

Snowmass2021 – Letter of Interest

REAL-TIME ADAPTIVE DEEP-LEARNING WITH EMBEDDED SYSTEMS FOR DISCOVERY SCIENCE

Instrumentation Frontier – Thematic Areas:

- ☐ (IF1) Quantum Sensors
- ☐ (IF2) Photon Detectors
- ☐ (IF3) Solid State Detectors and Tracking
- ☒ (IF4) Trigger and DAQ
- ☐ (IF5) Micro Pattern Gas Detectors (MPGDs)
- ☐ (IF6) Calorimetry
- ☒ (IF7) Electronics/ASICs
- ☐ (IF8) Noble Elements
- ☐ (IF9) Cross Cutting and Systems Integration

Computational Frontier – Thematic Areas:

- ☐ (CompF1) Experimental Algorithm Parallelization
- ☐ (CompF2) Theoretical Calculations and Simulation
- ☒ (CompF3) Machine Learning
- ☒ (CompF4) Storage and processing resource access
- ☐ (CompF5) End user analysis
- ☐ (CompF6) Quantum computing
- ☐ (CompF7) Reinterpretation and long-term preservation of data and code

Contact Information:

David W. Miller, University of Chicago, IL
 Contact Email: David.W.Miller@uchicago.edu

Authors:

Andrew A. Chien, University of Chicago and Argonne National Laboratory, IL
 Kristin Dona, University of Chicago, IL
 David W. Miller, University of Chicago, IL
 Jan Offermann, University of Chicago, IL
 Emily Smith, University of Chicago, IL
 Maximilian Swiatlowski, TRIUMF, Vancouver, Canada
 Fukun Tang, University of Chicago, IL
 Nhan Tran, Fermi National Accelerator Laboratory, Batavia, IL

Abstract:

Real-time machine learning holds significant promise for data-intensive science to filter and analyze extreme data rates and bandwidths to enable discoveries. This *Letter of Interest* aims to highlight the need for a comprehensive research program into the many avenues of development of deep-learning algorithm inference for real-time data processing and filtering using both multi-processor system-on-chip technology (MPSoC) and reprogrammable integrated circuits (FPGA) for physics. In particular, we would like to emphasize the potential importance of

heterogeneous on-chip devices such as MPSoC devices, as well as the new generations of devices that specifically target architectures that are optimized for “artificial intelligence” inference tasks. These platforms offer the possibility of powerful physics performance under extreme low latency and high bandwidth constraints, while also providing a wide array of functionality and interfaces, along with high-level software stacks that are able to be implemented and used by physicists. We hope that these ideas can be incorporated into either a dedicated *White Paper* on real-time processing architectures for physics, or a joint report for Snowmass summarizing the potential new directions trigger and data acquisition systems for future physics facilities.

Letter of Interest:

Data-intensive science is increasingly reliant on real-time processing capabilities and machine learning (ML) workflows in order to filter and analyze extreme volumes of data and perform feature extraction and pattern recognition at numerous time scales and levels of complexity. This is especially true at the energy and intensity frontiers of particle physics where bandwidths of raw data can exceed 100 Tb/s of heterogeneous, high-dimensional event records. Individual online filtering algorithms targeted at discovery science often operate at the level of selecting or filtering 1 part in 10^5 , and offline algorithms must then search efficiently through the petabytes of data collected; the exabyte scale is quickly becoming a practical reality. While particle physics is our domain focus, we believe real-time machine learning has significant potential across numerous high-throughput scientific domains. Inference for “on-detector” processing with mixed architecture chips is extremely powerful and broadly applicable for any high-speed processing tasks across the intensity, cosmic, energy frontier physics and beyond, even including the basic energy sciences, nuclear physics, and more.

