectors and corresponding output data which can be used to optimize and validate simulations of HEP analysis in different high performance computing (HPC) environments. These simulations are designed to address questions of data handling, cache optimization and workflow structures that are the prerequisites for modern HEP analysis chains to be mapped and optimized to run on the next generation of leadership class exascale computing facilities.

We will address the use of the SciDAC-Data distributions acquired from over 5.6 million analysis workflows and corresponding to over 410,000 HEP datasets, as the input to detailed queuing simulations that model the expected data consumption and caching behaviors of the work running in HPC environments. In particular we describe in detail how the SAM data handling system in combination with the dCache/Enstore based data archive facilities have been analyzed to develop the radically different models of the analysis of collider data and that of neutrino datasets. We present how the data is being used for model output validation and tuning of these simulations. The paper will address the next stages of the SciDAC-Data project which will extend this work to more detailed modeling and optimization of the models for use in real HPC environments.

Tertiary Keyword (Optional)

High performance computing

Secondary Keyword (Optional)

Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory)

Computing models

Authors: Dr NORMAN, Andrew (Fermilab); ALIAGA SOPLIN, Leonidas (College of William and Mary (US))

Co-authors: Dr LYON, Adam (Fermilab); Dr TSARIS, Aristeidis (Fermilab); Dr ALIAGA SOPLIN, Leonidas (Fermilab); Dr MUBAREK, Misbah (Argonne); DING, Pengfei (The University of Manchester); Dr ROSS, Robert (Argonne)

Presenter: ALIAGA SOPLIN, Leonidas (College of William and Mary (US))

Session Classification: Track 4: Data Handling

Track Classification: Track 4: Data Handling