|  |  |  |  |
|---|---|---|---|
| **PRINCIPAL INVESTIGATOR** | **Jin Huang (PO)** | **PHONE** | **x5898** |
| **DEPARTMENT/DIVISION** | **PO/NPP** | **DATE** | **June 2022** |
| **OTHER INVESTIGATORS** | **Yihui Ren, Yi Huang, Shinjae Yoo (CSI/ML)**<br>**ByungJun Yoon (CSI/Math), Adolfy Hoisie (CSI/ACL)**<br>**Chris Pinkenburg, Martin Purschke (PO/sPHENIX)**<br>**Torre Wenaus (PO/NPPS)** |  |  |
| **TITLE OF PROPOSAL** | **Real-time Information Distillation on Novel AI Hardware** |  |  |
| **TYPE A** |  |  |  |
| **PROPOSAL TERM (month/year)** | From **Oct 2022** | Through | **Sept 2025** |

## SUMMARY OF PROPOSAL

**Description of Project:**

Nuclear science requires diverse experimental data with stringent systematic uncertainty control. These scientific requirements drive the need for streaming data acquisition (DAQ) in the next generation of nuclear experiments such as the detectors at the Electron Ion Collider (EIC). Distinct from most existing experiments, these new experiments demand a high throughput data reduction in real-time which preserves all collision signal information while filtering out background and noise reliably. In this proposal, we will address this unique challenge by building a set of AI-based information distillation algorithms that perform noise filtering, feature building, and compression. The algorithms will be co-designed with and demonstrated on a selection of the emerging AI hardware, the first in the field.

Compared to traditional methods, our approach is constructed to utilize novel AI hardware, resulting in high throughput and energy efficiency. The approach is designed to be robust with built-in safeguard metrics communicating with human operators. BNL is well positioned to be a world-leader in such research as the host institution of the EIC with the added knowledge of the leading-edge AI algorithms and hardware through CSI. The proposed work will be carried out by a productive team with established publication and invited talk records in this area. This work directly advances the high-priority experimental programs in the 2015 NSAC Long-Range Plan.

It is also timely to fund this research to demonstrate a prototype with the first and only streaming DAQ run at RHIC in 2024, to influence the conceptual and technical design of both of the EIC detectors, and to keep the momentum for research at BNL and for collaboration with hardware vendors (support letters appended). This proposal addresses the FY23 LDRD-A topical areas of *EIC Detector-2* and *Human-AI-Facility Integration*.

**Expected Results:**

We will demonstrate the fidelity and throughput of AI-based information distillation algorithms on testbeds of novel AI hardware using simulated EIC data and real data collected from streaming DAQs in Relativistic Heavy Ion Collider. Comparing with the traditional systems, our concept has the potential to improve recorded physics events and throughput by one order of magnitude, as indicated by early tests. This work aims to position BNL to lead the research on utilizing novel AI hardware in real-time computing. Our uncertainty/competence-aware AI will significantly enhance the trustworthiness of the AI predictions, paving the way towards efficient human-AI-facility integration for EIC and beyond. And the topic of AI in experimental nuclear science has been one of the new funding areas at DOE, to which we will target future external funding.

# PROPOSAL

## 1 Introduction

Modern large-scale nuclear physics (NP) experiments in high-energy particle colliders use streaming-readout electronics to readout detector responses at $O(10)$ Tbps bandwidth. Prominent examples at BNL include the sPHENIX experiment at the Relativistic Heavy Ion Collider (RHIC) [1], now close to completion, and the experiments of the Electron-Ion Collider (EIC) [2,3], planned for the 2030s. One of the main challenges for these streaming-readout systems is to manage the data rate with sufficient data reduction in real-time, so the end data fit persistent storage for offline reconstruction, which typically is $O(100)$ times reduced to $O(100)$ Gbps [1, 4]. Such a reduction is traditionally achieved via triggering that save a small subset of collisions of interest. Although triggering is applicable to most collider experiments, it is insufficient for the next-generation NP experiments that study diverse collision topologies [2, 4, 5].

This challenge opens up an opportunity and necessity for the utilization of an AI-directed information distillation algorithm co-designed with novel AI hardware for real-time data reduction, that performs reliable noise filtering, feature extraction, and lossy compression in an integrated step. Despite the distinct signal and background features of various detectors and experiments, a commonality for the input data stream is they can be formatted as sparse-encoded 3D tensor of time frames on the FELIX streaming DAQ [1, 4, 8] as illustrated in Figure 1. Therefore, it is suitable for a common algorithm-hardware architecture to address all detectors as in this proposal. And in contrast to traditional scientific compression algorithms [9–11], using data-point-level precision as the criterion, AI-directed algorithm can achieve a greater compression ratio and run faster than traditional compressors as demonstrated by our exploratory work [12].

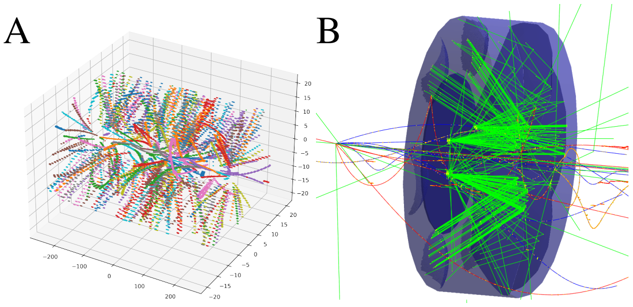


Figure 1: We propose to use a common AI algorithm-hardware architecture to process a wide array of detectors' data, as they can be formulated as sparse-encoded 3D tensor of time frames with distinct patterns of signal and background. A: simulated sPHENIX TPC data frame in $x$-$y$-time with 3 MHz $p$+$p$ collisions [1]; B: EIC dRICH detector with an $e$+$p$ collision [6, 7] whose data frame is a tensor of 2D(pixel)+time.

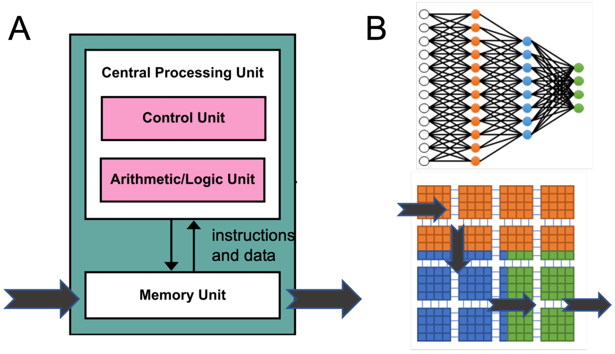


Figure 2: A. von Neumann Architecture. B. Dataflow Architecture [13]. Computing units (CU) (small grid cells) form a grid network (Cerebras) and each CU is equipped with a local memory unit. A three-layer neural network can be mapped to different regions of computing grid.

