Smart sensors using artificial intelligence for on-detector electronics and ASICs

Sandeep Miryala, BNL
Nhan Tran, FNAL
On behalf of the authors

July 21, 2022
IF07 - Seattle Snowmass Summer Meeting 2022

Snowmass 2021

Smart sensors using artificial intelligence for on-detector electronics and ASICs

Gabriella Carini, Grzegorz Deptuch, Jin Huang, Soumyajit Mandal, Sandeep Miryala, Veljko Radeka, Yihui Ren
Brookhaven National Laboratory
Jennet Dickinson, Farah Fahim, Christian Herwig, Cristina Mantilla Suarez, Benjamin Parpillon, Nhan Tran
Fermi National Accelerator Laboratory
Philip Harris, Dylan Rankin
Massachusetts Institute of Technology
Dionisio Doering, Angelo Dragone, Ryan Herbst, Lorenzo Rota, Larry Ruckman
SLAC National Accelerator Laboratory
Allison McCarn Deiana
Southern Methodist University
F. Mitchell Newcomer
University of Pennsylvania

ABSTRACT

Cutting edge detectors push sensing technology by further improving spatial and temporal resolution, increasing detector area and volume, and generally reducing backgrounds and noise. This has led to a explosion of more and more data being generated in next-generation experiments. Therefore, the need for near-sensor, at the data source, processing with more powerful algorithms is becoming increasingly important to more efficiently capture the right experimental data, reduce downstream system complexity, and enable faster and lower-power feedback loops. In this paper, we discuss the motivations and potential applications for on-detector AI. Furthermore, the unique requirements of particle physics can uniquely drive the development of novel AI hardware and design tools. We describe existing modern work for particle physics in this area. Finally, we outline a number of areas of opportunity where we can advance machine learning techniques, codesign workflows, and future microelectronics technologies which will accelerate design, performance, and implementations for next generation experiments.

Submitted to the Proceedings of the US Community Study
on the Future of Particle Physics (Snowmass 2021)

