Project Proposal: Leveraging ServiceX to Transform PHYSLITE Data into Flat N-tuples with Systematics

Alexander Schmidt

May 2, 2024

1 Project Summary

1.1 Problem Statement

Analysis of data from high-energy collision experiments and Monte Carlo simulations is at the core of HEP. This project focuses on access to end-user analysis of data from the ATLAS experiment in the High Luminosity LHC era. The experiment expects its main analysis data storage format to be PHYSLITE, which are a variety of ROOT files in the experiment's "xAOD" format that try to be readable as possible without the experiment's C++ I/O libraries. However, the code for determining systematic uncertainties in the experiment is written using those libraries, and is not yet ported to be usable when the data are presented to the user in a columnar format. Porting these libraries to enable columnar access is a significant challenge and a showstopper for the direct use of PHYSLITE in columnar analysis.

Several standard "ntuplers" exist in ATLAS to translate xAOD files into flat ntuples for analysis. Among other convenience items, these ntuplers compute systematic variations and store them in an easily-accessible format. In the future, it may be possible to run systematic uncertainties on the output of the ntuplers, in which case the systematic uncertainties do not need to be computed in the ntuple creation and the nuplers can be made simpler. This will allow a much gentler adoption of systematic variation code running on columnar data as the libraries become available.

The ServiceX project aims to build a general data transformation engine, with a focus on HL-LHC user data analysis. Multiple methods are being studied for how to best provide access to data stored in the PHYSLITE format, but they all leave in question how systematic uncertainties will be handled. Embedding a standard ATLAS ntupler in ServiceX is a potential route to solving this problem.

1.2 Solution

By writing and implementing a new transformer for the "TopCPToolkit" ntupler in ServiceX, the user experience when performing data analysis can be simplified. This transformer will be an alternative route through which data in the PHYSLITE format can be converted into flat ntuple data with systematic uncertainties included, allowing for more comprehensive columnar analysis. The transformer can also be adapted to other future ntuplers should they be developed.

To write and implement this transformer, there will be a number of steps. Firstly, TopCPToolkit will be setup locally. Next, it will be run in a Docker container. The interface to ServiceX (handled by an appropriate Python script) must be written and embedded in a container. This transformer will be test deployed in a functional ServiceX deployment. The proper usage of this transformer will be documented during development and made available.

2 Timeline

Time Frame	Objectives
Weeks 1-2	Familiarize with data formats, TopCPToolkit, ServiceX, etc.
Weeks 2-8	Write transformer script, and make container.
Weeks 4-12	Test results and write documentation

3 References

References

[And05] Andrew Eckart, Ben Galewsky, Rob Gardner, Lindsey Gray, Mark Neubauer, Jim Pivarski, Ilija Vukotic, Gordon Watts, Marc Weinberg, Michael Johnson. ServiceX - IRIS-HEP, 2005.

[CES$^+$23] James Catmore, Johannes Elmsheuser, Jana Schaarschmidt, Lukas Alexander Heinrich, Nurcan Ozturk, Alaettin Serhan Mete, and Nils Erik Krumnack. PHYSLITE - a new reduced common data format for ATLAS. 2023.

[Pat05] Chuck Patterson. Ntuples vs. TTrees, 2005.