Traditional von Neumann architectures, such as CPUs and GPUs, feature a memory hierarchy consisting of a large memory space (typically DRAM) and layers of caches (SRAM), Figure 2A. The latency of data access from memory or disk is orders of magnitude higher compared when data resides in caches. Delayed data availability at computing units is known as the "memory wall" [14]. In addition to the performance penalty, the energy cost of data movement across the memory hierarchy is much higher than that of arithmetic [15, 16]. Innovative computer architectures are being implemented that mitigate these inefficiencies for AI algorithms. For example, GraphCore [17], Cerebras [18], SambaNova [19], WaveComp [20], Untether [21] and Cambricon [22] developed the so-called "AI Chips" utilizing the dataflow architecture [23, 24], where in lieu of a global DRAM memory, each computation unit is accompanied by a local SRAM memory to achieve data locality. For neural networks computation, it means that the model weights stay close to the arithmetic units, and access to intermediary values (activations) is optimized for a logicalto-physical mapping that could be pre-compiled, as shown in Figure 2B. Optimized data movement also improves energy efficiency [25] and throughput performance [26, 27]. Provision of python packages compatible to popular AI software reduces the amount of efforts to port existing AI models to make use of such novel hardware.

**Project Objectives and Innovations.** We will demonstrate the fidelity and throughput of AI-based information distillation algorithms on testbeds of novel AI hardware using simulated (EIC) and real data (demonstration with sPHENIX data) from the streaming DAQs. Comparing with the traditional systems, our concept has the potential to improve recorded physics events and throughput by one order of magnitude, as indicated by early tests. The proposed work is innovative in multiple aspects: (1) The first application of a dataflow architecture AI processor in real-time data processing in collider experiments. (2) Inventing a set of novel AI-directed information distillation algorithms that is specifically designed to address the challenge of the NP experiments. (3) Our uncertainty/competence-aware AI will significantly enhance the trustworthiness of the AI predictions, paving the way towards efficient human-AI-facility integration for EIC and beyond.

We expect the above innovations will impact the online computing model for the nuclear experiments, in particular for the EIC. This work will also provide testbeds of the selected AI hardware to cross pollinate with other areas of AI development at BNL.

## 2 Innovative AI Algorithm for Reliable Real-time Information Distillation

**AI algorithms for compression and background filtering.** There are three unique problems we need to address when designing an AI compression and noise filtering algorithm for collision data: (1) streaming front-end applies a zero-suppression that squashes small non-zero values and induces a bimodal data distribution; (2) the desired compression ratio and throughput greatly exceeds that any sparse-coding or off-the-shelf compression algorithms can provide. (3) background hits can make large-scale patterns (e.g. highly curved delta-electron in TPC). To address these challenges, we proposed an autoencoder compressor called Bicephalous Convolutional Autoencoder (BCAE) as initially explored in Ref. [12], that extends a standard autoencoder network with bi-headed decoders (Figure 3A). The segmentation head $D_s$ differentiates between signal and background (including both zero and noise) and the regression head $D_r$ focuses on approximating ADC values for signal. To test BCAE's ability to address the bimodal data distribution, a BCAE with compression ratio 27 was designed for sPHENIX TPC data with Au+Au collisions. BCAE achieves a lower overall mean-squared error (MSE) than non-AI based compressors at a similar or higher compression ratio (See Figure 3B). To test BCAE's potential on noise-filtering, another BCAE with compression ratio 204 is designed for the TPC data with $p$+$p$ collisions as illustrated in Figure 1A. The threshold of the segmentation output can be adjusted to balance zero suppression, noise filtering, and signal preserving rates to meet the requirement for downstream applications (Figure 3C).

As an alternative to BCAE, which works well for high occupancy data (such as TPC in Au+Au collisions and calorimeter data), we will further explore Graph Neural Networks (GNN) [28] in this proposal to address the need of handling more sparse data. GNNs have been used extensively in particle physics [29–32] because they work with sparse data by default, but have rarely been studied for use with a TPC or PID detectors. With

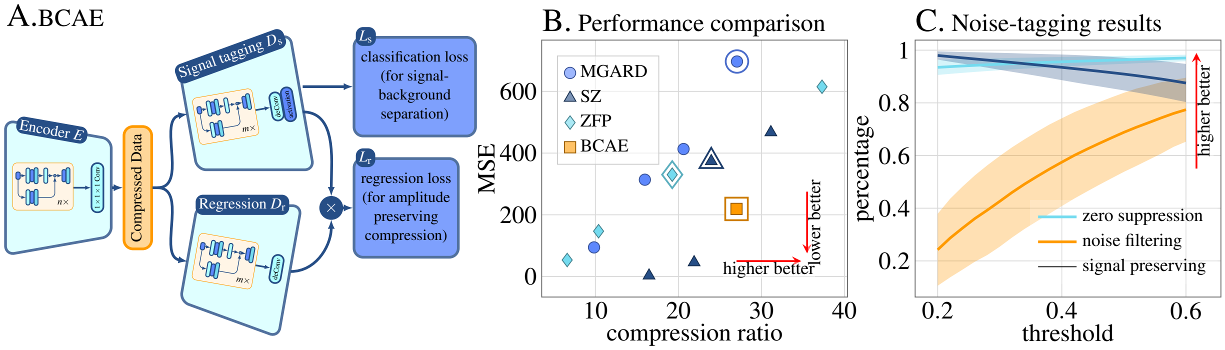


Figure 3: A. BCAE architecture. B. Its compression benchmark as in our feasibility study [12] C. Zero-suppression, noise-filtering, and signal preserving as functions of segmentation output threshold.

respect to the challenges mentioned above: (1) since GNN operates on nonzero bins only, it bypasses the bimodal distribution caused by zero-suppression; (2) we can train a GNN to extract trajectories (a sequence of locations that a particle most likely traversed TPC layers) from a cluster and hence compress it to few data points; (3) we can better utilize the global features of noise and differentiate noise from signal at graph level for efficient noise removal. Furthermore, lower occupancy implies that each tracklet within a detector sector (Figure 1A), form a small connected graph; and all tracklets can be processed in parallel.

**Enabling uncertainty quantification and competence-aware.** In order to design novel AI-algorithms that can be deployed in real-world scientific applications, it is critical to develop capabilities for quantifying the uncertainties in the predictions made by AI models and to effectively inform and assist domain experts to make robust uncertainty-aware decisions.

For this purpose, we plan to take a Bayesian approach for assessing the uncertainty of the predictions made by the proposed deep neural networks (DNNs). As uncertainty quantification (UQ) in deep learning models [33] can be computationally costly, we will consider several alternatives, compare their relative accuracy and computational efficiency, and select the ideal scheme for our AI algorithms to strike the right balance between accuracy and efficiency. More specifically, in this project, we will consider the following schemes for UQ. The first approach is based on Monte Carlo (MC) dropout technique, which randomly drops units in the trained model to estimate its impact on the prediction output, thereby aiming at quantifying the uncertainty in the prediction [34]. The second approach is to take an ensemble of DNNs, trained using different random seeds and/or training data, where the prediction variability across the DNNs in the ensemble can be used to quantify the uncertainty [35]. Finally, we may use the Bayesian last layer (BLL) approach [36, 37], where the feature space in the last layer of a trained deep network model is used to build a linear Bayesian model for uncertainty quantification. The quantified uncertainty from our UQ scheme will make our AI models aware of its competence and uncertainty. This will inform the human operators whether the proposed AI algorithms are reliably operating within "normal" operating conditions for which the AI model has been trained. If the prediction uncertainty exceeds (or the confidence level drops below) the specified tolerance, our AI algorithms will immediately raise a flag and notify the expert operator. Our uncertainty/competence-aware AI will significantly enhance the trustworthiness of the AI predictions, paving the way towards efficient human-AI-facility integration for EIC, RHIC, and beyond.

