Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

AI/ML Directed and Facility Integrated Informational Distillation and Feature Extraction for High Throughput Streaming DAQ for EIC Detector 2

Jin Huang (PO)

6/1/2022

@BrookhavenLab

FY2023 NPP LDRD Type A Proposal

Proposal title: AI/ML Directed and Facility Integrated Informational Distillation and Feature Extraction for High Throughput Streaming DAQ for EIC Detector 2

Primary Investigator: **Jin Huang (PO)**

Other Investigators:

Yihui Ren, Yi Huang, Shinjae Yoo (CSI/ML)
ByungJun Yoon (CSI/Math)
Adolfy Hoisie (CSI/ACL)
Chris Pinkenburg, Martin Purschke (PO/sPHENIX)
Torre Wenaus (PO/NPPS)

Indicate if this is a cross-directorate proposal. Yes _X__ No___
If yes, identify other directorates/organizations: **CSI, with strong support**

Program: **NP**

Proposal Term: 3 year **From: Oct 2022 To: Sept 2025**

Total funding per year in FY23, FY24 and FY25: **300k, 500k, 500k**

Brookhaven National Laboratory

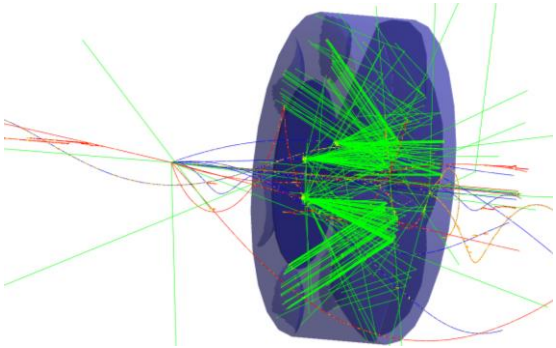
Targeted Topical Areas in FY23-A call

- Research and Development towards the Second Detector at the Electron-Ion Collider
- Discovery Science Driven by Human-AI-Facility Integration
  - 1) AI enhanced Detectors, Accelerators and Sensors

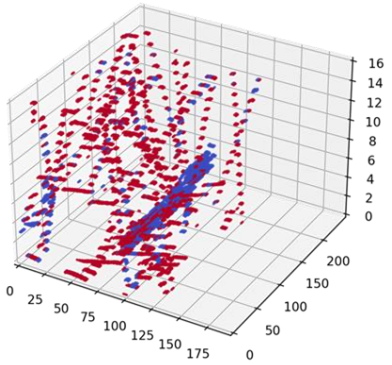
Motivation

- FELIX-type Streaming DAQ is well accepted in EIC (CDR, three detector proposals), reasonable to assume so for EIC detector 2
- A unique challenge to NP streaming DAQ: keep all collision signal while reliably filter out background and noise
  - Driven by physics requirement of low bias, and discovery potential
  - Especially so for detector 2, possibly operating at higher luminosity and background
- Streaming data from various subsystem can be presented to an AI algorithm in a consistent manner
  - Zero-suppressed 3-4 dimension sparse matrix of time-framed data, containing patterns of signal and background + noise
- Technical requirement
  - Requirement for high throughput: O(1)Tbps collision signal after zero suppression
  - Stringent bias control (systematical uncertainty control for)
  - Not much requirement for low latency (unlike triggering in LHC)

EIC SIDIS event on a dRICH detector, whose data can be presented in 3D zero-suppressed sparse matrix data frames of pixel X-Y, and time bin



sPHENIX TPC streaming data frame, also presented in 3D zero-suppressed sparse matrix (R-phi-time) marked with signal/background hits



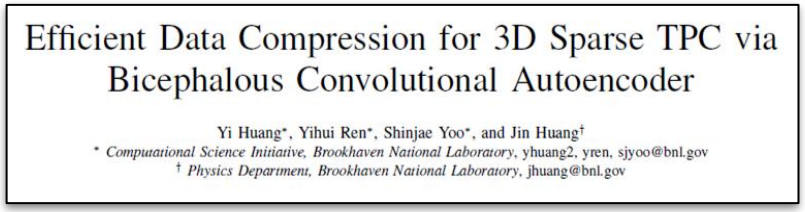
Brookhaven National Laboratory

Our methods and deliverables

- Algorithm based on bicephalous (and multi-headed) convolutional autoencoder
  - Data reduction/information distillation: noise filtering, feature extraction, and lossy compression;
  - Provide uncertainty quantification and real-time robustness-aware
- Test and deploy using cutting-edge AI accelerators
  - Non-von Neumann Architecture processor optimized for efficient and high throughput computing for large scale neutral networks
  - On the cusp of commercially availability; undergoing testing via our vendor relationships;
  - Promising to meet timeline for conceptual design of EIC D2
- Deliverables
  - AI-based data reduction and knowledge distillation algorithm designed for high-throughput real-time inference
  - Robust AI models with uncertainty quantification, out-of-distribution detection and periodical validation
  - Algorithm performance evaluation integrated through simulation, raw data, and reconstruction
  - Demonstrate throughput on multiple novel dataflow AI accelerators
- Budget request mainly for two postdocs (PO and CSI each) support: 300k, 500k, 500k

