CPAD RDC 5: Trigger and DAQ

Zeynep Demiragli (Boston University)

Jinlong Zhang (Argonne National Lab)

RDC 5 Timeline:

- First meeting (09/29)
  - To discuss key R&D areas and survey questions

- Circulate Survey (early Oct)
  - To identify and organize subgroups/subareas through survey
  - To understand relevant teams, existing activities, facilities/infrastructures

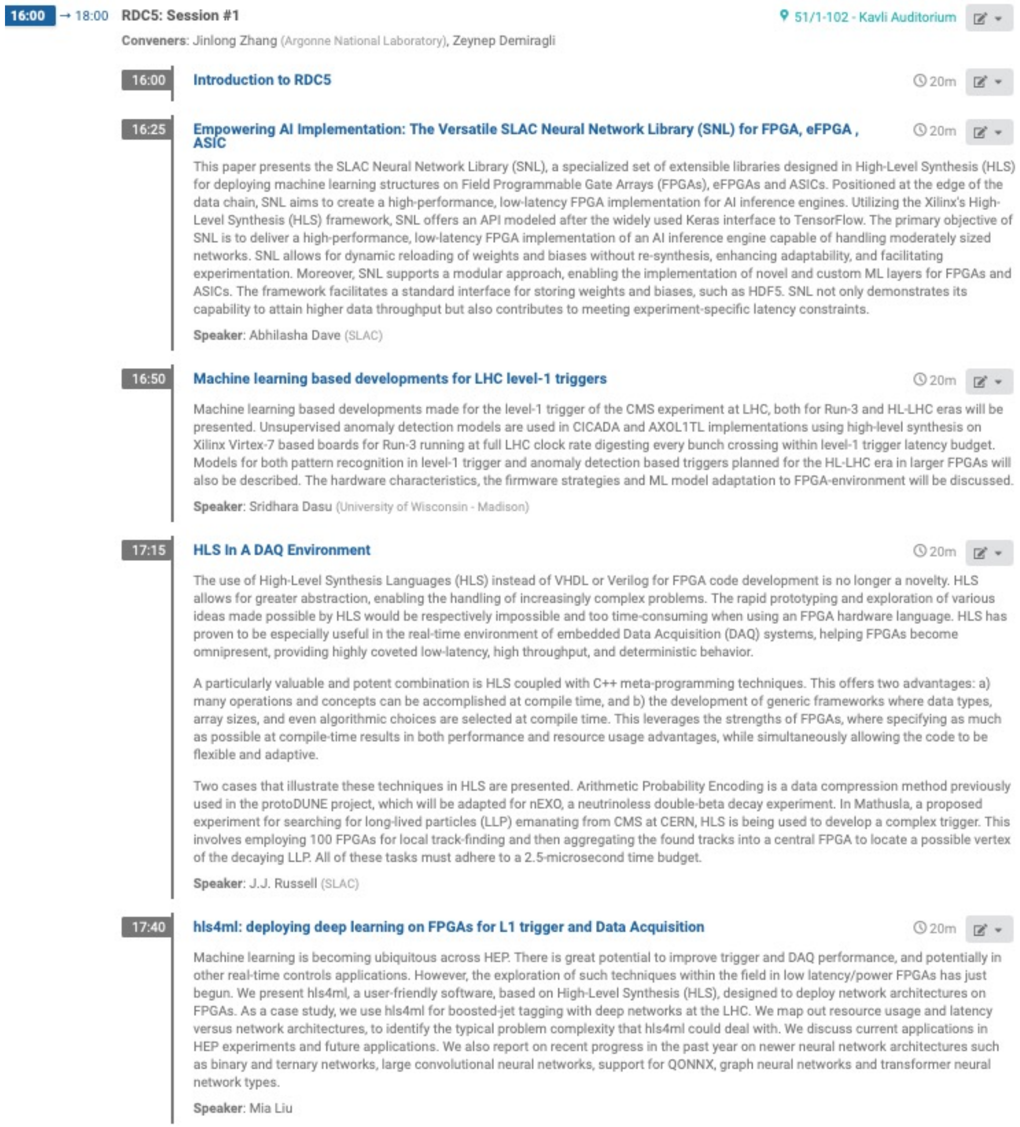
Only 1 person filled the survey…

- Second meeting (late Oct)
  - To formulate and discuss work packages

- CPAD workshop at SLAC (Nov 7-10)
  - https://indico.slac.stanford.edu/event/8288/registrations/539/
  - Present RDC 5 status at CPAD
  - In-person discussion to advance our preparation for proposal(s)

Workshop Schedule

Tuesday Nov 7



16:00 → 18:00 **RDC5: Session #1** 51/1-102 - Kavli Auditorium

Conveners: Jinlong Zhang (Argonne National Laboratory), Zeynep Demiragli

16:00 **Introduction to RDC5** 20m

16:25 **Empowering AI Implementation: The Versatile SLAC Neural Network Library (SNL) for FPGA, eFPGA , ASIC** 20m

This paper presents the SLAC Neural Network Library (SNL), a specialized set of extensible libraries designed in High-Level Synthesis (HLS) for deploying machine learning structures on Field Programmable Gate Arrays (FPGAs), eFPGAs and ASICs. Positioned at the edge of the data chain, SNL aims to create a high-performance, low-latency FPGA implementation for AI inference engines. Utilizing the Xilinx's High-Level Synthesis (HLS) framework, SNL offers an API modeled after the widely used Keras interface to TensorFlow. The primary objective of SNL is to deliver a high-performance, low-latency FPGA implementation of an AI inference engine capable of handling moderately sized networks. SNL allows for dynamic reloading of weights and biases without re-synthesis, enhancing adaptability, and facilitating experimentation. Moreover, SNL supports a modular approach, enabling the implementation of novel and custom ML layers for FPGAs and ASICs. The framework facilitates a standard interface for storing weights and biases, such as HDF5. SNL not only demonstrates its capability to attain higher data throughput but also contributes to meeting experiment-specific latency constraints.