Furthermore, we will investigate and develop strategies for improve our AI model to expand its capability to handle "out-of-distribution" data. To this end, we will adopt two data-efficient learning strategies [38], which are complementary to each other: a coreset selection strategy for semi-supervised learning and a Bayesian active learning strategy. Coresets [39] are small representative subsets of original data that approximates their characteristics, and effective coreset selection can lead to efficient training of deep learning models. In this project, we are especially interested in coreset selection for semi-supervised learning of our DNN to improve its performance for out-of-distribution data. While this coreset approach enables model improvement through the effective (and computationally efficient) use of unlabeled data, the Bayesian active learning (BAL) approach [40–42] aims to generate additional "smart" data to improve model performance. For this purpose, we will utilize the uncertainty-aware BAL scheme developed by our team [40–42].

## 3 AI algorithm Co-Design with Novel AI-computing Accelerator Hardware

Traditional numerically-intensive scientific computation workloads, such as Kalman filter tracking algorithm, exhibit data movement patterns and pose challenges different from those of deep neural networks (DNNs). DNN models consist a sequence of neural network layers, each operating on small size matrices. This requires frequent model weight loading and activations (intermediary layer outputs) caching, leading to higher latency and energy consumption on von Neumman architecture (Figure 2A). New AI-oriented hardware designs are based on a Dataflow architecture concept [43] (Figure 2B), which enables a much higher degree of parallelism at data level. The data parallel computation graph is then mapped onto a hardware architecture that enables the execution in an optimal fashion.

In this work, we propose to explore the novel AI-hardware designs and technologies for optimal co-design with the data and computation characteristics of real-time detector data. Some of the industry partners we will be considering include, but are not limit to, GraphCore, Untethered and LightMatter, with whom we have on-going collaborations using our open-sourced BCAE model [44].

**GraphCore™** (founded in 2016) is a relatively mature company and pioneered a non-GPU AI-oriented design named Intelligence Processing Unit (IPU). Its large in-processor static random-access memory (SRAM) provides fast and low-latency memory access [26]. Exploratory tests with LHCb simulation and data reconstruction also demonstrated significant improvements compared with GPUs [27]. We have observed similar improvements in the encoding rate on our BCAE model under certain conditions[1]. More optimization opportunities have been identified.

**Untethered™** (founded in 2018) specializes in dataflow architecture for AI model inference using the int8 data type. The current generation comes in a PCI-e card form factor equipped with four chips totalling 2 PetaOps. Preliminary emulation results showed promising throughput results[1]. Although quantized neural network models can maintain their accuracy in classification and regression tasks [45], in this project we will further investigate how to mitigate performance degradation due to quantization on an auto-encoder task.

**Lightmatter™** (founded in 2017) designs integrated optical circuits for AI model inference. Numeric values are coded in the phase of light beams and the phase-shift is determined by Mach–Zehnder interferometer. Their 4U server Envise claims a 7-fold improvement of inference throughput per Watt [46, 47] on BERT-base (a large natural language processing model) comparing to the NVIDIA DGX-A100.

GraphCore products are available out-of-shelf; Untethered and Lightmatter will make their products available during this proposal's performance period. Before then, they all have agreed to test our models on their prototypes and provide support to our proposed efforts (see attached support letters)[1].

## 4 Milestones, Timeline, and Resource Allocation

The project timeline and budget overview is summarized below. It is worth highlighting that our industry connections allow us to test a variety of AI chips and computing infrastructure at vendor facilities (support letters appended), which significantly improves cost-benefit of this program.

**Fiscal Year 2023** Feasibility demonstration. **Budget**: $300k, consist of ramping up FTE for two postdocs at PO and CSI (0.5 and 0.8 FTEs, respectively, which is lower than 1.0 each as recruiting is usually only completed mid-year), and PI's and other scientists' FTE.

- **Q1-Q2** Develop the initial set of AI algorithms
- **Q1-Q4** Perform feasibility tests by accessing the AI-computing accelerator chip testbeds at our industry partner. Sample support letter attached.

**Fiscal Year 2024** AI-experiment integration. **Budget**: $495k, dominated by manpower, plus one AI server for testing with sPHENIX $p$+$p$ collision Run-2024

- **Q1-Q2** Establish the first AI-computing testbed with relatively mature technology (such as GraphCore IPU)
- **Q3-Q4** Demonstration with offline (and optimistically online) TPC data stream in sPHENIX

**Fiscal Year 2025** Advanced development. **Budget**: $485k, dominated by manpower, plus second AI-computing testbed

- **Q1-Q2** Refine algorithm to optimize performance in fidelity and throughput
- **Q3-Q4** Establish second hardware test stand with mid-risk technology, and demonstrate performance with sPHENIX streaming data and the simulated EIC data

### 4.1 Timeliness of the Research

It is timely to fund this research now: (1) to keep and ramp up the momentum of research [12, 48]; (2) to continue and strengthen the connection with novel AI-hardware vendors (support letters appended); (3) to allow for a demonstrative test with the first and only streaming DAQ run at RHIC in 2024 when proton beams will be collided. For such collisions, streaming DAQ will significantly expand the physics reach of the sPHENIX experiment [5]. As a result, its data stream will be the one most closely resembling that of the EIC for the coming decade; (4) to influence the technical design of EIC project detector [2, 4] and the

[1] Further information can be made available within BNL under Commercial Nondisclosure Agreements (CNDAs).

conceptual design of the EIC Detector-2. Last, in response to this LDRD call, this proposal addresses two topical areas of *EIC Detector-2* and *Human-AI-Facility Integration*.

### 4.2 Project Team and Qualifications

This project is driven by a productive team. Since the start of exploration in 2021, five invited talks were presented and first paper on the exploratory work published in IEEE ICMLA [12, 44]. The next accepted talk is at IEEE RealTime 2022 [48]. The team member, expertise and responsibility are listed below:

**Jin Huang (PO/sPHENIX) + Postdoc** PI + domain integration and evaluation. Co-convener for EIC Detector1 global integration; manager for sPHENIX TPC readout; expertise in simulation and performance evaluation and heavy flavor physics

**Yi Huang (Postdoc), Yihui Ren, Shinjae Yoo (CSI/ML)** Leading AI/ML algorithm development on noise reduction, data compression, and network optimization.

