Snowmass 2021 Letter of Interest: Edge Computing Devices for Detectors Developed for Scientific Applications

Sandeep Miryala, Grzegorz Deptuch, Vamshi Manthena, and Gabriella Carini
Brookhaven National Laboratory
Jeremy Love
Argonne National Laboratory

Thematic Areas: IF7: Electronics/ASICs
IF4: Trigger and DAQ

New experiments are producing increasingly growing amounts of data that, in part, need to be stored, but also need to be processed on the fly, helping in critical decisions in the operation of experiments, such as for example triggering. Individual streams of data carry majorly 'zeros' and noise and the whole activity context is not known. The bandwidths of links cannot cope with streams of produced unprocessed data. Also sending of data entails significant energy expense that can be reduced if data is converted to information as early as possible. The streams start reflecting zoned activities in the experiment only when combined in data concentrators. Traditionally, data is minimally processed inside and processing needing context is outside real time or offline. In the regard of the above, it is worth considering techniques of reduction of volumes of raw data generated at detector frontends, and where streams of data get aggregated by embedding machine learning, often termed edge computing in custom designed detector readout electronics. Introduction of smartness by bringing processing inside with artificial neural networks is a future for Read Out Integrated Circuits (ROIC) ASICs. Addressing these topics with practical solutions is the subject of this Letter of Intent.

The focus of this proposal is the study and carry out development of conventional Von-Neumann and non Von-Neumann neuromorphic computing based AI ASICs for scientific data processing being targeted for the front-end electronics, respecting specific needs, such as extreme environments of operation, restriction on power dissipation or circuital resources. This approach postulates harnessing the co-designing methodology. As architectures and sizes of a neural processor fits only classes of problems, first, building and optimizing a neural network model with the tools available today: Tensor Flow, PyTorch or Caffe2 frameworks needs to be done. Then, through training of the neural network model to estimate kernel weights transition can be fed to hardware design. There, hardware constrained high-level synthesis to optimize the registry-transfer-level coding shows up as the right way for implementation due to large sizes of the resulting circuits. These steps need collaboration of experimental physicists, computer scientists and circuit designers to well understand and to work out practices that will be suited optimally for building neural processors for scientific data.

At present, High Energy Physics (HEP) experiments develop machine learning analysis software and GPUs, DSPs or FPGAs as hardware allowing re-programmability and versatility. Experiments such as Minerva, DUNE or experiments on the LHC use neural networks for data analysis [1, 2, 3]. Until today, however, Neuromorphic Computing algorithms, requiring non Von-Neumann hardware development have been in early stage, and have realistically been not used for any high-scale application. Neuromorphic computing based on GPUs or FPGAs has huge latency and is definitively not a power efficient solution as it involves an enormous number of read and write operations between memory and computation units.

The approach, consisting in sequential processing of computation sequences, has limited benefits for edge computing. ROIC can be developed, unlike general purpose processors, free of this limit. Parallelization, in-hardware embedded matrix operations and time-encoding of information [11], instead of electrical quantities (voltage or current), are postulated investigations. Here, in-memory computing architectures based on crossbar mesh can be given as an example of how the challenging matrix scalar algebra (dot-

product) can be achieved, overcoming the quoted limitations. In particular, such matrix operation architectures could be realized using conventional Static Random-Access Memory (SRAM) [5-6] or a non-volatile memory element such as a Memristor [7-8]. Non-volatile memories offer lower area and power requirements than traditional memory elements. Hence, they are a strong contender for using them as memory in edge computing hardware [4]. As it was mentioned earlier indicating the co-designing methodology as an important approach for conceptual development of AI circuits, co-designing and heterogeneity needs to be extended over developing hardware. Memristors are currently not included in any mainstream fabrication process. Their fabrication, characterization, and integration with conventional CMOS circuits is planned and if it is successful, it would be a major breakthrough in the HEP community.

Each AI method sets different architectural and programming requirements, and each manages data differently. There are also synergies, for example ASIC embedded AI overlaps with other parallelly developed concepts such as event-driven processing and data readout in the experiments all the way down to the underlying analog and digital circuits. Common is processing new events that carry information when possible, not raw data. There are several domains that could immediately benefit from AI based ASICs. The first would be waveform processing with digital interpolating filters for processing of sampled waveform, e.g. digital peak finding in readout circuits for liquid noble gas Time Projection Chamber (TPC) or time of arrival measurement in 4D tracking. The second would be analyzes of spatial distributions of signals with enhancing 2D or 3D spatial resolution and data reduction filtering, e.g. solving charge or light sharing problems in pixel detectors or extraction of depth of interaction that is desired in PET scanners. The last, here but not exhaustive in general, would be contextual analyzes, comprising processing of multi-source waveforms in radiation detection systems for on-the fly spatially sensitive event reconstruction, e.g. processing of signals from arrays of detector electrodes or from detector subsystems in concentrators.

As the last element, it is stressed that the proposed implementation methodology could be conducted either by exploiting High Level Synthesis (HLS) tools [9-10] or novel in-memory computing techniques. Creating benchmarks for a novel software and hardware co-design approach resulting in an energy-efficient, low-latency neuromorphic network that could be designed, fabricated and deployed at the detector readout circuitry is a part of the intended work [12].

References:

[1] J. Duarte et.al., “Fast inference of deep neural networks in FPGAs for particle physics”, JINST, 2018
[2] Microboone Collaboration,“Deep neural network for pixel-level electromagnetic particle identification in the MicroBooNE liquid argon time projection chamber,” Phys. Rev. D99 (2019) No. 9, 092001.
[3] Microboone Collaboration,“Convolutional Neural Networks Applied to Neutrino Events in a Liquid Argon Time Projection Chamber,” JINST 12 (2017) No. 03, P03011.
[4] C. Mead, "Neuromorphic electronic systems," in Proceedings of the IEEE, vol. 78, no. 10, pp. 1629-1636, Oct. 1990, doi: 10.1109/5.58356.
[5] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An Energy-Efficient SRAM With In-Memory Dot-Product Computation for Low-Power Convolutional Neural Networks," in *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 217-230, Jan. 2019
[6] J. Zhang, Z. Wang and N. Verma, "In-Memory Computation of a Machine-Learning Classifier in a Standard 6T SRAM Array," in *IEEE Journal of Solid-State Circuits*, vol. 52, no.4, pp. 915-924, April 2017.
[7] Y. Yang, B. Chen, and W. D. Lu, Adv. Mater. **27**, 7720 (2015). https://doi.org/10.1002/adma.201503202
[8] J.Yang, D.B.Strukov, and D. R. Stewart, Nat. Nanotechnol. **8**, 13 (2013). https://doi.org/10.1038/nnano.2012.240
[9] Mentor Graphics, CATAPULT, https://www.mentor.com/hls-lp/catapult-high-level-synthesis/
[10] Cadence, Tensilica, https://ip.cadence.com/ipportfolio/tensilica-ip
[11] R. Ivans and K. D. Cantley, "A Spatiotemporal Pattern Detector," *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS)*, Dallas, TX, USA, 2019, pp. 444-447, doi: 10.1109/MWSCAS.2019.8884799.
[12] G. Chakma, et al., "A mixed-signal approach to memristive neuromorphic system design," 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, 2017, pp. 547-550, doi: 10.1109/MWSCAS.2017.8052981.