Fast ML White Paper

Allison Deiana on behalf of the authors

Motivation

- Fast Machine Learning for Science Workshop was held 30 November – 3 December, hosted virtually by Southern Methodist University
  - Website available here: https://indico.cern.ch/event/924283/
- Workshop was interdisciplinary and attracted over 500 participants, talks on a wide variety of scientific applications.
- Workshop also included a hands-on tutorial session, to get people started on applications of fast machine learning.
- After the workshop, a community white paper has been prepared, and has been submitted to a special issue of Frontiers in AI.

Status of White Paper

Computer Science > Machine Learning

[Submitted on 25 Oct 2021]

Applications and Techniques for Fast Machine Learning in Science

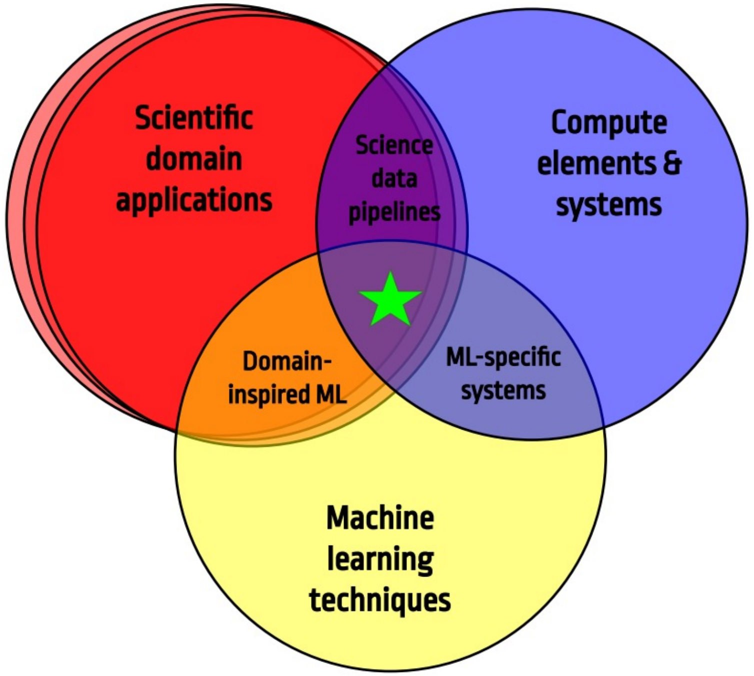
Allison McCarn Deiana (coordinator), Nhan Tran (coordinator), Joshua Agar, Michaela Blott, Giuseppe Di Guglielmo, Javier Duarte, Philip Harris, Scott Hauck, Mia Liu, Mark S. Neubauer, Jennifer Ngadiuba, Seda Ogrenci-Memik, Maurizio Pierini, Thea Aarrestad, Steffen Bahr, Jurgen Becker, Anne-Sophie Berthold, Richard J. Bonventre, Tomas E. Muller Bravo, Markus Diefenthaler, Zhen Dong, Nick Fritzsche, Amir Gholami, Ekaterina Govorkova, Kyle J Hazelwood, Christian Herwig, Babar Khan, Sehoon Kim, Thomas Klijnsma, Yaling Liu, Kin Ho Lo, Tri Nguyen, Gianantonio Pezzullo, Seyedramin Rasoulinezhad, Ryan A. Rivera, Kate Scholberg, Justin Selig, Sougata Sen, Dmitri Strukov, William Tang, Savannah Thais, Kai Lukas Unger, Ricardo Vilalta, Belinavon Krosigk, Thomas K. Warburton, Maria Acosta Flechas, Anthony Aportela, Thomas Calvet, Leonardo Cristella, Daniel Diaz, Caterina Doglioni, Maria Domenica Galati, Elham E Khoda, Farah Fahim, Davide Giri, Benjamin Hawks, Duc Hoang, Burt Holzman, Shih-Chieh Hsu, Sergo Jindariani, Iris Johnson, Raghav Kansal, Ryan Kastner, Erik Katsavounidis, Jeffrey Krupa, Pan Li, Sandeep Madireddy, Ethan Marx, Patrick McCormack, Andres Meza, Jovan Mitrevski, Mohammed Attia Mohammed, Farouk Mokhtar, Eric Moreno, Srishti Nagu, Rohin Narayan, Noah Palladino, Zhiqiang Que, Sang Eon Park, Subramanian Ramamoorthy, Dylan Rankin, Simon Rothman, Ashish Sharma, Sioni Summers, Pietro Vischia, Jean-Roch Vlimant, Olivia Weng

