Snowmass 2021 Letter Of Interest: FPGA Based Artificial Intelligence Inference In Triggered Detectors

Ryan Herbst, Angelo Dragone, Michael Kagan, L. J. Kaufman,
Brian Mong, and Rafael Teixeira De Lima
SLAC National Accelerator Laboratory

Thematic areas:

- (IF4) Trigger and DAQ
- (IF7) Electronics/ASICs

Contact Information:

Ryan Herbst rherbst@slac.stanford.edu

Modern detector electronics generate data volumes much larger than can be feasible processed and stored by back end data acquisition systems. As a result experiments are typically triggered to select windows of data, or events, from the generated data. These triggered events are readout globally across the various sources of data in a detector, coordinated by a central trigger decision node that makes a decision based on all inputs.

In a typical physics detector, each readout channel contains electronics which hosts a local ring buffer where the raw data is stored temporarily while the central trigger decision is generated. At the same time the electronics may provide “trigger primitives” which identify interesting features in the raw data possibly indicating the presence of an important physics event to the trigger node. The logic which generates these trigger primitives must be fast and the resulting stream of data must be small compared to the incoming data bandwidth. This first level data processing logic is a prime candidate for distributed edge AI engines implemented either in the front end ASIC hardware or in FPGAs closely connected to the ASIC or ADC.

The amount of time, or latency, allowed for a central trigger decision is dependent on the size of the ring buffer and therefore the time before the data is overwritten. This latency requirement can greatly limit its complexity which may result in excessive false positive triggers. The typical latency requirements for these trigger systems range from 10us - 100us depending on the type of detector, the raw data rate, and technologies used in the front end. This limit makes software and GPU based AI engines infeasible, requiring the use of ASIC and FPGA based trigger processing systems. FPGAs are the more preferred choice due to their flexibility and lower non-recurring engineering costs. The nature of the FPGA hardware makes it possible to have multiple calculations run in parallel in a pipeline. In other words, the data arrives at the input layer where each node can calculate its result independently. The results are forwarded to the next layer where more sophisticated calculations can be performed, and this ganging can be repeated as necessary to build sufficiently complex triggering decisions. Once the detector is running at full rate, all layers will be calculating simultaneously. This pipelined approach supports both high frame rates and low latency applications.

SLAC intends to continue development of a framework for deploying AI inference engines onto FPGAs, utilizing the modern tools provided by the FPGA vendors and the acceleration engines available in modern FPGAs. This framework will allow non professional FPGA developers to define the neural network structure to be deployed on the FPGA as an input to this framework. This framework is being funded with an existing LDRD at SLAC focused on the development of fast inference engines for detectors deployed at LCLS-2, with a view toward applications in other others, including trigger systems for HEP detectors such as ATLAS, DUNE & nEXO. We intend to investigate how this framework can be applied to the trigger systems for these detectors, allowing for more efficient trigger systems to be deployed utilizing artificial intelligence, while meeting the low latency requirements of these detector systems.