Snowmass 2021

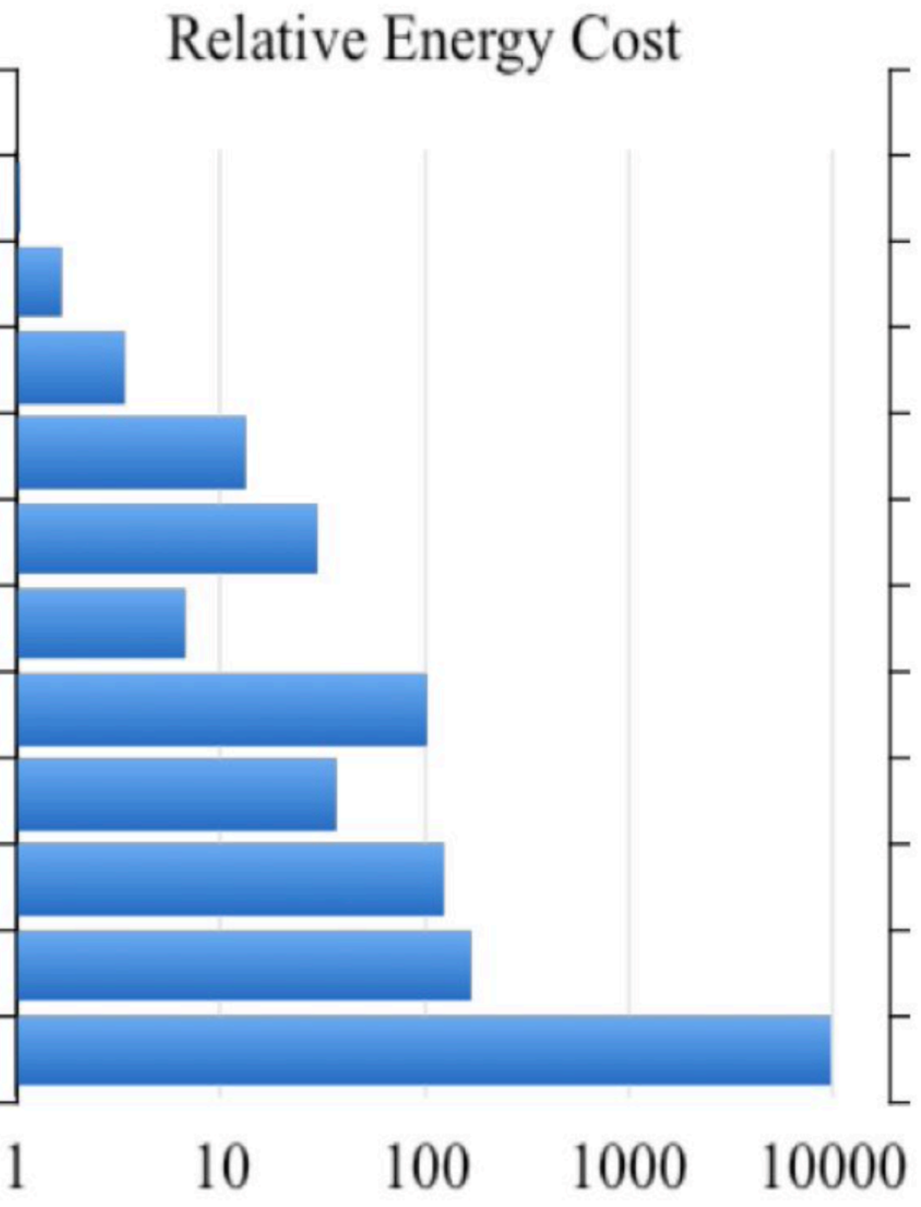
Contents

Snowmass 2021

2. Science drivers

- Scientific discoveries are enabled by probing nature at higher spatial and temporal precision
  - Results in rapidly growing scientific data pipelines!
    - Complex and rich data - powerful algos
  - Data transmission is far less efficient than data processing
- **Explore the power of AI-at-source!**