Brookhaven National Laboratory

Initial exploration on compression stage alone published in IEEE ICMLA
arXiv:2111.05423
https://github.com/BNL-DAQ-LDRD/NeuralCompression

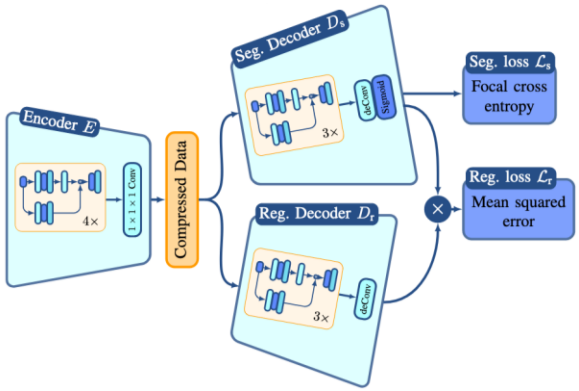

Efficient Data Compression for 3D Sparse TPC via Bicephalous Convolutional Autoencoder

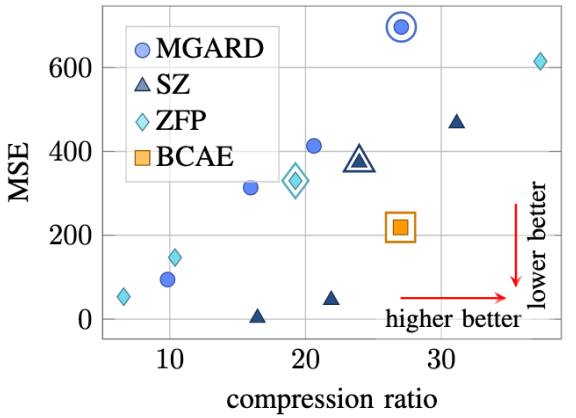
Yi Huang*, Yihui Ren*, Shinjae Yoo*, and Jin Huang†
* Computational Science Initiative, Brookhaven National Laboratory, yhuang2, yren, sjyoo@bnl.gov
† Physics Department, Brookhaven National Laboratory, jhuang@bnl.gov

Intellectual merit 1
Strategy of the algorithm design

- Neural network-based data compression and information distillation has better inference throughput, reconstruction fidelity and leading-edge hardware support.
- Fixed latency execution, favored by real-time computing application at EIC
- Extend the model to take sparse-encoding data directly to mitigate communication bottleneck.
- Extend to probabilistic models for uncertainty quantification and real-time robustness-aware.
- Branch-free and dataflow hardware-friendly neural network optimization for fast deployment on and integration with advanced hardware

IEEE ICMLA, arXiv:2111.05423





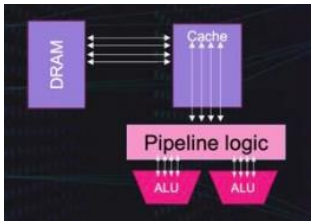
Brookhaven National Laboratory

Intellectual merit 2 - Why Dataflow AI accelerators?

- Comparing with von Neumann Architectures (CPU/GPU), Dataflow processors are
  - Designed for NN computing
  - Massive on-chip activation/weight storage on sRAM
  - Good integration with popular AI tools
  - Energy efficient and high throughput
- Two family of accelerators under eval and benchmarked against GPUs:
  - Dataflow architecture digital processors, such as GraphCore IPUs
    - GraphCore tested our network [arXiv:2111.05423] on 2nd gen IPU showing x(10) improvement of speed comparing to GPUs (A6000/LHCb Alan like)
  - Electro-phontoncs processor, such as LightMatter Envise
    - On-going, network and data shared

