# Machine Learning Inference in Athena with ONNX

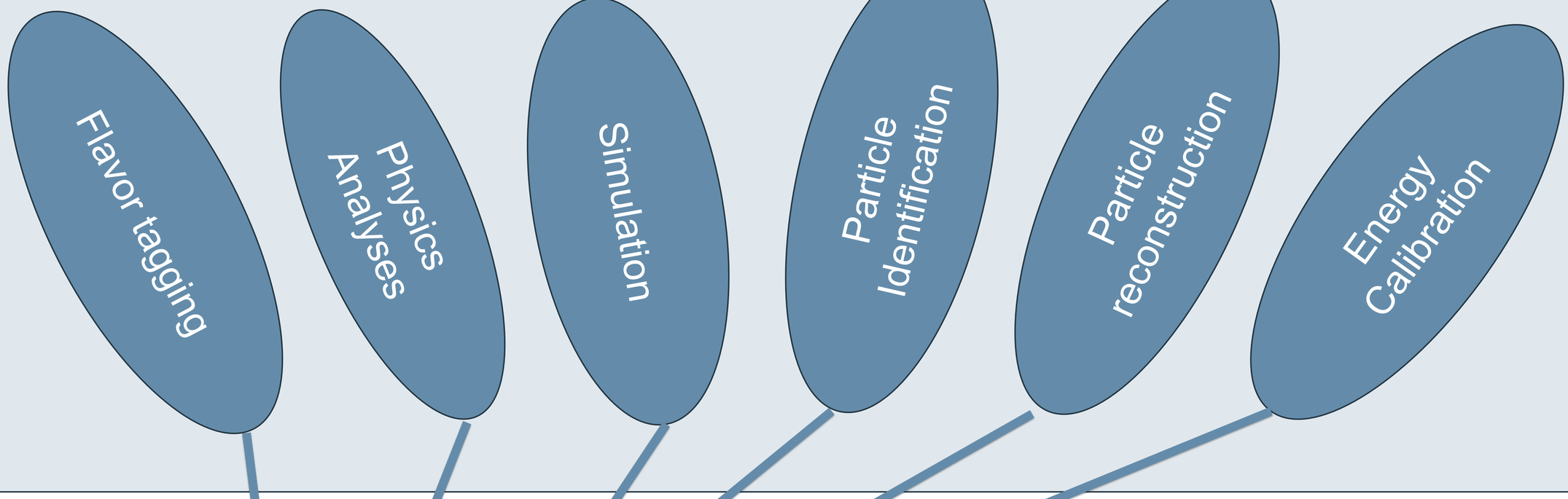
Attila Krasznahorkay[2], Beojan Stanislaus[1], Charles Leggett[1], Debottam Bakshi Gupta[3], Johannes Elmsheuser[4], Julien Esseiva[1], Paolo Calafiura[1], Vakho Tsulaia[1], Xiangyang Ju[1]