This *Letter of Interest* seeks to communicate the need for a comprehensive research program of the many avenues of development of deep-learning algorithm inference for real-time data processing and filtering using both multi-processor system-on-chip technology (MPSoC) [1] and reprogrammable integrated circuits (FPGA) [2, 3] for high energy particle physics, as well as novel ML-focused architectures [4, 5]. Such a program needs to balance the instrumentation challenges, the computational hurdles, and the physics goals of these systems. A trigger-less data acquisition (DAQ) system for use in a detector system at a future e^+e^- collider [6] may benefit from significantly different architectures or implementations for real-time pattern recognition compared to multi-tiered trigger system for rejecting event pile-up and “background” events typical of a detector at a hadron-hadron collider [7, 8]. A muon collider, which could have potentially large “beam-induced” backgrounds which arrive out of time and with different angles than particles from a hard-scatter collision, would have again a different set of considerations optimized at removing very different kinds of background. However, all three systems may yet benefit from the same ability to deploy deep convolutional neural networks for low-level feature identification even if they have significantly different latency requirements – and thus constraints on their levels of sophistication – for higher-level event reconstruction and classification. In a different application of SoCs, PID (Proportional-Integral-Derivative) control loops are prevalent in real-time, low-latency controls of accelerator operations for tasks like magnet power supply modulation for ramping and focusing; for example, Arria10 SoCs are deployed for such tasks in the Fermilab accelerator complex. Dynamic, continuous learning AI algorithms have the potential to improve and automatize future accelerator control systems [9–23]. We feel that a dedicated effort to survey the status and future of these approaches will be extremely important for understanding the range of instrumentation possibilities for trigger, DAQ, and control systems at future physics facilities.

Recent advances in both the *algorithm* architectures that enable high-performance machine learning as well as the *hardware* architectures that can perform these tasks provide a unique opportunity at the intersection of particle physics, computer science (CS), and electronics engineering [24–27]. By embracing this nexus of scientific sub-domains it is conceivable to develop systems that can process input data from the detectors at the particle colliders more efficiently and with greater accuracy than current approaches. Even more specifically, the authors would like to advocate for the development and deployment of machine learning algorithms for processing high-rate signals using high-performance FPGA/MPSoC architectures including real-time adaptation and dynamic model implementation. Scientific and engineering investment is needed to understand the potential applications of the

numerous available hardware and software architectures and systems engineering choices to the physics challenges that lie ahead. For example, flexible hardware structures that implement machine learning accelerators may be coupled with flexible software structures and yet still satisfy hard-real time computation requirements.

In addition to algorithm architectures and hardware systems, the software stacks and workflows that are necessary to develop and deploy algorithms to embedded hardware systems is equally as important. Industry leaders such as Intel and Xilinx provide software development platforms for their proprietary hardware systems. Moreover, physicists and engineers have developed the HLS4ML [24] suite of tools for ease of deploying ML models to these devices.

Prototype trigger and processing systems using both FPGAs and MPSoC are currently being developed by multiple groups around the US and the world for use in Runs 3 and 4 of the LHC. These systems – along with the workflows required to develop, implement, deploy, commission, validate, and use deep learning algorithms for physics results – should be carefully evaluated and assessed for much more widespread usage at future facilities that might rely on such approaches for the fundamental framework of their trigger and DAQ systems.

An eventual White Paper for Snowmass 2021 should aim to review the current status of the development of real-time embedded systems for use in Trigger and DAQ architectures, as well as to establish a set of key milestones and demonstrator systems for future physics facilities. Specific consideration should be given to the likely prevalence of highly heterogeneous compute systems: Systems with a mixture of both data-center-like and cloud-like resources; specialized architectures like GPUs; real-time embedded systems like FPGAs; on-chip heterogeneous systems like MPSoC, Xilinx Versal, and Intel Agilex devices.

In summary, discoveries in particle physics may hinge on the ability to parse exponentially larger datasets using commensurately improved algorithms and at low latencies. Real-time machine-learning inference may have a significant impact on the discovery potential of the LHC experiments. Clear directions towards this future are already obvious, however the space of possibilities is enormous and the technical hurdles are significant. Given the growing importance of edge computing and algorithm codesign in the broader machine learning community, this effort is timely and should connect to broader instrumentation opportunities within HEP and beyond.