In this community review report, we discuss applications and techniques for fast machine learning (ML) in science -- the concept of integrating power ML methods into the real-time experimental data processing loop to accelerate scientific discovery. The material for the report builds on two workshops held by the Fast ML for Science community and covers three main areas: applications for fast ML across a number of scientific domains; techniques for training and implementing performant and resource-efficient ML algorithms; and computing architectures, platforms, and technologies for deploying these algorithms. We also present overlapping challenges across the multiple scientific domains where common solutions can be found. This community report is intended to give plenty of examples and inspiration for scientific discovery through integrated and accelerated ML solutions. This is followed by a high-level overview and organization of technical advances, including an abundance of pointers to source material, which can enable these breakthroughs.

Comments:	66 pages, 13 figures, 5 tables
Subjects:	Machine Learning (cs.LG); Hardware Architecture (cs.AR); Data Analysis, Statistics and Probability (physics.data-an); Instrumentation and Detectors (physics.ins-det)
Report number:	FERMILAB-PUB-21-502-AD-E-SCD
Cite as:	arXiv:2110.13041 [cs.LG]
	(or arXiv:2110.13041v1 [cs.LG] for this version)

- Available on arXiv at the link: https://arxiv.org/abs/2110.13041
- Currently in the review process with Frontiers in AI

Content of White Paper



Contents

This report aims to summarize the progress in the community to understand how our scientific challenges overlap and where there are potential commonalities in data representations, ML approaches, and technology, including hardware and software platforms. Therefore, **the content of the report includes the following: descriptions of a number of different scientific domains including existing work and applications for embedded ML; potential overlaps across scientific domains in data representation or system constraints; and an overview of state-of-the-art techniques for efficient machine learning and compute platforms, both cutting-edge and speculative technologies**.

Section 2: Domain Examples

- Large section on Large Hadron Collider for:
  - Event Reconstruction
  - Event Simulation
  - Heterogeneous Computing
  - Real-Time Analysis at 40 MHz
  - Bringing ML to Detector Front-End

Example use cases are not comprehensive, but representative.

Discussion included on tools used for fast machine learning – hls4ml and conifer.

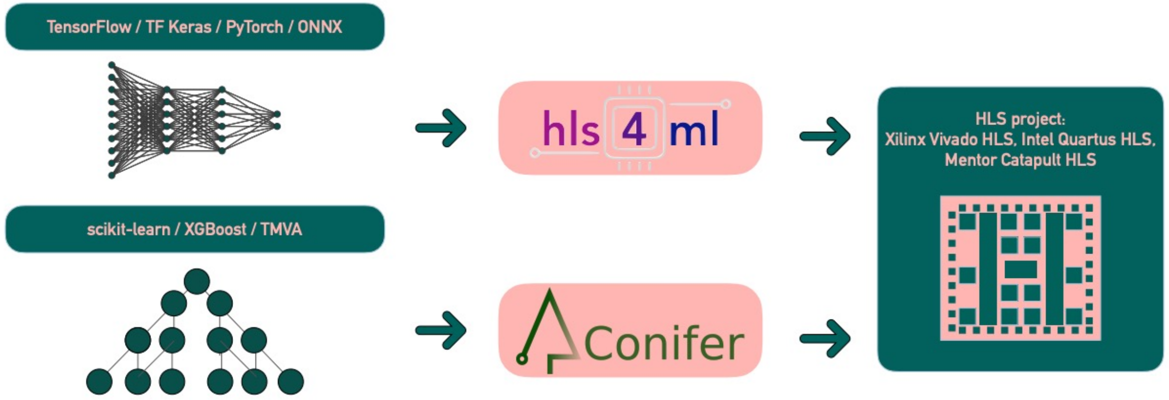


Figure 3. Two dedicated libraries for the conversion of Machine Learning algorithms into FPGA or ASIC firmware: `hls4ml` for deep neural network architectures and `Conifer` for Boosted Decision Tree architectures. Models from a wide range of open-source ML libraries are supported and may be converted using three different high-level synthesis backends.