Speaker: Abhilasha Dave (SLAC)

16:50 **Machine learning based developments for LHC level-1 triggers** 20m

Machine learning based developments made for the level-1 trigger of the CMS experiment at LHC, both for Run-3 and HL-LHC eras will be presented. Unsupervised anomaly detection models are used in CICADA and AXOL1TL implementations using high-level synthesis on Xilinx Virtex-7 based boards for Run-3 running at full LHC clock rate digesting every bunch crossing within level-1 trigger latency budget. Models for both pattern recognition in level-1 trigger and anomaly detection based triggers planned for the HL-LHC era in larger FPGAs will also be described. The hardware characteristics, the firmware strategies and ML model adaptation to FPGA-environment will be discussed.

Speaker: Sridhara Dasu (University of Wisconsin - Madison)

17:15 **HLS In A DAQ Environment** 20m

The use of High-Level Synthesis Languages (HLS) instead of VHDL or Verilog for FPGA code development is no longer a novelty. HLS allows for greater abstraction, enabling the handling of increasingly complex problems. The rapid prototyping and exploration of various ideas made possible by HLS would be respectively impossible and too time-consuming when using an FPGA hardware language. HLS has proven to be especially useful in the real-time environment of embedded Data Acquisition (DAQ) systems, helping FPGAs become omnipresent, providing highly coveted low-latency, high throughput, and deterministic behavior.

A particularly valuable and potent combination is HLS coupled with C++ meta-programming techniques. This offers two advantages: a) many operations and concepts can be accomplished at compile time, and b) the development of generic frameworks where data types, array sizes, and even algorithmic choices are selected at compile time. This leverages the strengths of FPGAs, where specifying as much as possible at compile-time results in both performance and resource usage advantages, while simultaneously allowing the code to be flexible and adaptive.

