Contribution ID: **250** Type: **presentation**

Columnar data processing for HEP analysis

Tuesday 10 July 2018 11:00 (15 minutes)

In the last stages of data analysis, only order-of-magnitude computing speedups translate into increased human productivity, and only if they're not difficult to set up. Producing a plot in a second instead of an hour is life-changing, but not if it takes two hours to write the analysis code. Fortunately, HPC-inspired techniques can result in such large speedups, but unfortunately, they can be difficult to use in a HEP setting.

These techniques generally favor operating on columns—arrays representing a single attribute across events, rather than whole events individually—which allows data to stream predictably from disk media to main memory and finally to CPU/GPU/KNL onboard memory (e.g. L* cache) for prefetching and sometimes allows for for vectorization. However, the need to work with variable-length structures in HEP, such as different numbers of particles per event, makes it difficult to apply this technique to HEP problems.

We will describe several new software tools to make it easier to compute analysis functions with columnar arrays in HEP: array-at-a-time I/O in ROOT ("BulkIO") and Python/Numpy ("uproot"), compiling object-oriented analysis code into columnar operations ("oamap" for "object-array mapping"), and storage solutions with columnar granularity. We will show performance plots and usage examples.

Authors: PIVARSKI, Jim (Princeton University); ELMER, Peter (Princeton University (US)); NANDI, Jaydeep; LANGE, David (Princeton University (US))

Presenter: PIVARSKI, Jim (Princeton University)

Session Classification: T6 - Machine learning and physics analysis

Track Classification: Track 6 –Machine learning and physics analysis