Section 2: Domain Examples

- High-intensity Accelerators: Belle II, Mu2e
- Materials Discovery: Materials Synthesis, Scanning Probe Miscroscopy
- Fermilab Accelerator Controls
- Neutrino/Dark Matter Experiments: e.g. DUNE, MINERvA, Direct Detection Dark Matter
- Electron-Ion Collider
- Gravitational Waves
- Health: Biomedical Engineering and Health Monitoring
- Cosmology
- Plasma Physics
- Wireless Networking and Edge Computing

Section 3: Data Representation

Domain	Spatial	Point Cloud	Temporal	Spatio-Temporal	Multi/Hyper-spectral	Examples
LHC	✓✓	✓✓	✓	✓	–	detector reconstruction
Belle-II/Mu2e	✓✓	✓✓	–	–	–	track reconstruction
Material Synthesis	✓	–	✓	✓✓	✓✓	high-speed plasma imaging
Accelerator Controls	✓	–	✓✓	–	–	beam sensors
Accelerator neutrino	✓✓	✓✓	✓	✓	–	detector reconstruction
Direct detection DM	✓✓	✓✓	✓	✓	–	energy signatures
EIC	✓✓	✓✓	✓	✓	–	detector reconstruction
Gravitational Waves	✓	–	✓✓	–	–	laser inference patterns
Biomedical engineering	✓✓	–	–	✓✓	–	cell and tissue images
Health Monitoring	✓	–	✓✓	✓	✓	physiological sensor data
Cosmology	✓✓	✓✓	✓✓	✓	✓✓	lensing/radiation maps
Plasma Physics	✓	–	✓✓	✓	–	detector actuator signals
Wireless networking	–	–	✓✓	–	–	electromagnetic spectrum

Types of data representation that are relevant for different domains.

Section 3: System Constraints

Table 2. Domains and practical constraints: systems are broadly classified as soft (software-programmable computing devices: CPUs, GPUs, and TPUs) and custom (custom embedded computing devices: FPGAs and ASICs)

Domain	Event Rate	Latency	Systems	Energy-constrained
Detection and Event Reconstruction				**No**
LHC & intensity frontier HEP	10s Mhz	ns-ms	Soft/custom	
Nuclear physics	10s kHz	ms	soft	
Dark matter & neutrino physics	10s MHz	μs	Soft/custom	
Image Processing				
Material synthesis	10s kHz	ms	Soft/custom	
Scanning probe microscopy	kHz	ms	Soft/custom	
Electron microscopy	MHz	μs	Soft/custom	
Biomedical engineering	kHz	ms	Soft/custom	Yes (mobile settings)
Cosmology	Hz	s	soft	
Astrophysics	kHz–MHz	ms-us	Soft	Yes (remote locations)
Signal Processing				
Gravitational waves	kHz	ms	Soft	
Health monitoring	kHz	ms	Custom	Yes
Communications	kHz	ms	Soft	Yes (mobile settings)
Control Systems				
Accelerator controls	kHz	ms–μs	Soft/custom	
Plasma physics	kHz	ms	Soft	

Section 4: Efficient ML

- A discussion of strategies for improving ML efficiency to enable lower latency.
  - Designing new efficient ML architectures
  - NN & hardware co-design
  - Quantization
  - Pruning and sparse inference
  - Knowledge distillation

- Discussion of automation of the NN architecture design process (Neural Architecture Search).

