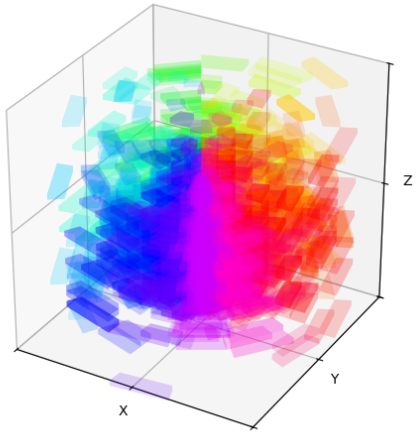
ML Inference Benchmarking & OT Flow Matching for Fast CaloSim

CERN Summer School 2024, Paul Wollenhaupt


Machine Learning Inference Benchmarking

- ML-based Reconstruction, Event Selection, Feature Extraction, Fast simulation, ...
- High cost in time and compute
- Different framworks and hardware
  - Keras, Torch, ONNX, Sofie
  - CPU (single-/multithreaded), GPU

▷ Benchmarking for performance is needed!



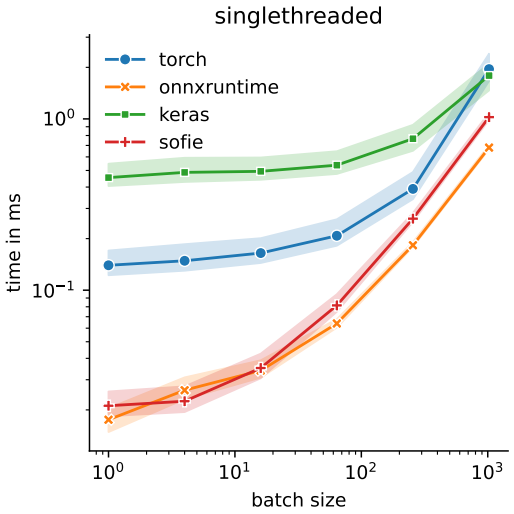
The Neural Network Frameworks

- **Torch** flexible, focused on research
- **Keras** high level API for TensorFlow
- **SOFIE** ML inference codegen in ROOT
- **ONNX** portable format for ML models

▷ ... many more

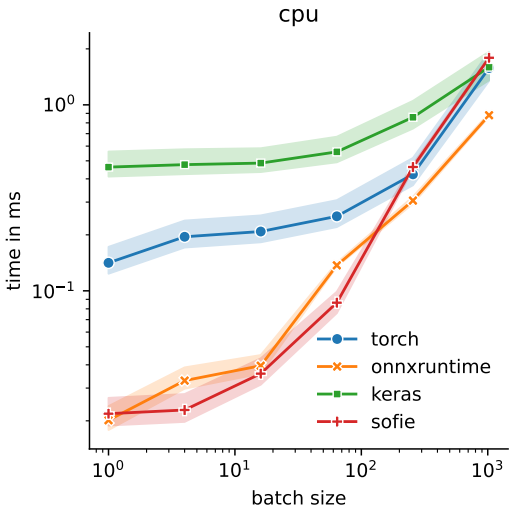
MLP - Singlethreaded CPU Results

- Keras and Torch are slowest
- SOFIE and ONNX comparable
- ONNX better for high batch sizes
- ▷ ONNX best choice for CPU inference



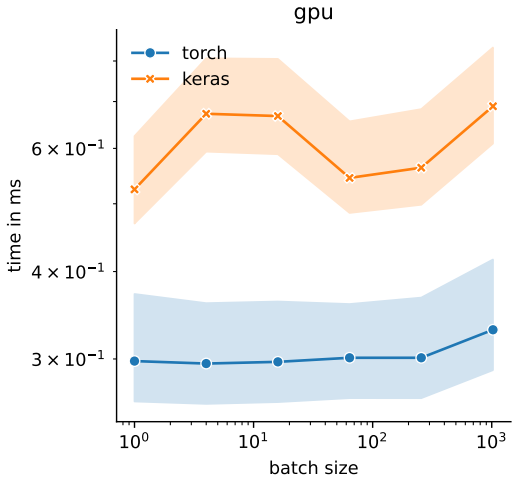
MLP - Multithreaded CPU Results

- ONNX seems to oddly underperform
- Similar to singlethreaded
- For large enough batch sizes keras and torch are competitive



MLP - GPU Results

- No results for SOFIE and ONNX
- GPU scales better with batch size
- Keras and Torch are close
- Torch is consistently slightly faster

