Contribution ID: **402** Type: **Oral**

Nested data structures in array and SIMD frameworks

Thursday 14 March 2019 16:50 (20 minutes)

Nested data structures are critical for particle physics: it would be impossible to represent collision data as events containing arbitrarily many particles in a rectangular table (without padding or truncation, or without relational indirection). These data structures are usually constructed as class objects and arbitrary length sequences, such as vectors in C++ and lists in Python, and data analysis logic is expressed in imperative loops and conditionals. However, code expressed this way can thwart auto-vectorization in C++ and Numpy optimization in Python, and may be too explicit to automatically parallelize. We present an extension of the "array programming" model of APL, R, MATLAB, and Numpy, which expresses regular operations on large arrays in a concise syntax. Ordinarily, array programming only applies to flat arrays and rectangular tables, but we show that it can be extended to collections of arbitrary length lists ("jagged arrays"), nested records, polymorphic unions, and pointers. We have implemented such a library in Python called awkward-array, and we will show how it can be used to fit particle physics data into systems designed for Numpy data, such as Pandas (for analysis organization), Numba (for just-in-time compilation), Dask (for parallel processing), and CuPy (array programming on the GPU). We will also show how a proper set of primitives enables non-trivial analyses, such as combinatorial searches for particle candidates, in SIMD environments.

Primary author: PIVARSKI, Jim (Princeton University)

Co-author: ELMER, Peter (Princeton University (US))

Presenter: PIVARSKI, Jim (Princeton University)

Session Classification: Track 1: Computing Technology for Physics Research

Track Classification: Track 1: Computing Technology for Physics Research