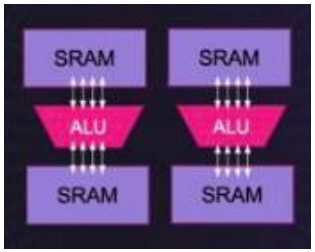
CPU/GPU

von Neumann Architectures

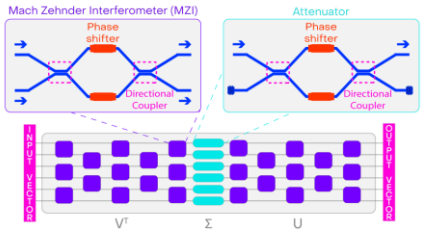


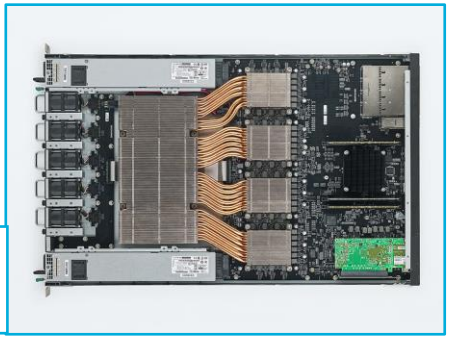
IPU

Dataflow Architectures



Electro-Photonics





Brookhaven National Laboratory

Our recent test based on second generation IPU from GraphCore, TSMC 7nm, 1GB sRAM
See also LHCb sim/reco tests arXiv:2008.09210

Why this team? Why now?

- The team
  - Jin Huang (PO/sPHENIX) + Postdoc hiring: Domain integration and evaluation. Co-convener EIC Detector1 global integration; manager-sPHENIX TPC readout; expertise in simulation and performance evaluation and HF physics
  - Yihui Ren, Yi Huang, Shinjae Yoo (CSI/ML): Lead AI/ML algorithm development on noise reduction, data compression, and network optimization.
  - ByungJun Yoon (CSI/Math): AI Uncertainty Quantification; PI for ASCR award for data reduction
  - Adolfy Hoisie (CSI/ACL): advisor, leading Advanced Computing Lab (ACL), connection to cutting-edge hardware vendors
  - Chris Pinkenburg, Martin Purschke (PO/sPHENIX): advisor; coordinators for sPHENIX computing and DAQ
  - Torre Wenaus (PO/NPPS): advisor; co-convener EIC UG software group, leadership roles in NP/HEP computing
- Productive team
  - Five invited talks since starting last year
  - First paper published in IEEE ICMLA: arXiv:2111.05423
  - Coming talk/paper accepted: IEEE RealTime 2022

Why fund now?

- Keep and expand the momentum for the research, currently supported under LDRD19-028 (ending in one month)
- Keep and enhance the connection with novel hardware vendors, who are testing our network/data
- At the end of LDRD: bring real-time application of dataflow AI accelerator to the maturity that meets the conceptual design stage of EIC detector 2
- Address the FY23 LDRD-A topical areas of Human-AI-Facility Integration and EIC detector-2

Return of investment

- Demonstrate a working prototype, as a ladder to external FOA for construction of production system for
  - EIC Detector 2 construction project for online computing ($O(10) M)
  - Opportunistic upgrade of EIC Detector 1 and sPHENIX (few M$)
- Saving in tape storage and offline computing need for reaching same physics goals $O(1M)

Broader impact for BNL

- Building expertise on AI/ML directed data reduction algorithm and the corresponding high throughout computing system
- Directly supporting and enhancing high priority BNL research in nuclear physics
- Direct support for the BNL AI strategy, in particular the components on real-time experimental application, and Human-AI-Facility Integration
- Testbed with AI accelerators for broad use cases at BNL
- Put BNL at a leadership role in application of novel AI accelerator in real-time NP/HEP experiment (next slides)

BNL positioned to lead novel AI-accelerator in NP/HEP real-time computing

- sPHENIX : data stream most resembles EIC
  - Streaming readout for full tracking system
  - 60M charge particle per second ~ EIC upgraded to 10^{35} /cm/s << LHC
  - TPC present high throughput data as 2+1D zero suppressed data stream, a generic data representation for streaming detectors from algorithm development point of view
- BNL leads recent ASCR award on data reduction
  - Focusing on generic algorithms and uncertainty quantification, with early application focus on cryo-EM and climate science applications.
  - Synergy to our application at EIC
- Connection to AI Hardware vendors through Advanced Computing Lab @ BNL
  - Connection to multiple vendors with wide range of technology maturity. (details availability under NDA)
  - Allow for evaluation of our algorithm on multiple vendor hardware

Summary

- Neural network-based data compression and information distillation with high throughput, reconstruction fidelity, and aimed to run on leading-edge AI hardware.
- Addresses real-time computing need for EIC Detector-2 streaming DAQ
- Backed by a productive team: multiple invited talks and publications
- Synergize well with multiple CSI research activities
- Position BNL to lead novel AI-accelerator in NP/HEP real-time computing
- Deliverables :
  - AI-based data reduction and knowledge distillation algorithm designed for high-throughput real-time inference
  - Robust AI models with uncertainty quantification, out-of-distribution detection and periodical validation
  - Algorithm performance evaluation integrated through simulation, raw data, and reconstruction
  - Demonstrate throughput on multiple novel dataflow AI accelerators
- Budget request mainly for two postdocs (PO and CSI each) support: 300k, 500k, 500k