Theory Needs

FNAL AI Infrastructure Planning

George T. Fleming, Apr. 6, 2023

Current Theory AIML R&D Efforts

- Monte Carlo event generation
- Numerical integration
- Density estimation:
  - Non-parametric likelihood functions and Bayesian posterior probabilities.
  - Anomaly detection (finding events that don’t look like background).
- AIML for generating ensembles for lattice quantum field theory. Currently, equivariant normalizing flows are popular but other architectures under investigation.
- A current theme in a lot of the research at FNAL is applications of normalizing flows.
- Researchers: Jim Simone, GTF, Josh Isaacson, Dan Hackett (starting Fall ’23)

Workflow for Monte Carlo Event Generators

Josh Isaacson

- Current AIML algorithmic development proceeding on Wilson cluster on single GPUs. Competition for scarce A100 GPU’s can be a problem. Access to more A100 or H100 GPUs for rapid development would be helpful.

- As algorithms advance beyond proof-of-concept stage, some training of some larger neural systems may need to be moved to NERSC or Aurora.

- In a production environment, cost of generating events will be borne by experimental collaborations using the production code.

- Possibility of theorists pre-generating a library of hard interactions using a large cluster and distributing to experimentalists.

Workflow for Lattice Generation

Jim Simone, GTF, Dan Hackett

- Current R&D efforts focus on scaling of algorithms that work generating small lattices on single GPU's to large lattices on 100's-1000's of GPUs.
- Need rapid turn around on undersubscribed local clusters on jobs ranging from single GPU's to 10-100's of GPU's with fast interconnects to test scaling.
- Once algorithms are stable and shown to scale, NERSC or INCITE/ALCC time can be requested for generating large ensembles (year timescale).
- What is a large lattice? 200B quad-prec numbers. Min job size (Frontier): 8x184=1472 GPUs
- Medium-scale ensembles can be generated using allocated time on USQCD clusters (e.g. LQ2).
- Architectures that work best for AIML are same as what works best for standard LQFT codes (fast half/…/quad precision, fast interconnect, large memory a plus). Ultimately AIML will likely augment standard algorithms as learned components to accelerate algorithms.

Data Management for Lattice Generation

- In algorithmic development phase, very large ensembles of relatively small lattices need to be generated for high-precision comparisons of ML flows with standard algorithms: need for O(100) TB of short term scratch space.
- Currently, training data is generated on the fly, no need for fast loading of pre-generated data. May change in future.
- Similar to how lattices are stored now, trained models will become a valuable resource worth storing long-term.
- Having a centralized database of trained models accessible (read/write) by compute resources very useful during R&D:
  - Omniboard (https://github.com/vivekratnavel/omniboard)
  - Weights and Biases (https://wandb.ai/site/dashboard)