References

- [1] XILINX. Zynq UltraScale+ MPSoC. URL: <https://www.xilinx.com/products/silicon-devices/soc/zynq-ultrascale-mpsoc.html>
- [2] INTEL. Intel FPGA Family. URL: <https://www.intel.com/content/www/us/en/products/programmable/fpga.html>
- [3] XILINX. Xilinx Ultrascale+ FPGA. URL: <https://www.xilinx.com/products/silicon-devices/fpga/virtex-ultrascale-plus.html>
- [4] XILINX. Zynq Versal ACAP. URL: <https://www.xilinx.com/products/silicon-devices/acap/versal.html>
- [5] INTEL. Intel Agilex FPGAs and SoCs Family. URL: <https://www.intel.com/content/www/us/en/products/programmable/fpga/agilex.html>
- [6] A. ABADA et al. “FCC-ee: The Lepton Collider”. *The European Physical Journal Special Topics*, **228**:2, 261–623, 2019.
- [7] ATLAS COLLABORATION. *Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System*. Tech. rep. CERN-LHCC-2017-020. ATLAS-TDR-029. Geneva: CERN, Sept. 2017
- [8] CMS COLLABORATION. *The Phase-2 Upgrade of the CMS Level-1 Trigger*. Tech. rep. CERN-LHCC-2020-004. CMS-TDR-021. Final version. Geneva: CERN, Apr. 2020
- [9] M.-R. BOUGUELIA, Y. BELAÏD, and A. BELAÏD. “A stream-based semi-supervised active learning approach for document classification”. In: *2013 12th International Conference on Document Analysis and Recognition*. IEEE. 2013. 611–615
- [10] W. CHU, M. ZINKEVICH, L. LI, A. THOMAS, and B. TSENG. “Unbiased online active learning in data streams”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011. 195–203
- [11] C. C. LOY, T. M. HOSPEDALES, T. XIANG, and S. GONG. “Stream-based joint exploration-exploitation active learning”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. 2012. 1560–1567
- [12] B. SETTLES. *Active Learning*. Synthesis digital library of engineering and computer science. Morgan & Claypool Publishers, 2012.
- [13] Y. LIU. “Active learning with support vector machine applied to gene expression data for cancer classification”. *Journal of chemical information and computer sciences*, **44**:6, 1936–1941, 2004.

- [14] S. DASGUPTA. “Analysis of a greedy active learning strategy”. In: *Advances in Neural Information Processing Systems 17*. 2005.
- [15] M.-F. BALCAN, A. BEYGELZIMER, and J. LANGFORD. “Agnostic Active Learning”. *J. Comput. Syst. Sci.*, **75**:1, 78–89, 2009.
- [16] K. KONYUSHKOVA, R. SZNITMAN, and P. FUA. “Learning active learning from data”. In: *Advances in Neural Information Processing Systems*. 2017. 4225–4235
- [17] Y. CHEN, S. H. HASSANI, and A. KRAUSE. Near-optimal Bayesian Active Learning with Correlated and Noisy Tests. Tech. rep. arXiv, May 2016
- [18] Y. CHEN, S. JAVDANI, A. KARBASI, J. A. BAGNELL, S. SRINIVASA, and A. KRAUSE. “Submodular Surrogates for Value of Information”. In: *Proc. Conference on Artificial Intelligence (AAAI)*. Jan. 2015.
- [19] Y. CHEN and A. KRAUSE. “Near-optimal Batch Mode Active Learning and Adaptive Submodular Optimization”. In: *International Conference on Machine Learning (ICML)*. 2013.
- [20] Y. CHEN, A. SINGLA, O. M. AODHA, P. PERONA, and Y. YUE. “Understanding the Role of Adaptivity in Machine Teaching: The Case of Version Space Learners”. In: *Proc. Conference on Neural Information Processing Systems (NeurIPS)*. Dec. 2018.
- [21] Y. CHEN, J.-M. RENDERS, M. H. CHEHREGHANI, and A. KRAUSE. “Efficient Online Learning for Optimizing Value of Information: Theory and Application to Interactive Troubleshooting”. *arXiv preprint arXiv:1703.05452*, 2017.
- [22] C. SZEPESVÁRI. “Algorithms for reinforcement learning”. *Morgan and Claypool*, 2009.
- [23] S. SHALEV-SHWARTZ et al. “Online learning and online convex optimization”. *Foundations and trends in Machine Learning*, **4**:2, 107–194, 2011.
- [24] J. DUARTE et al. “Fast inference of deep neural networks in FPGAs for particle physics”. *JINST*, **13**:07, P07027, 2018.
- [25] S. SUMMERS et al. “Fast inference of Boosted Decision Trees in FPGAs for particle physics”. *JINST*, **15**:05, P05026, 2020.
- [26] V. LONCAR et al. “Compressing deep neural networks on FPGAs to binary and ternary precision with HLS4ML”. 2020.
- [27] C. N. COELHO, A. KUUSELA, H. ZHUANG, T. AARRESTAD, V. LONCAR, J. NGADIUBA, M. PIERINI, and S. SUMMERS. “Ultra Low-latency, Low-area Inference Accelerators using Heterogeneous Deep Quantization with QKeras and hls4ml”. 2020.