Statement of the Problem. The Field-Programmable Gate Arrays (FPGA) gains overwhelming attention in edge computing: it serves as alternative deep learning (DL) inference engine to GPU because of its speed and energy efficiency. FPGA for DL is severely constrained because existing FPGA accelerator kernels are designed for Deep Convolutional Neural Network (CNN) of pictures and videos with n-d dense array while many science applications deal with unstructured and sparse data. There are no FPGA DL compilers supporting the sparse data networks and inferences, i.e., Graph Convolutional Neural Networks (GNN).

How the problem is Being Addressed. We will extend XiLinx FPGA to support more deep learning architectures, including GNN and Deep Reinforcement Learning (DRL). We will use the DNNDK (Deep Neural Networks Development Kit), the High-Level Synthesis for Xilinx FPGA, to ease programming complexity and perform three main functions in software and hardware integration: scheduling, allocation, resource mapping, and binding. One cannot leave all these technical details to the HLS compiler. In particular, we will manually merge neural network layers and re-arrange the execution order of tasks for better parallelism and data granularity and locality.

What was done in Phase I: We implemented software trigger system based on GNN for sPHENIX and confirmed that the trigger system detects right tracks with 97%, and find the interaction point within 200 μm from the ground truth. The smart trigger identified 70% of trigger events in simulation data. In addition, we implemented a DRL-based NSLS-II orbit controller to correct off-track electron beams from operator console.

What is planned for the Phase II project: We will implement a highly optimized hardware-level GNN layer for real-time requirements. We will tackle four tasks to enable end-to-end DL solutions from AI software to FPGA hardware/firmware for two DOE programs. We will implement a cost-effective embedded system comprised of NVIDIA Jetson and FPGA to support on-chip training and inference. Ultimately, a **Hybrid GPU+FPGA** prototype combines hardware acceleration and algorithmic optimization to reduce latency at edge computing environment to meet the real-time requirements of science applications.

Commercial Applications and Other Benefits We will improve the efficiency of physics data acquisition by at least a factor of 700, which translates to a 700-fold reduction in data volume while retaining the same amount of signals. Our research will enable co-designing marketable products that are innovative, scalable, and transformative, and extend the successes of high-performance computing in DOE to the edge environment.

Keywords: Deep Learning; Field-programmable gate array, Autonomous Control.

Summary for Members of Congress: Any hardware upgrade to the decision engine of the data acquisition and control systems because the hardware circuits are fixed and cannot be reused. Our proposed solution adopts machine intelligent and customizable hardware to allow the control module to be flexible and support upgrades as simple as button clicks.