Contribution ID: **143** Type: **oral presentation**

Automated recovery of data-intensive jobs in D0 and CDF using SAM

Monday, 13 February 2006 16:40 (20 minutes)

SAM is a data handling system that provides Fermilab HEP experiments of D0, CDF and MINOS with the means to catalog, distribute and track the usage of their collected and analyzed data. Annually, SAM serves petabytes of data to physics groups performing data analysis, data reconstruction and simulation at various computing centers across the world. Given the volume of the detector data, a typical physics analysis job consumes terabytes of information during several days of running at a job execution site. At any stage of that process, non systematic failures may occur, leaving a fraction of the original dataset unprocessed. To ensure convergence to completion of the computation request, a facility user has to employ a procedure to identify pieces of data that need to be re-analyzed in a manner that guarantees completeness without duplication in the final result. It is common that these issues are addressed by analyzing the output of the job. Such an approach is fragile, since it depends critically on the (changeable) output file format, and time-consuming. The approach that is reported in this article saves the user's time and ensures consistency in results. We present an automated method that uses SAM data handling to formalize distributed data analysis by defining a transaction based model of the physics analysis job work cycle to enable robust recovery of the unprocessed data.

Primary author: Mr BARANOVSKI, Andrew (FNAL)

Co-authors: Mr LYON, Adam (FNAL); Mr BENJAMIN, Doug (FNAL); Mr LIPELES, Elliot (FNAL); Mr SFILIGOI, Igor (FNAL); Mr GENSER, Krzysztof (FNAL); Ms BARTSCH, Valeria (FNAL)

Presenter: BARTSCH, Valeria (FERMILAB / University College London)

Session Classification: Distributed Data Analysis

Track Classification: Distributed Data Analysis