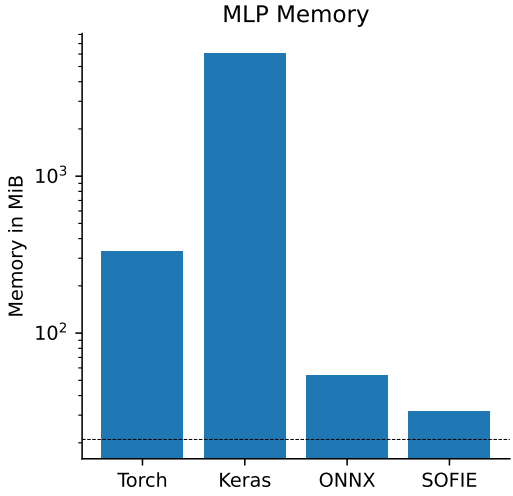


Python Implementation Details

- Load all models from ONNX files
- Convert ONNX models to PyTorch and Keras
- Get SOFIE model from the Keras model
- Set threading options for singlethreaded case
- Synchronize after each inference call
- Some options can only be set once per execution
  - ▷ Python file with arguments run from bash script

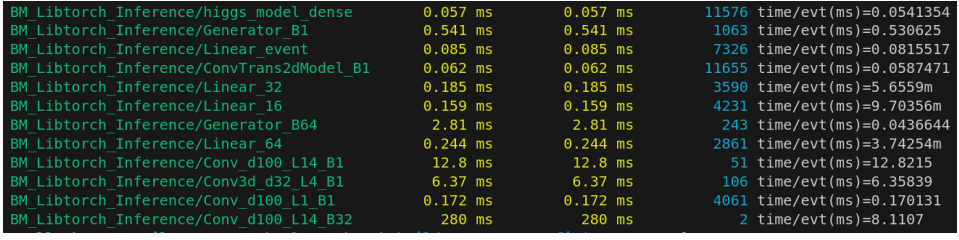
Memory Benchmark

- Memory profiling using `memray`
- Traces and records all function calls
- Handles calls from python to C/C++
- Keras and Torch have unreasonably high memory usage
- ONNX and SOFIE are in OOM of the theoretical minimum
- SOFIE has the lowest memory usage



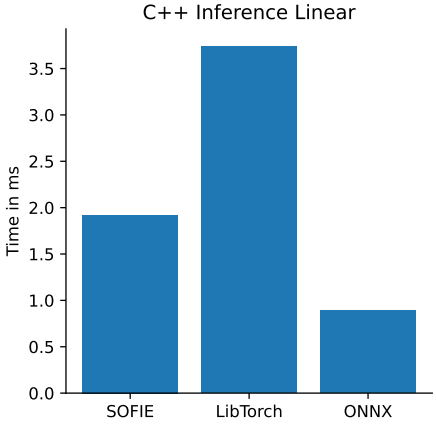
Rootbench - ROOT Benchmarking Tool

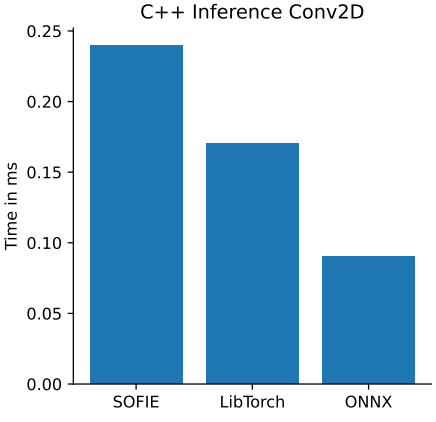
- ROOT benchmarking tool based on Google Benchmark
- ONNX model inferenec with ONNXruntime and SOFIE
- Added support for inference in LibTorch for some models
- Python script converts ONNX to TorchScript



```
BM_Libtorch_Inference/higgs_model_dense        0.057 ms        0.057 ms        11576 time/evt(ms)=0.0541354
BM_Libtorch_Inference/Generator_B1              0.541 ms        0.541 ms         1063 time/evt(ms)=0.530625
BM_Libtorch_Inference/Linear_event              0.085 ms        0.085 ms         7326 time/evt(ms)=0.0815517
BM_Libtorch_Inference/ConvTrans2dModel_B1       0.062 ms        0.062 ms        11655 time/evt(ms)=0.0587471
BM_Libtorch_Inference/Linear_32                 0.185 ms        0.185 ms         3590 time/evt(ms)=5.6559m
BM_Libtorch_Inference/Linear_16                 0.159 ms        0.159 ms         4231 time/evt(ms)=9.70356m
BM_Libtorch_Inference/Generator_B64               2.81 ms         2.81 ms          243 time/evt(ms)=0.0436644
BM_Libtorch_Inference/Linear_64                 0.244 ms        0.244 ms         2861 time/evt(ms)=3.74254m
BM_Libtorch_Inference/Conv_d100_L14_B1           12.8 ms         12.8 ms           51 time/evt(ms)=12.8215
BM_Libtorch_Inference/Conv3d_d32_L4_B1           6.37 ms         6.37 ms          106 time/evt(ms)=6.35839
BM_Libtorch_Inference/Conv_d100_L1_B1           0.172 ms        0.172 ms         4061 time/evt(ms)=0.170131
BM_Libtorch_Inference/Conv_d100_L14_B32           280 ms          280 ms            2 time/evt(ms)=8.1107
```

Rootbench - Results





Results on CaloSim coming on monday!