Two cases that illustrate these techniques in HLS are presented. Arithmetic Probability Encoding is a data compression method previously used in the protoDUNE project, which will be adapted for nEXO, a neutrinoless double-beta decay experiment. In Mathusla, a proposed experiment for searching for long-lived particles (LLP) emanating from CMS at CERN, HLS is being used to develop a complex trigger. This involves employing 100 FPGAs for local track-finding and then aggregating the found tracks into a central FPGA to locate a possible vertex of the decaying LLP. All of these tasks must adhere to a 2.5-microsecond time budget.

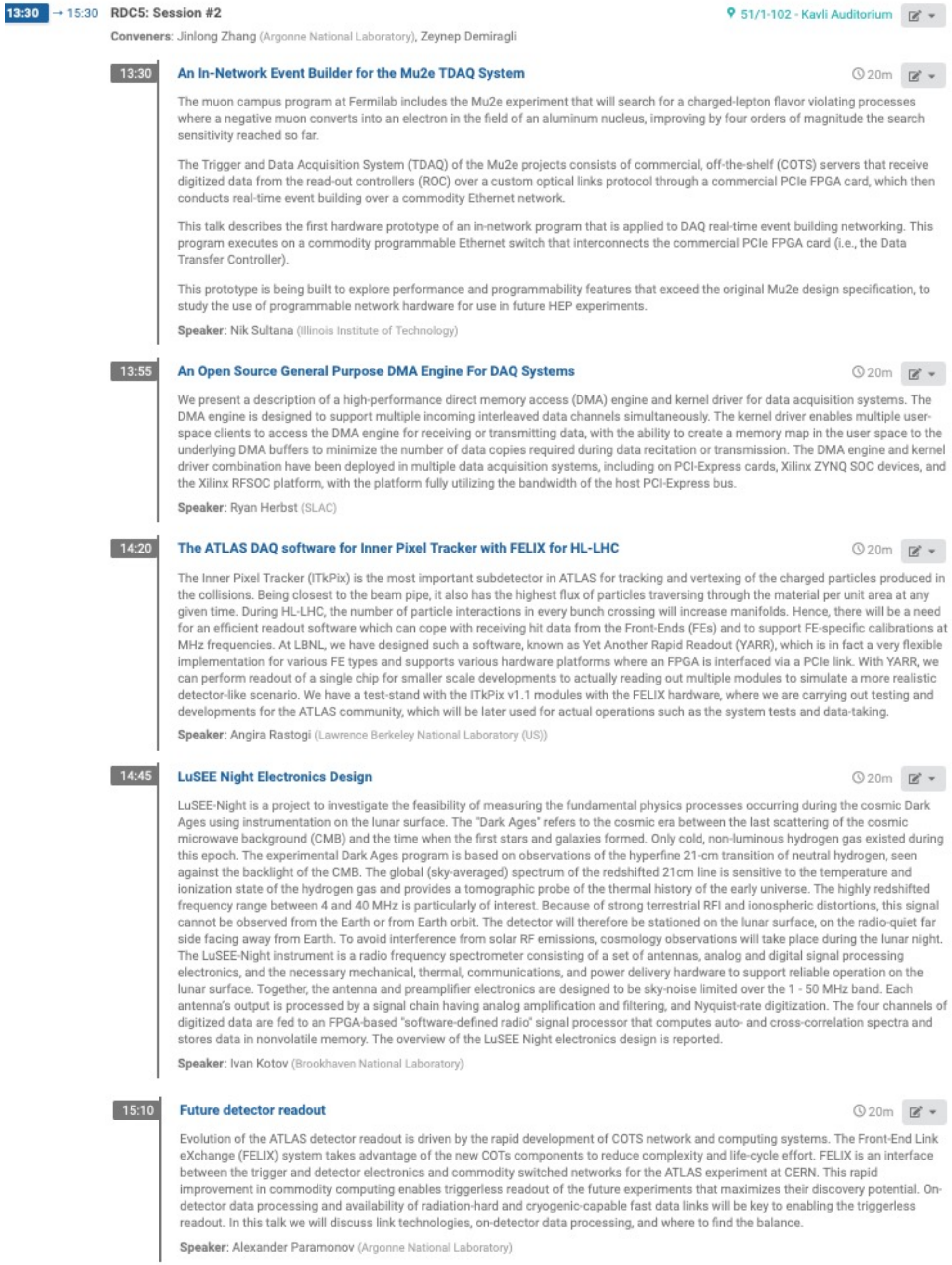
Speaker: J.J. Russell (SLAC)

