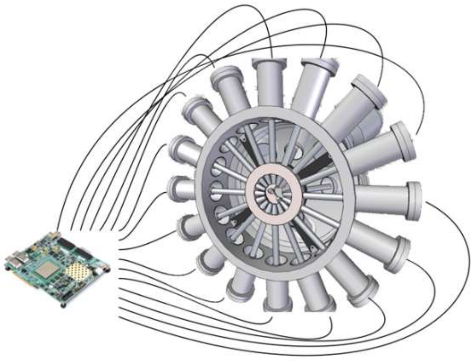
Machine Learning at the Edge: Intelligent Data Triage in Real Time

Audrey Corbeil Therrien

Several large physics experiments face an increasingly large "data firehose" problem. Raw data generation exceeds TB/s rates for several existing and planned experiments, generating untenable data sets in very little time. The raw data often contain limited information; vetoing and extracting this relevant information online would reduce the offline storage requirements by several orders of magnitude. Additionally, ultra low latency data analysis can be used to drive a fast feedback control system to adjust the experiment in real time, including decisions on data acquisition conditions, detector parameter adjustments and source operation modifications. Many of the current analysis algorithms use computationally expensive operations that require uploading the data to a CPU or GPU compute nodes, adding compute and data transfer latency to the decision loop and limiting the performance of online data analysis. However, with appropriate training, machine learning can categorize data and extract relevant information from raw data using simple – less computationally expensive – arithmetic operations. Placing these fast inference models on FPGAs near the detector – at the edge – would reduce and the data velocity at the source and provide very low latency (1-100 μs) feedback information. We are developing an implementation of Edge Machine Learning (EdgeML) for LCLS-II targeting the CookieBox, an angular streaking detector which can be used for online non-destructive beam diagnostics.



The CookieBox EdgeML extracts beam characteristics (energy and time profile) from the CookieBox raw data. This snapshot can then be used to select appropriate data compression schemes in real-time and control downstream detectors. This approach can be applied to a large variety of physics detectors generating high data velocities or requiring very low power consumption.