EXPAND

Characterizing NanoAOD Read Latency

Jonathan Guiang — Würthwein Group (Spring, 2021)

What you will do:

Recently, CMS physicists have transitioned to using a lightweight data structure called “NanoAOD” (a smaller version of “MiniAOD”) which provides only the essentials for physics analysis. Just like all data that LHC scientists analyze, NanoAOD is stored in the form of ROOT files, a custom binary file-type that is optimized for large datasets similar to Parquet or HDF5. Like some of its industry equivalents, data in a ROOT file is chunked into “baskets” allowing for partial reads, reducing RAM usage. However, for certain important use-cases, one can have a one-hour job that takes days or weeks to finish due to a fundamental choice in the size of these baskets for NanoAOD files. This problem is troubling and has never been explored, but our group here at UCSD is uniquely poised to address it. In addition to our physics work, we run one of three Tier-2 computing clusters in the United States (which services the entirety of the US CMS collaboration), and thus support a variety of expertise in high throughput computing as well as a diverse collection of modern resources. **We intend to have the EXPAND mentee(s) characterize the issue with NanoAOD in detail, and potentially provide a globally useful prescription for it.** This work will be of interest to the CMS computing community at large and could be further developed into a poster presentation for one of their conferences (e.g. ICHEP).

Skills you will acquire:

- Understanding of Big Data file structures and high-throughput computing infrastructures
- Fluency in standard programming languages (Python and C++)
- Familiarity with the command line
- Capability to perform rigorous data analysis

THE CENTER
Chemistry & Biochemistry | Mathematics | Physics

EXPAND

UC San Diego
PHYSICAL SCIENCES