17:40 **hls4ml: deploying deep learning on FPGAs for L1 trigger and Data Acquisition** 20m

Machine learning is becoming ubiquitous across HEP. There is great potential to improve trigger and DAQ performance, and potentially in other real-time controls applications. However, the exploration of such techniques within the field in low latency/power FPGAs has just begun. We present hls4ml, a user-friendly software, based on High-Level Synthesis (HLS), designed to deploy network architectures on FPGAs. As a case study, we use hls4ml for boosted-jet tagging with deep networks at the LHC. We map out resource usage and latency versus network architectures, to identify the typical problem complexity that hls4ml could deal with. We discuss current applications in HEP experiments and future applications. We also report on recent progress in the past year on newer neural network architectures such as binary and ternary networks, large convolutional neural networks, support for QONNX, graph neural networks and transformer neural network types.

Speaker: Mia Liu

Thursday Nov 9



13:30 → 15:30 **RDC5: Session #2** 51/1-102 - Kavli Auditorium

Conveners: Jinlong Zhang (Argonne National Laboratory), Zeynep Demiragli

13:30 **An In-Network Event Builder for the Mu2e TDAQ System** 20m

The muon campus program at Fermilab includes the Mu2e experiment that will search for a charged-lepton flavor violating processes where a negative muon converts into an electron in the field of an aluminum nucleus, improving by four orders of magnitude the search sensitivity reached so far.

The Trigger and Data Acquisition System (TDAQ) of the Mu2e projects consists of commercial, off-the-shelf (COTS) servers that receive digitized data from the read-out controllers (ROC) over a custom optical links protocol through a commercial PCIe FPGA card, which then conducts real-time event building over a commodity Ethernet network.

This talk describes the first hardware prototype of an in-network program that is applied to DAQ real-time event building networking. This program executes on a commodity programmable Ethernet switch that interconnects the commercial PCIe FPGA card (i.e., the Data Transfer Controller).

This prototype is being built to explore performance and programmability features that exceed the original Mu2e design specification, to study the use of programmable network hardware for use in future HEP experiments.

Speaker: Nik Sultana (Illinois Institute of Technology)

13:55 **An Open Source General Purpose DMA Engine For DAQ Systems** 20m

We present a description of a high-performance direct memory access (DMA) engine and kernel driver for data acquisition systems. The DMA engine is designed to support multiple incoming interleaved data channels simultaneously. The kernel driver enables multiple user-space clients to access the DMA engine for receiving or transmitting data, with the ability to create a memory map in the user space to the underlying DMA buffers to minimize the number of data copies required during data recitation or transmission. The DMA engine and kernel driver combination have been deployed in multiple data acquisition systems, including on PCI-Express cards, Xilinx ZYNQ SOC devices, and the Xilinx RFSOC platform, with the platform fully utilizing the bandwidth of the host PCI-Express bus.

Speaker: Ryan Herbst (SLAC)

14:20 **The ATLAS DAQ software for Inner Pixel Tracker with FELIX for HL-LHC** 20m

The Inner Pixel Tracker (ITkPix) is the most important subdetector in ATLAS for tracking and vertexing of the charged particles produced in the collisions. Being closest to the beam pipe, it also has the highest flux of particles traversing through the material per unit area at any given time. During HL-LHC, the number of particle interactions in every bunch crossing will increase manifolds. Hence, there will be a need for an efficient readout software which can cope with receiving hit data from the Front-Ends (FEs) and to support FE-specific calibrations at MHz frequencies. At LBNL, we have designed such a software, known as Yet Another Rapid Readout (YARR), which is in fact a very flexible implementation for various FE types and supports various hardware platforms where an FPGA is interfaced via a PCIe link. With YARR, we can perform readout of a single chip for smaller scale developments to actually reading out multiple modules to simulate a more realistic detector-like scenario. We have a test-stand with the ITkPix v1.1 modules with the FELIX hardware, where we are carrying out testing and developments for the ATLAS community, which will be later used for actual operations such as the system tests and data-taking.

