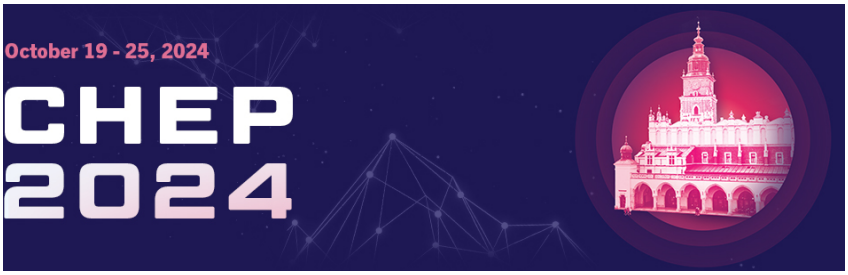

Contribution ID: **499** Type: **Talk**

Reshaping Analysis for Fast Turnaround

Thursday 24 October 2024 17:09 (18 minutes)

In the data analysis pipeline for LHC experiments, a key aspect is the step in which small groups of researchers —typically graduate students and postdocs—reduce the smallest, common-denominator data format down to a small set of specific histograms suitable for statistical interpretation. Here, we will refer to this step as "analysis"with the recognition that in other contexts, "analysis"might include other pieces, such as the actual computation required to extract statistical interoperation from the histograms. Analysis is a very important part of the pipeline as it is the step where individual researchers exercise their creativity in trying new ideas in the pursuit of discovery. Therefore, a critical metric for the analysis step is turnaround time because it determines how rapidly researchers can explore their space of ideas. We demonstrate our experience reshaping late-stage analysis applications on thousands of nodes with the goal of minimizing turnaround time. It is not enough merely to increase scale: it is necessary to make changes throughout the stack, including storage systems, data management, task scheduling, and application design. We demonstrate these changes when applied to CMS analysis applications built using the Coffea framework, leveraging Dask and TaskVine to scale out to distributed resources. We evaluate the performance of the applications on opportunistic campus clusters, showing effective scaling up to 7200 cores, thus producing significant improvement in turnaround time.

Primary authors: TOWNSEND, Austin (University of Notre Dame (US)); SLY-DELGADO, Barry (University of Notre Dame); TOVAR LOPEZ, Benjamin (University of Notre Dame); MOORE, Connor (University of Notre Dame (US)); THAIN, Douglas (University of Notre Dame); ZHOU, Jin (University of Notre Dame); LANNON, Kevin Patrick (University of Notre Dame (US))

Presenter: LANNON, Kevin Patrick (University of Notre Dame (US))

Session Classification: Parallel (Track 9)

Track Classification: Track 9 - Analysis facilities and interactive computing