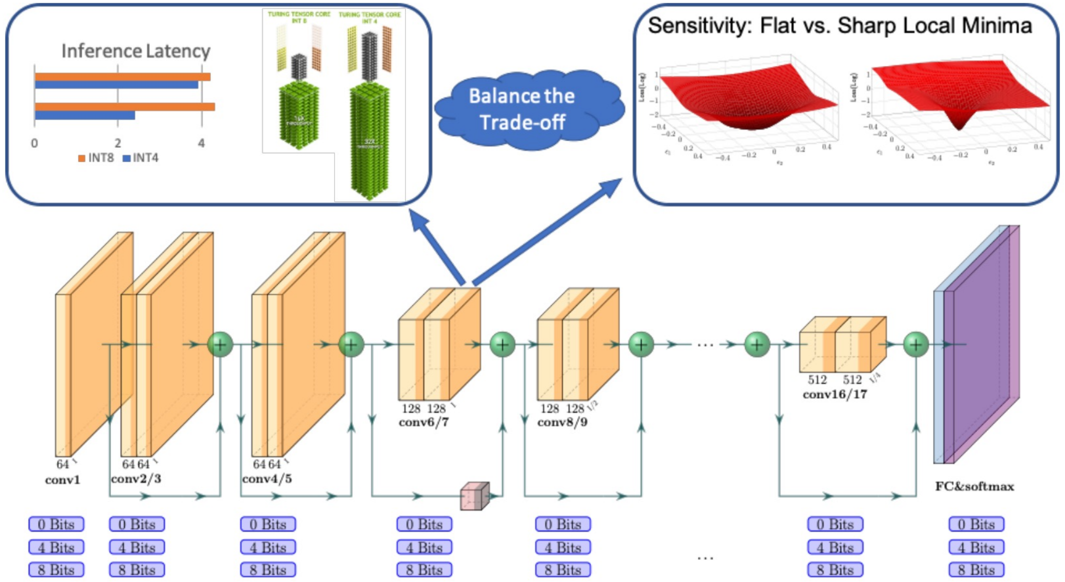


Figure 9. The illustration of hardware-aware quantization and pruning. A given NN model can be compressed by using low precision quantization instead of single precision. The extreme case is to use 0-bit quantization which is equivalent to removing/pruning the corresponding neurons. The goal of compression is to find the best bit-precision setting for quantization/pruning to reduce model footprint/latency on a target hardware with minimal generalization loss.

Section 4: Hardware Architecture

- Discussion of different computing architectures: CPU, GPU, FPGA/ASIC
- DPU: Deep learning processing unit, customized for CNNs. These can be implemented on FPGAs or ASICs.

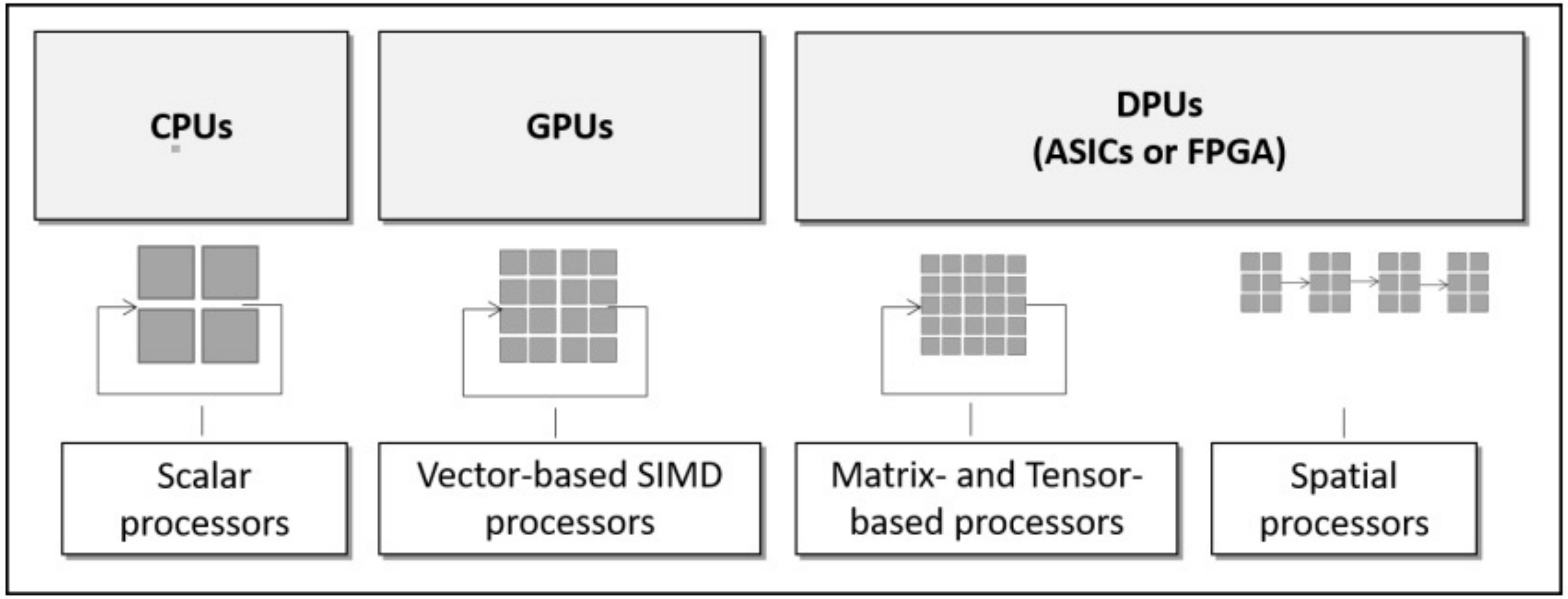


Figure 10. Taxonomy of compute architectures, differentiating CPUs, GPUs and DPUs