**Byung-Jun Yoon (CSI/Math)** AI Uncertainty Quantification; PI for ASCR award for objective-driven data reduction

**Adolfy Hoisie (CSI/ACL)** Advisor, leading advanced computing architectures and technologies R&D

**Chris Pinkenburg, Martin Purschke (PO/sPHENIX)** Advisor; coordinators for sPHENIX computing and DAQ, respectively

**Torre Wenaus (PO/NPPS)** Advisor; co-convener EIC user group software working group, leadership roles in NP/HEP computing

### 4.3 Challenges and Risk mitigation

We recognize deploying the novel AI hardware is high-risk high-reward research: many such AI chips are in their pre-commercial phase and produced by start-up companies [17–22]. And their futures are variable, influenced by the external market forces. This risk is mitigated by (1) investigating a spectrum of AI hardware of various maturity as discussed in Section 3, notably the current generation of GraphCore IPU that is commercially available [17]; (2) strong connections with the hardware vendor's research teams (as evident by attached support letters) and gaining access to pre-commercial information via NDAs; (3) solid fall-back solution of high-throughput GPU systems available at BNL through CSI.

### 4.4 Data Management Plan

The team will publish findings in reviewed journals and code via GitHub with the aim of transparency and reproducibility. Examples include this team's first publication (paper [12], code [44]). The simulated and real data used in the research will be archived at tape storage facility at BNL SDCC, taking advantage of the data management plan of the hosting facilities.

## 5 Why and Benefit to BNL

This work will take full advantage of the unique facilities and capabilities at BNL: (1) BNL is the host lab for EIC. (2) Through CSI, BNL has on-going connection and non-disclosure agreement with several companies developing the AI hardware, which will give the team rare access to variety of pre-commercial and pre-release AI hardware (e.g. support letter attached). (3) Benefit from the expertise developed under ASCR award on data reduction (co-PI ByungJun Yoon ) to apply to the domain of streaming DAQ. (4) Benefit from the expertise of streaming DAQ operation at BNL, in particular the first and only streaming DAQ run at RHIC in 2024 at the sPHENIX experiment.

In return, this work will benefit BNL in multiple aspects: (1) Establish BNL as the leader in the application of novel AI-computing accelerator hardware in real-time application (2) Address the challenges of reliable real-time data reduction in the high priority scientific facilities of the EIC, which directly support and enhance high priority BNL research in nuclear physics (3) BNL is well positioned to capitalize on this proposed development with production hardware project via DOE construction funding. (4) We will establish testbeds with two distinct novel AI hardware technologies, which are available for general use of accelerating of AI adoption at BNL. (5) This research is aligned with BNL AI/ML strategy on real-time experiment application and Human-AI-Facility Integration.

This work directly advances the high-priority areas in the 2015 NSAC Long-Range Plan. And the topic of AI in experimental nuclear science has been one of the new funding areas at DOE, to which we will target future external funding.

## References

[1] sPHENIX, Technical design report: sphenix experiment at rhic (2019).
URL https://indico.bnl.gov/event/5905/

[2] Conceptual design report: Electron ion collider (2021).
URL http://www.eicug.org/web/sites/default/files/EIC_CDR_Final.pdf

[3] ECCE Consortium, Design of the ECCE Detector for the Electron Ion Collider, a proposal to EIC Detector Proposal Advisory Panel, and to be published in Nucl. Instrum. Methods A (2021).

[4] ECCE Consortium, ECCE Electronics and Readout/DAQ, ecce-note-det-2021-05 (2021).
URL https://www.ecce-eic.org/ecce-internal-notes

[5] sPHENIX, sPH-TRG-2022-001: sPHENIX Beam Use Proposal for RHIC PAC 2022 (2022).
URL https://indico.bnl.gov/event/15845/

[6] ECCE Consortium, Selected topics in ECCE software and simulation, ecce-note-comp-2021-02 (2021).
URL https://www.ecce-eic.org/ecce-internal-notes

[7] ECCE Consortium, ECCE Particle Identification, ecce-note-det-2021-04 (2021).
URL https://www.ecce-eic.org/ecce-internal-notes

[8] K. Chen, H. Chen, J. Huang, F. Lanni, S. Tang, W. Wu, A Generic High Bandwidth Data Acquisition Card for Physics Experiments, IEEE Trans. Instrum. Measur. 69 (7) (2019) 4569–4577. doi:10.1109/TIM.2019.2947972.

[9] S. Di, F. Cappello, Fast error-bounded lossy hpc data compression with sz, in: 2016 ieee international parallel and distributed processing symposium (ipdps), IEEE, 2016, pp. 730–739.

[10] J. Chen, L. Wan, X. Liang, B. Whitney, Q. Liu, D. Pugmire, N. Thompson, M. Wolf, T. Munson, I. Foster, et al., Accelerating multigrid-based hierarchical scientific data refactoring on gpus, arXiv preprint arXiv:2007.04457 (2020).

[11] P. Lindstrom, Fixed-rate compressed floating-point arrays, IEEE transactions on visualization and computer graphics 20 (12) (2014) 2674–2683.

[12] Y. Huang, Y. Ren, S. Yoo, J. Huang, Efficient Data Compression for 3D Sparse TPC via Bicephalous Convolutional Autoencoder, in: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2021, pp. 1094–1099. arXiv:2111.05423, doi:10.48550/arXiv.2111.05423.

[13] How cerebras works — software documentation (version 1.3.0).
URL https://docs.cerebras.net/en/latest/cerebras-basics/how-cerebras-works.html

[14] W. A. Wulf, S. A. McKee, Hitting the memory wall: Implications of the obvious, ACM SIGARCH computer architecture news 23 (1) (1995) 20–24.

[15] W. J. Dally, Y. Turakhia, S. Han, Domain-specific hardware accelerators, Communications of the ACM 63 (7) (2020) 48–57.

[16] G. Kestor, R. Gioiosa, D. J. Kerbyson, A. Hoisie, Quantifying the energy cost of data movement in scientific applications, in: 2013 IEEE international symposium on workload characterization (IISWC), IEEE, 2013, pp. 56–65.

[17] G. Ltd, Dell DSS 8440 Graphcore IPU White Paper.
URL https://www.graphcore.ai/dell-dss-8440-graphcore-ipu-white-paper

[18] Cerebras.
URL https://www.cerebras.net/

[19] SambaNova Systems.
URL https://sambanova.ai/

[20] WaveComp.
URL https://wavecomp.ai/

[21] Untether AI.
URL https://www.untether.ai

[22] F. ID, MLU270-S Intelligent Processing Card User Manual MLU270-S Series Intelligent Processing Card Cambricon Technologies.
URL https://fccid.io/2ARVF-MLU270-S/User-Manual/MLU270-S-Series-Intelligent-Processing-Ca

[23] D. C. Nicol, A Dataflow Processing Chip for Training Deep Neural Networks 25.
URL https://old.hotchips.org/wp-content/uploads/hc_archives/hc29/HC29.22-Tuesday-Pub/HC29.22.60-NeuralNet1-Pub/HC29.22.610-Dataflow-Deep-Nicol-Wave-07012017.pdf

[24] A. H. Veen, Dataflow machine architecture 18 (4) 365–396. doi:10.1145/27633.28055.
URL https://doi.org/10.1145/27633.28055

[25] Z. Jia, B. Tillman, M. Maggioni, D. P. Scarpazza, Dissecting the Graphcore IPU Architecture via Microbenchmarking. arXiv:1912.03413.

