SML
DRAFT Parallel Session Report

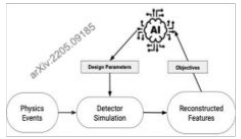
Paolo Calafiura, Walter Hopkins

LBL AHM
July 24, 2024

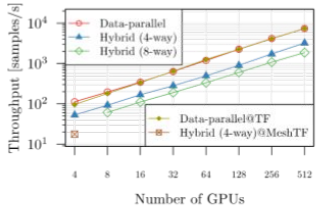
What have been the two most significant Year 1 accomplishments so far?

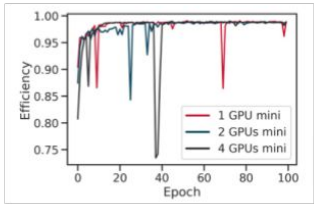
HEP-CCE

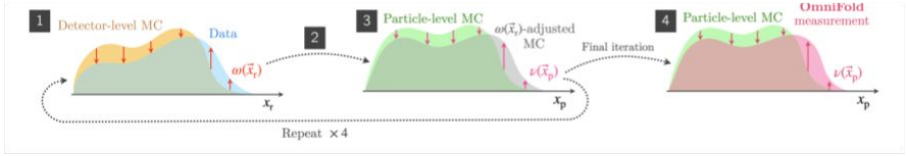
Highlights from SML bi-weekly meetings



- Wen Guan: HPO and detector optimization using iDDS (slides)
- Peter Nugent: Introduction to LBANN (slides)
- Alina Lazar: Scaling up Distributed GNN Training (slides)
- Ben Nachman: Unbinned Unfolding with Omnifold (slides)
- Aishik, Jay, and Rafael:
  Neural Simulation Based Inference on HPCs (notes)

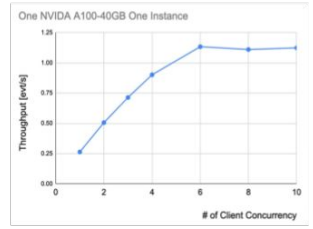






HEP-CCE

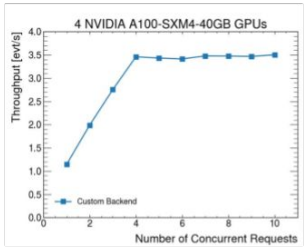
Year 1: Inference as a Service (IaaS)



Why is it useful?
- Clean interface to abstract host/device communication, even across WAN (study performance impact, usability, security on DOE HPC systems)
- Fully utilize GPU
- Scale out to multiple GPU and nodes

What is SML role?
- Build on existing IaaS efforts (including Exa.TrkX and ACTS as a Service)
- Deploy IaaS on DOE HPCs (Perlmutter as a start)

Collaborate with **PAW** on optimization&portability?



What are the focus areas of the work in your technical area right now?

Continue IaaS Studies

- Explore synergies with SONIC work at NERSC NESAP

Model Selection

- Representative, benefit from scaling, stakeholder involvement

HEP FASST Whitepaper

- Present HEP as a use case
  - Open Data
  - (Towards) HEP Foundation Models

Open Data Discussion triggered by FASST white paper

FASST asked for models training on “billions of tokens”

For each domain, estimate time for 'Frontier model' ready training data at 1 Billion tokens (then 10 Billion tokens, etc.). 'Frontier model' training ready data means a representation that is known to work in training models either through small scale experiments, fine-tuning experiments on existing models, RAG experiments or some other direct evidence that the representation works.

GNN4ITk (supervised GNN) training on ~1B hits

10K G4 evts (ATLAS closed), OR Open Data Detector. Could grow to ~10B hits

OmniLearn (supervised foundation) training on ~1B jets

Jet Class dataset (100M Delphes events) could easily grow 10-100x

Michael K wants >**1B G4 “FCC-ee” events** to train generic classifier.

Would require ~100M core hours. HL-LHC 10-100x more

ATLAS released 7B+2B Run 2 events (65 TB PHYSLITE ~columnar format)

More HEP foundation models for FASST white paper

Xiangyang's tracking Language Model (see also Track-formers)
OmniLearn jet generative+ classifier (supervised)
Michael's pre-training foundation model (detector-level. self-supervised)
Kazu's HEP simulation

What are the two or three main priorities for the next 18 months

1. **Complete performance studies of Distributed Training (and Inference) for HEP GNNs.**
   a. Publish lessons learned, possibly “HOW-TO”.
2. **Automate ensemble model development for SBI unbinned analyses**
   a. Submission of O(10K) training jobs
   b. Evaluation of single model performance, class performance, ensemble performance
   c. (stretch goal) As inner loop for HPO
3. **Explore growing landscape of HEP foundation models**