Speaker: Angira Rastogi (Lawrence Berkeley National Laboratory (US))

14:45 **LuSEE Night Electronics Design** 20m

LuSEE-Night is a project to investigate the feasibility of measuring the fundamental physics processes occurring during the cosmic Dark Ages using instrumentation on the lunar surface. The "Dark Ages" refers to the cosmic era between the last scattering of the cosmic microwave background (CMB) and the time when the first stars and galaxies formed. Only cold, non-luminous hydrogen gas existed during this epoch. The experimental Dark Ages program is based on observations of the hyperfine 21-cm transition of neutral hydrogen, seen against the backlight of the CMB. The global (sky-averaged) spectrum of the redshifted 21cm line is sensitive to the temperature and ionization state of the hydrogen gas and provides a tomographic probe of the thermal history of the early universe. The highly redshifted frequency range between 4 and 40 MHz is particularly of interest. Because of strong terrestrial RFI and ionospheric distortions, this signal cannot be observed from the Earth or from Earth orbit. The detector will therefore be stationed on the lunar surface, on the radio-quiet far side facing away from Earth. To avoid interference from solar RF emissions, cosmology observations will take place during the lunar night. The LuSEE-Night instrument is a radio frequency spectrometer consisting of a set of antennas, analog and digital signal processing electronics, and the necessary mechanical, thermal, communications, and power delivery hardware to support reliable operation on the lunar surface. Together, the antenna and preamplifier electronics are designed to be sky-noise limited over the 1 - 50 MHz band. Each antenna's output is processed by a signal chain having analog amplification and filtering, and Nyquist-rate digitization. The four channels of digitized data are fed to an FPGA-based "software-defined radio" signal processor that computes auto- and cross-correlation spectra and stores data in nonvolatile memory. The overview of the LuSEE Night electronics design is reported.

Speaker: Ivan Kotov (Brookhaven National Laboratory)

15:10 **Future detector readout** 20m

Evolution of the ATLAS detector readout is driven by the rapid development of COTS network and computing systems. The Front-End Link eXchange (FELIX) system takes advantage of the new COTs components to reduce complexity and life-cycle effort. FELIX is an interface between the trigger and detector electronics and commodity switched networks for the ATLAS experiment at CERN. This rapid improvement in commodity computing enables triggerless readout of the future experiments that maximizes their discovery potential. On-detector data processing and availability of radiation-hard and cryogenic-capable fast data links will be key to enabling the triggerless readout. In this talk we will discuss link technologies, on-detector data processing, and where to find the balance.

Speaker: Alexander Paramonov (Argonne National Laboratory)

Ideas for Work Packages (based on BRN Report)

PRD 21

- Real-time / low-latency data reduction and feature extraction
- Fast artificial intelligence and neuromorphic computing on real-time hardware
- High-bandwidth, rad-hard, low-power optical link (>50Gbps)
- Wireless readout

PRD 22

- Intergrading modern computing architecture and emergingtechnologies
- Self-running DAQ system

PRD 23

- Timing distribution with picosecond synchronization (1ps over 1km)

Notes on the funding Opportunities

FOA

- https://science.osti.gov/grants/FOAs/-/media/grants/pdf/foas/2023/DE-FOA-0003177.pdf
- This is an open call.
  - Not enough time for FY24,
  - Is likely be better to target FY25.
- In the previous meeting, Helmut indicated there may also be a specific FOA next year.

Collaboration Proposals