[26] Z. Jia, B. Tillman, M. Maggioni, D. P. Scarpazza, Dissecting the graphcore IPU architecture via microbenchmarking, CoRR abs/1912.03413 (2019). arXiv:1912.03413.
URL http://arxiv.org/abs/1912.03413

[27] L. R. M. Mohan, A. Marshall, S. Maddrell-Mander, D. O'Hanlon, K. Petridis, J. Rademacker, V. Rege, A. Titterton, Studying the potential of graphcore ipus for applications in particle physics (2020). doi:10.48550/ARXIV.2008.09210.
URL https://arxiv.org/abs/2008.09210

[28] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Transactions on Neural Networks 20 (1) (2009) 61–80. doi:10.1109/TNN.2008.2005605.

[29] J. Pata, J. Duarte, J.-R. Vlimant, M. Pierini, M. Spiropulu, Mlpf: efficient machine-learned particle-flow reconstruction using graph neural networks, The European Physical Journal C 81 (5) (2021) 1–14.

[30] S. R. Qasim, J. Kieseler, Y. Iiyama, M. Pierini, Learning representations of irregular particle-detector geometry with distance-weighted graph networks, The European Physical Journal C 79 (7) (2019) 1–11.

[31] J. Shlomi, P. Battaglia, J.-R. Vlimant, Graph neural networks in particle physics, Machine Learning: Science and Technology 2 (2) (2020) 021001.

[32] X. Ju, S. Farrell, P. Calafiura, D. Murnane, L. Gray, T. Klijnsma, K. Pedro, G. Cerati, J. Kowalkowski, G. Perdue, et al., Graph neural networks for particle reconstruction in high energy physics detectors, arXiv preprint arXiv:2003.11603 (2020).

[33] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion 76 (2021) 243–297.

[34] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, Neurocomputing 338 (2019) 34–45.

[35] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Advances in neural information processing systems 30 (2017).

[36] N. Weber, J. Starc, A. Mittal, R. Blanco, L. Màrquez, Optimizing over a bayesian last layer, in: NeurIPS workshop on Bayesian Deep Learning, 2018.

[37] J. Watson, J. A. Lin, P. Klink, J. Pajarinen, J. Peters, Latent derivative bayesian last layer networks, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1198–1206.

[38] C. Coleman, C. Yeh, S. Mussmann, B. Mirzasoleiman, P. Bailis, P. Liang, J. Leskovec, M. Zaharia, Selection via proxy: Efficient data selection for deep learning, arXiv preprint arXiv:1906.11829 (2019).

[39] K. Killamsetty, X. Zhao, F. Chen, R. Iyer, Retrieve: Coreset selection for efficient and robust semi-supervised learning, Advances in Neural Information Processing Systems 34 (2021) 14488–14501.

[40] G. Zhao, E. Dougherty, B.-J. Yoon, F. J. Alexander, X. Qian, Bayesian active learning by soft mean objective cost of uncertainty, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 3970–3978.

[41] G. Zhao, E. Dougherty, B.-J. Yoon, F. Alexander, X. Qian, Uncertainty-aware active learning for optimal bayesian classifier, in: International Conference on Learning Representations (ICLR 2021), 2021.

[42] G. Zhao, E. Dougherty, B.-J. Yoon, F. Alexander, X. Qian, Efficient active learning for gaussian process classification by error reduction, Advances in Neural Information Processing Systems 34 (2021) 9734–9746.

[43] A. H. Veen, Dataflow machine architecture 18 (4) 365–396. doi:10.1145/27633.28055.
URL https://doi.org/10.1145/27633.28055

[44] Y. Huang, Y. Ren, S. Yoo, J. Huang, Github repository and data release for Efficient Data Compression for 3D Sparse TPC via Bicephalous Convolutional Autoencoder, https://github.com/BNL-DAQ-LDRD/NeuralCompression (2021).

[45] S. Miryala, S. Mittal, Y. Ren, G. Carini, G. Deptuch, J. Fried, S. Yoo, S. Zohar, Waveform processing using neural network algorithms on the front-end electronics, Journal of Instrumentation 17 (01) (2022) C01039. doi:10.1088/1748-0221/17/01/c01039.
URL https://doi.org/10.1088/1748-0221/17/01/c01039

[46] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, M. Soljačić, Deep learning with coherent nanophotonic circuits 11 (7) 441–446. doi:10.1038/nphoton.2017.93.
URL https://doi.org/10.1038/nphoton.2017.93

[47] C. Demirkiran, F. Eris, G. Wang, J. Elmhurst, N. Moore, N. C. Harris, A. Basumallik, V. J. Reddi, A. Joshi, D. Bunandar, An electro-photonic system for accelerating deep neural networks, number: arXiv:2109.01126. arXiv:2109.01126[cs], doi:10.48550/arXiv.2109.01126.
URL http://arxiv.org/abs/2109.01126

[48] Y. Huang, T. Marshall, Y. Ren, S. Yoo, J. Huang, Real-time Compression and Noise Filtering for Time Projection Chamber Data via Bicephalous Convolutional Autoencoder, in: 2022 23rd IEEE Real Time Conference (RT22), Accepted, Not yet presented, IEEE, 2022.

# **VITA** (Jin Huang)

## Education

| | |
|---|---|
| Massachusetts Institute of Technology, Department of Physics | Cambridge, MA |
| *Ph.D. in Physics* | Feb. 2012 |
| University of Science and Technology of China, Special Class for the Gifted Young | Hefei, China |
| *Bachelor of Science in Physics* | July 2006 |

## Professional Experience

| | |
|---|---|
| Brookhaven National Laboratory (BNL), (s)PHENIX Group | Upton, NY |
| *Physicist with the Distinction of Tenure* | Jan 2021 – Present |
| *Physicist* | Aug 2019 – Jan 2021 |
| *Associate Physicist* | Oct 2016 – Aug 2019 |
| *Assistant Physicist* | Jan 2014 – Oct 2016 |
| Los Alamos National Laboratory, Subatomic Physics Group | Los Alamos, NM |
| *Postdoctoral Research Associate* | Oct 2011 – Dec 2013 |

## Selected Service Experience

| | |
|---|---|
| 2022 | Co-convener for Global Detector Integration WG, EIC Project Detector |
| 2021 | Co-convener for Simulation WG, EIC Comprehensive Chromodynamics Experiment |
| 2018 – 2022 | Co-Principal Investigator, LDRD 19-028 High-throughput FELIX DAQ |
| 2017 – Present | Nucl. Phys. Coordinator – International Organization of Chinese Physicists and Astro. |
| 2017 – Present | L3-Project Manager: Data Aggregation Module for sPHENIX Time Projection Chamber |
| 2017 – Present | sPHENIX Executive Committee |
| 2016 – Present | Co-convener for sPHENIX Heavy-flavor Topical Group and sPHENIX Simulation |
| 2011 – 2013 | Co-spokesperson for Jefferson Lab Experiment E12-11-007 |