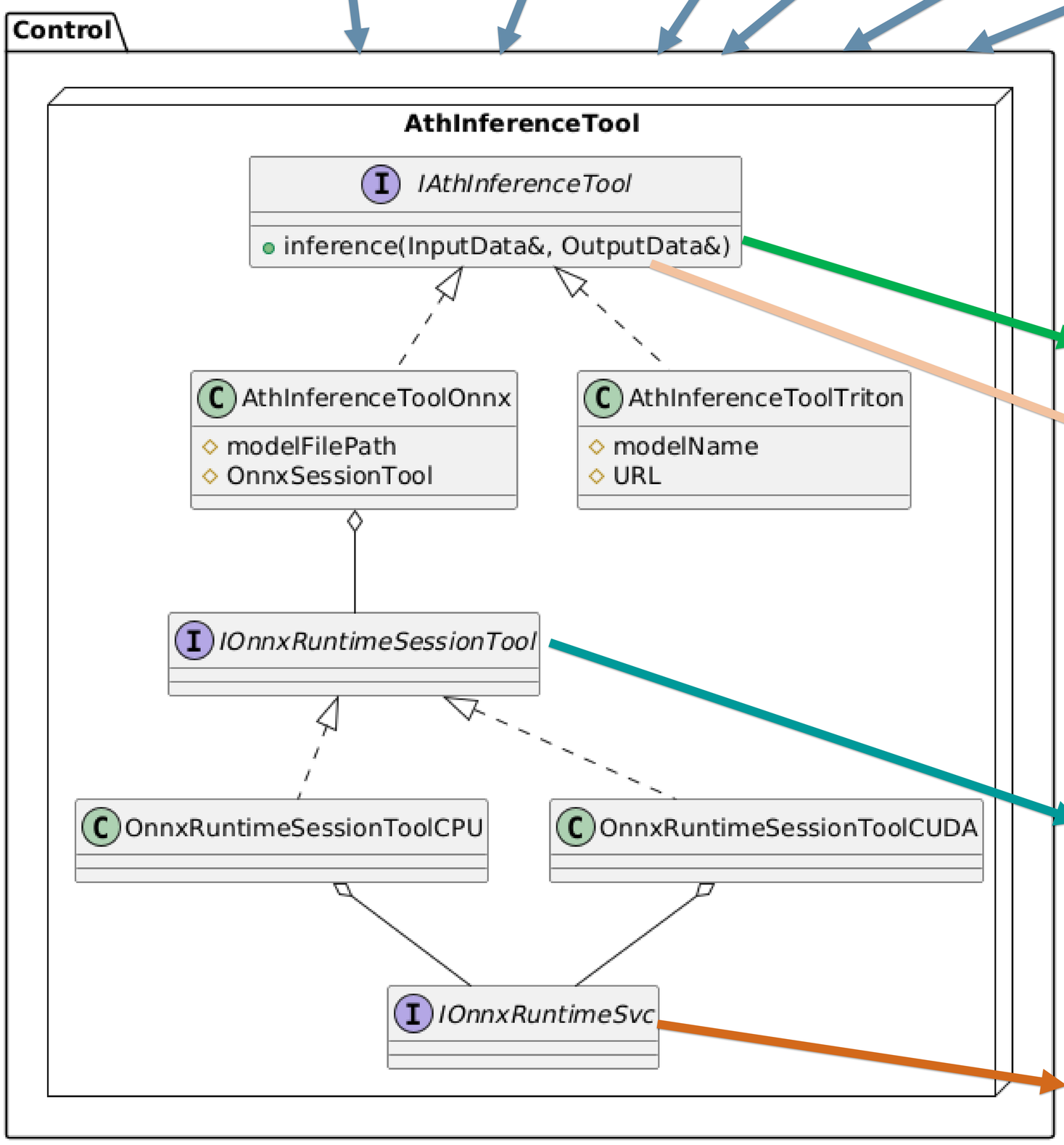
on behalf of the ATLAS computing activity

[1]Lawrence Berkeley National Laboratory, [2]University of Massachusetts Amherst,
[3]University of Texas at Arlington, [4] Brookhaven National Laboratory

27th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2024)

## ML applications in *ATLAS*




***OnnxRuntime*** [1] is a high-performance inference engine designed to accelerate machine learning models across various platforms and hardware, supporting the ONNX (Open Neural Network Exchange) format for interoperability.

***A common interface*** for Machine Learning inference in Athena that are implemented by OnnxRuntime and Inference as a Service [2].

Input and output data is a ***dictionary*** that maps input and output names to a data object that contains the tensor shape and data. The dictionary allows dynamic number of input and output data.

Tensor data is represented by ***std::variant***, allowing different data types.

Lightweight ***Athena Tool***, each for a specific backend such as CPU or NVIDIA GPU, allows the possibility that other backends can be used.

One ORT::Env as ***Athena Service***, shared by all Athena Algorithms. Its functionality includes 1) logging, 2) managing thread pools for running ONNX Sessions (To-be-done), 3) managing memory allocations for ONNX models (To-be-done)

[1]: OnnxRuntime: https://onnxruntime.ai/.
[2]: J. Duarte, et al. FPGA-accelerated machine learning inference as a service for particle physics computing, https://arxiv.org/abs/1904.08986