Section 4: Hardware/Software Co-Design

- Discussion of design, and of frameworks specifically created for the ML domain where they automate the process of hardware generation for the end-user thus hiding the associated design complexity of FPGAs and enabling them for the previously discussed end applications.
  - hls4ml
  - FINN

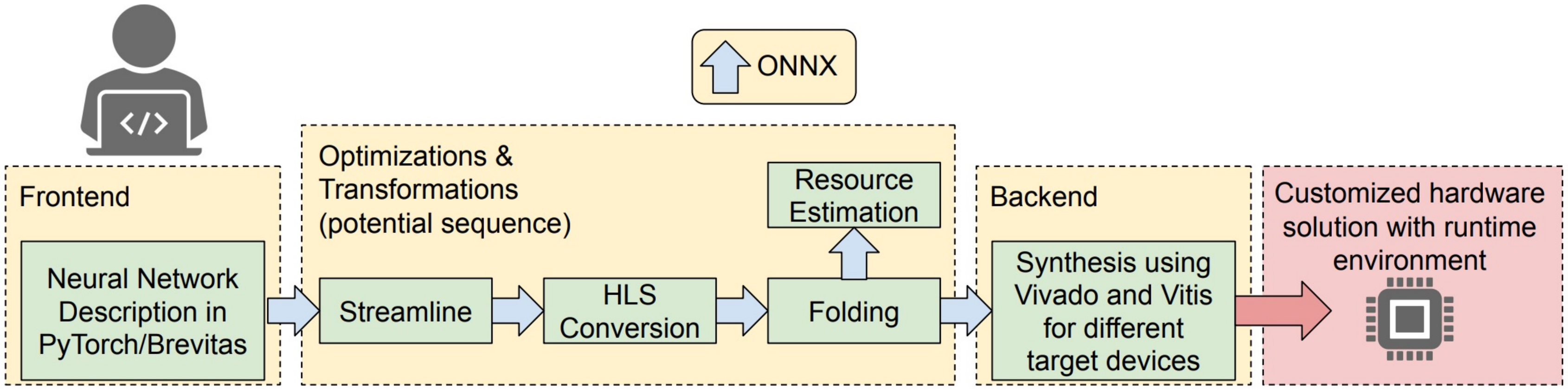


Figure 12. FINN Compiler Flow

Section 4: Beyond-CMOS Neuromorphic Hardware

- In this section, the most prominent emerging technology proposals, including those based on emerging dense analog memory device circuits, are grouped according to the targeted low-level neuromorphic functionality.
  - Analog Vector-by-Matrix Multiplication
  - Stochastic Vector-by-Matrix Multiplication
  - Spiking Neuron and Synaptic Plasticity
  - Reservoir Computing
  - Hyperdimensional Computing / Associative Memory

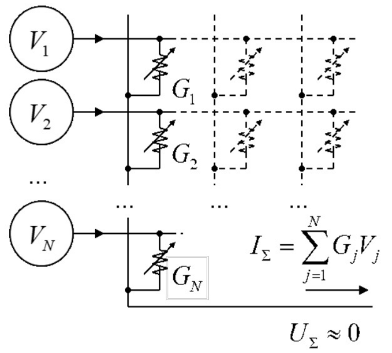


Figure 13. Analog vector-by-matrix multiplication (VMM) in a crossbar circuit with adjustable crosspoint devices. For clarity, the output signal is shown for just one column of the array, while sense amplifier circuitry is not shown. Note that other VMM designs, e.g. utilizing duration of applied voltage pulses, rather than their amplitudes, for encoding inputs/outputs, are now being actively explored – see, e.g., their brief review in Ref. [551]

Conclusion

Reminder: Full document is available on arXiv for those interested:
https://arxiv.org/abs/2110.13041

White Paper is not comprehensive but does cover many example use cases of fast machine learning, overlap between scientific domains, and a review of state-of-the-art technology.

Connection to Snowmass process: Can summarize/borrow from most relevant parts of full white paper (with an updated introduction more aligned to Snowmass process to be submitted as a Snowmass white paper → Contact person: Javier Duarte

Thank you for your attention!