## Selected Awards

| | |
|---|---|
| 2021 | RHIC & AGS Merit Award |
| 2018 | Sambamurti Award Lecture |
| 2013 | First Dissertation Award by the Hadronic Physics Group of the American Physical Society |
| 2012 | Los Alamos Award Program on post-doctoral work |

## Selected Publications

[1] "Efficient Data Compression for 3D Sparse TPC via Bicephalous Convolutional Autoencoder", Y. Huang, Y. Ren, S. Yoo, **J. Huang**, IEEE ICMLA 2021, pp. 1094–1099 arXiv: 2111.05423 [cs.LG]

[2] "A Generic high bandwidth data acquisition card for physics experiments", K. Chen, H. Chen, **J. Huang**, F. Lanni, S. Tang, W. Wu, IEEE Transactions on Instrumentation and Measurement, DOI: 10.1109/TIM.2019.2947972

[3] "Nuclear dependence of the transverse single-spin asymmetry in the production of charged hadrons at forward rapidity in polarized p+p, p+Al, and p+Au collisions at √(s_NN )=200 GeV", PHENIX collaboration, Phys. Rev. Lett. 123 (2019) no.12, 122001, arXiv:1903.

[4] "Design and Beam Test Results for the sPHENIX Electromagnetic and Hadronic Calorimeter Prototypes", (corresponding author), sPHENIX Collaboration, arXiv:1704.01461

[5] "B-meson production at forward and backward rapidity in p+p and Cu+Au collisions at √($s_{NN}$)=200 GeV", PHENIX collaboration, Phys. Rev. C96 (2017) no.6, 064901, arXiv:1702.01085

[6] "Spin asymmetries for vector boson production in polarized p+p collisions", **J. Huang** *et al.*, Phys.Rev. D93 (2016) no.1, 014036, arXiv:1511.06764

[7] "The PHENIX Forward Silicon Vertex Detector", C. Aidala, *et al.*, Nucl. Instrum. Meth. A755 (2014)

[8] "Beam-Target Double Spin Asymmetry $A_{LT}$ in Charged Pion Production from Deep Inelastic Scattering on a Transversely Polarized $^3$He Target at $1.4 < Q^2 < 2.7\ \mathrm{GeV}^2 1.4 < Q^2 < 2.7\ \mathrm{GeV}^2$", **J. Huang** *et al*., Phys. Rev. Lett. 108, 052001 (2012), arXiv:1108.0489

## 1. ALIGNMENT WITH THE LABORATORY MISSION AND VISION

This proposal direct address the FY23 LDRD-A topical areas of

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

And this proposal aligns with lab initiatives of

- Nuclear Physics
- Artificial Intelligence and Data Science, as well as the BNL AI strategy

## 2. POTENTIAL FUTURE FUNDING

The topic of AI in experimental nuclear science has been one of the new areas at DOE, to which we will target future external and construction funding.

Specifically, we expect the following **Return on Investment**

- Demonstrate a working prototype, as a ladder to external FOA for construction of production system from Department of Energy
    - **$10-20M**: EIC Detector 2 construction project for online computing which is at time scale of 2030
    - **Approximately $5M ×2**: Possible upgrade of EIC Detector 1 and sPHENIX, which is at time scale of coming years
- **Approximately $1M**: Saving in tape storage and offline computing need for reaching same physics goals

## 3. BUDGET JUSTIFICATION

Majority of the budget is allocated to support two postdocs, who carry out the bulk of work. We intend to establish two set of testbeds using novel AI-accelerators at BNL of increasing risks, which allow a demonstration with sPHENIX Run-2024 streaming-DAQ data. It is worth highlight that our industry connections allow us to test a variety of AI accelerators at the vendor facility (support letters appended), which significantly improves cost-benefit of this program.

More specifically:

- Fiscal Year 2023: $300k, dominated by manpower of scientists, and recruiting and ramping FTE for two postdocs at PO and CSI (0.5 and 0.8 FTE each)
- Fiscal Year 2024: $495k, dominated by manpower, and AI-computing accelerator server quoted at $40k (budgetary quote attached)
- Fiscal Year 2025: $485k, dominated by manpower, and second AI-computing accelerator-chip testbed

**4. NAME OF SUGGESTED BNL REVIEWERS**

- **Brett Viren**, PO - DUNE computing/AI expert
- **Hucheng Chen**, PO - ATLAS DAQ Lead
- **Jeffery Landgraf**, M, PO – EIC Detector-1 collaboration DAQ/electronics WG coordinator
- **Lingda Li**, CSI
  **Thomas Flynn**, CSI – Expertise in hardware and machine learning

**5. EQUIPMENT** (Reference: DOE Order 413.2C Chg. 1 (Min Chg) for guidance on equipment restrictions)

Will LDRD funding be used to purchase equipment? Y/N Y

If "Yes," provide cost and description of equipment

Year 1 - $ 0k

Year 2 - $ **40k** (Description: Testbed1 with relative mature technology, based on attached budgetary quote.)

Year 3 - $ **40k** (Description: Testbed2 with riskier technology, to be down selected)

**6. HUMAN SUBJECTS** (Reference: DOE Order 443.1C)

Are human subjects involved from BNL or a collaborating institution? Human Subjects is defined as "A living individual from whom an investigator obtains either (1) data about that individual through intervention or interaction with the individual, or (2) identifiable, private information about that individual".

If **yes**, attach copy of the current Institutional Review Board Approval and Informed Consent Form from BNL and/or collaborating institution. Y/N N

**7. VERTEBRATE ANIMALS**

Are live, vertebrate animals involved? Y/N N

If **yes**, attach copy of approval from BNL's Institutional Animal Care and Use Committee.

## 8. NEPA REVIEW

Are the activities proposed similar to those now carried out in the Department/Division which have been previously reviewed for potential environmental impacts and compliance with federal, state, local rules and regulations, and BNL's Environment, Safety, and Health Standards? (Therefore, if funded, proposed activities would require no additional environmental evaluation.) Y/N Y

If **no**, has a NEPA review been completed in accordance with the National Environmental Policy Act (NEPA) and Cultural Resources Evaluations Subject Area and the results documented? Y/N ________

(**Note:** If a NEPA review has not been completed, submit a copy of the work proposal to the BNL NEPA Coordinator for review. No work may commence until the review is completed and documented.)

## 9. ES&H CONSIDERATIONS

Does the proposal provide sufficient funding for appropriate decommissioning of the research space when the experiment is complete? Y/N Y

Is there an available waste disposal path for project wastes throughout the course of the experiment? Y/N Y

Is funding available to properly dispose of project wastes throughout the course of the experiment? Y/N Y

Are biohazards involved in the proposed work? If yes, attach a current copy of approval from the Institutional Biosafety Committee. Y/N N

Can the proposed work be carried out within the existing safety envelope of the facility (Facility Use Agreement, Nuclear Facility Authorization Agreement, Accelerator Safety Envelope, etc.) in which it will be performed? Y/N Y

If **no**, attach a statement indicating what has to be done and how modifications will be funded to prepare the facility to accept the work.

## 10. TYPE OF WORK

Select Basic, Applied or Development Basic, Development

**APPROVALS - NPP**

Business Operations Manager

James L. Desmond III
Jim Desmond

Department Chair/Division Manager
To the Department Chair/Division Manager: Please indicate if this project is a sensitive technology under the S&T Risk Matrix. (Note: Red projects require an Access Management Plan.)

☒ Green
☐ Yellow
☐ Red
☐ Not Applicable

Hong Ma
Hong Ma

Associate Laboratory Director
for Nuclear and Particle Physics

Haiyan Gao

**APPROVALS - CSI**

Business Operations Manager

Print Name

Department Chair/Division Manager
To the Department Chair/Division Manager: Please indicate if this project is a sensitive technology under the S&T Risk Matrix. (Note: Red projects require an Access Management Plan.)

☐ Green
☐ Yellow
☐ Red
☐ Not Applicable

Print Name

Associate Laboratory Director
for Nuclear and Particle Physics

Kerstin Kleese van Dam

# Real-time information Distillation on Novel AI-Hardware
# LDRD Type A
# PI:  Jin Huang

| Resource Category | DESCRIPTION | | FY23 | | FY24 | | FY25 |
|---|---|---|---|---|---|---|---|
| 050 | Salary - Scientific | | 22,730 | | 23,679 | | 62,039 |
| 051 | Salary - Research Assoc | | 127,323 | | 204,849 | | 158,361 |
| 050 | Salary - Professional | | 0 | | 0 | | 0 |
| 050 | Salary -Technical | | 0 | | 0 | | 0 |
| 050 | Salary - Management & Admin. | | 0 | | 0 | | 0 |
| | **Total FTEs** | | **1.32** | | **1.97** | | **1.66** |
| **TOTAL SALARY/WAGE & FRINGE** | | | **150,052** | | **228,528** | | **220,400** |
| various | Contracts - Low Value | | 0 | | 0 | | 0 |
| 280 | Foreign Travel | | 7,000 | | 7,000 | | 10,000 |
| 290 | Domestic Travel | | 0 | | 0 | | 0 |
| various | Purchase - computers & misc | | 10,590 | | 2,512 | | 1,245 |
| **TOTAL MSTC** | | | **17,590** | | **9,512** | | **11,245** |
| 170 | Relocation Expense | | 10,000 | | 0 | | 0 |
| 240 | Registration Fees | | 0 | | 0 | | 0 |
| 271 | Communications | | 1,200 | | 1,230 | | 1,261 |
| **TOTAL COM/MISC** | | | **11,200** | | **1,230** | | **1,261** |
| various | **Contracts** | | 0 | | 0 | | 0 |
| 340 | Special Procurements | | 0 | | 0 | | 0 |
| **TOTAL SPECIAL PURCHASES** | | | **0** | | **0** | | **0** |
| 312/314 | Equipment Low - value | | 0 | | 40,000 | | 40,000 |
| **TOTAL EQUIPMENT** | | | **0** | | **40,000** | | **40,000** |
| **TOTAL OTH-ALLOCS** | | | **0** | | **0** | | **0** |
| 480 | Space | | 1,500 | | 1,535 | | 1,570 |
| **TOTAL SPACE** | | | **1,500** | | **1,535** | | **1,570** |
| **TOTAL DIRECT COSTS** | | | **180,342** | | **280,804** | | **274,476** |
| 251 | Electric Distributed (Electric Power Burden) | | 1,501 | | 2,285 | | 2,204 |
| 700/701/481 | Organizational Burden | | 30,293 | | 43,147 | | 42,640 |
| **TOTAL ORGANIZATIONAL BURDEN** | | | **31,793** | | **45,432** | | **44,844** |
| 745 | Procurement (Material Handling) | | 1,231 | | 3,466 | | 3,587 |
| 710 | G&A Burden | | 0 | | 28,049 | | 27,507 |
| 711 | Adjs to G&A Burden | | | | | | |
| 720 | Common Support | | 86,633 | | 122,831 | | 120,459 |
| 722 | Safeguards & Security Assess | | 0 | | 0 | | 0 |
| 746 | Adjs to Procurement Burden | | | | | | |
| **TOTAL LABORATORY BURDEN** | | | **87,864** | | **154,346** | | **151,554** |
| 705 | LDRD Burden | | 0 | | 14,417 | | 14,126 |
| **TOTAL PROGRAM COSTS** | | | **300,000** | | **495,000** | | **485,000** |

*** Note:  there is a 7% lab support for post docs**

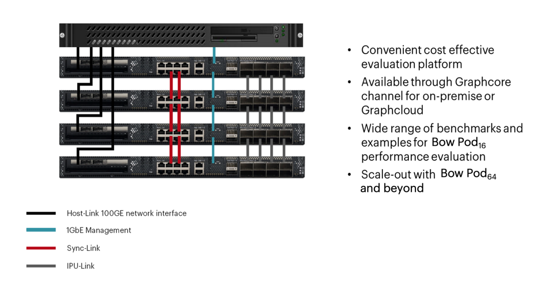
| Labor | | FY23 | | FY24 | | FY25 | |
|---|---|---|---|---|---|---|---|
| **Band** | **Name** | **FTE** | **Amount** | **FTE** | **Amount** | **FTE** | **Amount** |
| RA2 | Post Doc (CSI) * | 0.74 | 78,190 | 0.93 | 102,424 | 0.93 | 105,574 |
| RA2 | Post Doc (PO) * | 0.47 | 49,133 | 0.93 | 102,424 | 0.47 | 52,787 |
| SCI1 | R. Yihui Ren (CSI) | 0.03 | 5,524 | 0.03 | 5,755 | 0.03 | 5,932 |
| SCI1 | B. Yoon (CSI) | 0.03 | 5,524 | 0.03 | 5,755 | 0.03 | 5,932 |
| SCI2 | J. Huang | 0.05 | 11,681 | 0.05 | 12,170 | 0.20 | 50,176 |
| | Total | 1.32 | 150,052 | 1.97 | 228,528 | 1.66 | 220,400 |

EXXACT
46221 Landing Parkway, Fremont, CA, 94538
Fremont, CA, 94539
Tel: (510) 226-7366 Fax: (510) 226-7367 www.exxactcorp.com

| | |
|---|---|
| Quote # : | 119317-1 |
| Issue / Expiration Dates : | 06/06/2022 / 07/06/2022 |
| ETA : | |
| Terms : | Net 30 |
| Ship Via : | Freight |
| Tax Rate : | None |
| Salesperson: | Mike Chen |
| Sales Email: | mchen@exxactcorp.com |
| Sales Phone: | 510-226-7366 x204 |

## QUOTE

| Billing Information | | Shipping Information |
|---|---|---|
| Company: | Brookhaven National Laboratory | Brookhaven National Laboratory |
| Attention: | | |
| Street Address: | | |
| | | |
| City, State, Zip: | | |
| Contact: | | |
| Phone Number: | | |
| Email: | | |

| Description | Quantity | Unit Price | Extended Amount |
|---|---|---|---|
| **1x BOW-2000 1U**<br>**1x Graphcore BOW-2000 w/ 3 Year Support Package** | 1 | $40,000.16 | $40,000.16 |
| ---------------------------- Graphcore BOW-2000 ---------------------------- | 0 | | |
| 1U BOW-2000 IPU System<br>- 4x BOW IPUs - 1.4 PetaFLOPS FP16.16 AI Compute<br>- 5,888 processor cores<br>- 35,000 independent parallel threads<br>- Up to ~450GB Exchange Memory - Up to 448GB Streaming Memory, 3.6GB In-Processor-Memory<br>- IPU-Fabric for compiled in network comprised of<br>- IPU-LinkTM - 512Gbps for intra IPU-POD64 communication<br>- GW-Link - 2x 100Gbps Gateway-Links for inter IPU-POD64 communication<br>- Sync-Link - dedicated hardware signalling for BSP, low jitter on IPU to IPU synchronisation<br>-Host-Link - PCIe Gen4 RoCEv2 NIC/SmartNIC Interface for IPU-M2000 to server communication | 1 | | |
| BOW IPU 3 Year Support | 1 | | |
| <br> Host-Link 100GE network interface<br>1GbE Management<br>Sync-Link<br>IPU-Link<br>• Convenient cost effective evaluation platform<br>• Available through Graphcore channel for on-premise or Graphcloud<br>• Wide range of benchmarks and examples for Bow Pod$_{16}$ performance evaluation<br>• Scale-out with Bow Pod$_{64}$ and beyond | 0 | | |
| **GOV Discount Applied** | **1** | | |

| |
|---|
| **Unless otherwise specifically noted, assembled computer systems are covered under the 3-Year Exxact Standard System Warranty, available here** |
| Configuration Notes: |
| |

| Notes/Comments/Special Instructions: | Subtotal: $ 40,000.16<br>Tax: $ 0.00<br>Shipping : $ TBD<br>Total: $ 40,000.16 |
|---|---|
| | |

intel Technology Provider Platinum 2019 GSA Contract Holder

ISO 9001:2008 | Intel ITP Certificate

**Terms and Conditions:**
**NOTICE: SHARING THIS QUOTE WITH ANYONE OTHER THAN THE INDIVIDUAL THAT EXXACT SENDS THE QUOTE TO (INTENDED RECIPIENT) MAY INVALIDATE THIS QUOTATION.**
All purchases of Exxact Corporation products are subject to the following Terms and Conditions of Sale. (A) Please allow two weeks from receipt of purchase order before shipment. (B) Shipping may not be included on the Quote but will be included on the invoice if applicable. (C) Applicable sales tax (tax rate based on ship-to location) will be charged on shipments in California if Exxact Corporation does not receive a Tax Exemption Certificate prior to shipment of the order. (D) Warranty, Terms of Payment and quote Expiration Date that appear in the quote. (E) Unless otherwise stated, Exxact system builds are custom, built to order and are NCNR (Not Cancelable, Not Returnable). (F) Exxact Corporation is not responsible for potential unpredictable market volatility and reserves the right to update the quotation, even if the quotation is within the quotes stated Issue/Expiry dates due to this.

Tickets may be raised at Exxact Support Portal

GRAPHCORE

Dr. Jin Huang
Brookhaven National Lab
Physics Department
P.O. Box 5000
Upton, NY 11973

Re: Real-time Information Distillation on Novel AI-Accelerators

Dear Dr. Huang,

The purpose of this letter is to confirm the status of discussions between Graphcore and Brookhaven National Lab. At Graphcore, we are interested in collaborating with you to facilitate the development of low-cost computing platform utilizing probabilistic processor architectures. We will be happy to help your team optimize on our Intelligence Processing Units (IPUs) as part of your proposal.

The Parties will collaborate as follows:

- The parties will enter into a two-way Commercial Nondisclosure Agreement (CNDA) to allow both parties to disclose and discuss future development directions for pre-commercial, pre-release, and pre-publication plans and products.
- The Parties will share test and evaluation software from both parties through free license agreements for the lifetime of the project.
- Brookhaven National Lab will provide feedback to Graphcore on new and planned features and application programming interface (API).
- Graphcore will provide feedback to Brookhaven National Lab on designed algorithms and their technical feasibility and provide technical support in the implementation stage.
- Graphcore will provide one-month access to an off-site server at least for one user at Brookhaven National Lab to evaluate the IPU platform.
- The Parties will engage in common public dissemination of the research results directly related to this collaboration, where authorship and attributions are going to be determined mutually based on the contributions of participants from The Parties.

The Parties agree that the terms of this Letter of Support are not binding on either Party and do not create any legal rights or obligations for either Party, notwithstanding those that will be covered by the terms under the CNDA. The collaboration outlined above is only intended to facilitate discussions and preparation of any further documentation describing the final understanding of The Parties. The Parties, in good faith, will negotiate the terms of a Memorandum of Understanding, based on the described outline and in concordance with applicable institutional rules and regulations.

We are looking forward to a fruitful collaboration with you on this project, and we wish you good luck in your application.

Sincerely,

Jacob Moulton
Head of Federal Sales
jacobm@graphcore.ai
(860)-576-1606

Lightmatter, Inc.
100 Summer St.,
Boston, MA 02110

June 16th, 2022

Brookhaven National Laboratory
20 Brookhaven Ave
Upton, NY 11973

To Whom it may concern:

This letter is intended to demonstrate Lightmatter's support and interest in collaborating with Brookhaven National Laboratory on solving scientific computing problems using Lightmatter's photonic compute and interconnect technology. As Lightmatter's technology is designed as a general purpose accelerator, there are many areas of exploration possible. Specifically, the following areas are of particular interest:

1. Exploring real-time scientific computing problems which require high computational throughput and low round-trip latency
2. Investigating the effects of quantization and analogue noise inherent to photonic accelerators as it relates to accuracy for scientific computing problems
3. Researching the impacts of analogue noise during training and how they influence model robustness, and security against malicious attacks.

Given Brookhaven's expertise in solving challenging scientific problems and strong technical acumen, Lightmatter considers partnering with Brookhaven National Laboratory on these topics a priority.  We look forward to supporting the Brookhaven team in demonstrating next generation compute being brought to bear on the world's most challenging scientific problems.

Sincerely,

Bradford Turcott
Director Field Applications Engineering at Lightmatter, Inc.
480-620-2683
bradford@lightmatter.co

UNTETHER AI

# Support Letter

Untether AI and Brookhaven National Lab

The team at Untether AI is hopeful to work with Brookhaven National Lab (BNL) to explore artificial intelligence acceleration. The collaboration is an opportunity to solve some of BNL's most demanding compute challenges with Untether AI hardware and software, while giving Untether AI valuable experience and exposure to the cutting edge of high performance computing. BNL would have the opportunity to influence the Untether AI technology roadmap, and apply to participate in the Early Access Program for next generation silicon and software.

BNL's expertise will directly contribute to Untether AI's ability to build products for high performance computing and scientific computing, and would be a valuable partner to increase the awareness of our solutions in this space.

Sincerely,

George Totolos
Business Development Manager, Untether AI