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640



Adapted from Horowitz

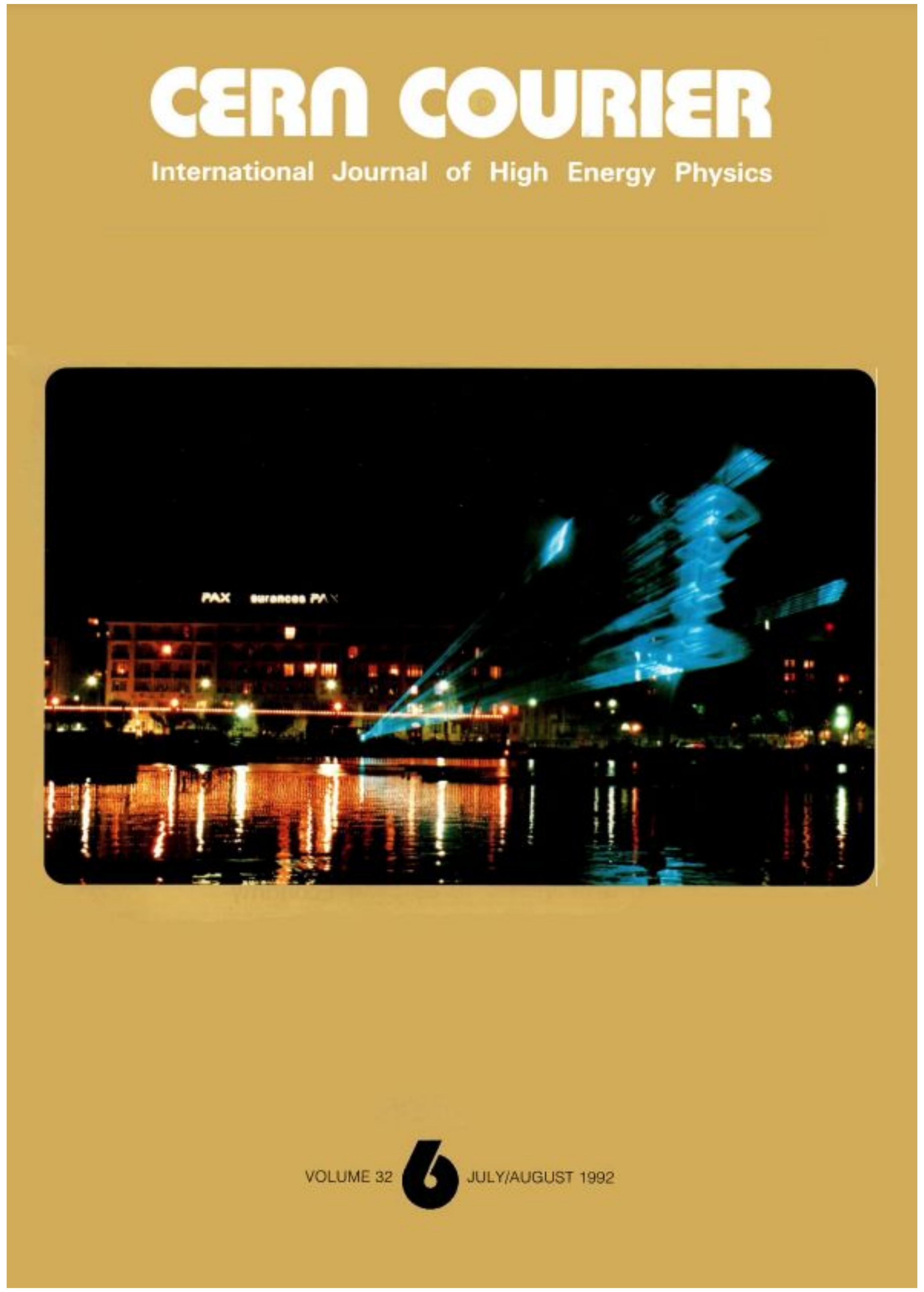
2. Science drivers

- Extreme environments in HEP experiments (power, rate, radiation, cryo,…)
  - AI in near-detector electronics is natural evolution
  - can be a driver for progress in other scientific domains
- **Benefits:**
  - ML algorithms can enable powerful and efficient non-linear data reduction or feature extraction techniques, **preserves the physics content** that would otherwise be lost;
  - **Reduce the complexity** of down stream processing systems and **transmit less overall information**
  - Enables real-time data filtering and triggering which would otherwise not be possible or be much less efficient; or in the case of cryogenic systems, creates less data bandwidth from cold to warm electronics and thus reduce the system complexity;
  - **Enable faster feedback loops** - e.g., in continuous learning applications where data is part of control or operations loop such as in quantum information systems or particle accelerators

3. Community needs

https://cds.cern.ch/record/1732048

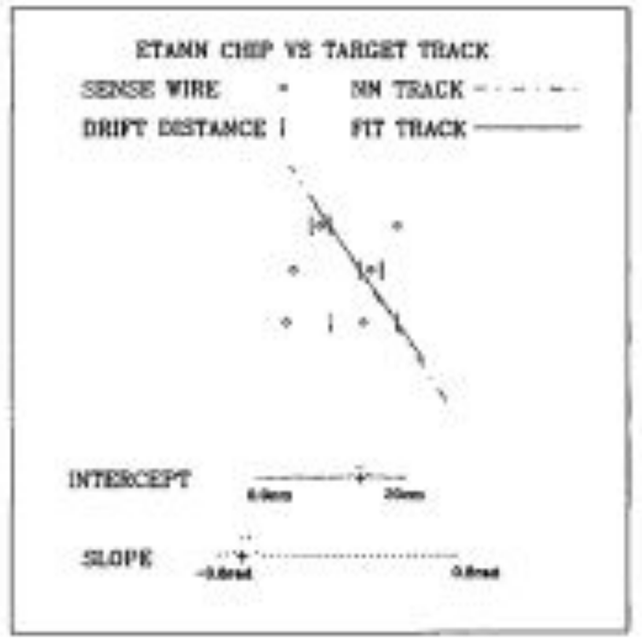
- This is not a new idea :)



ware of a high energy physics experiment.

The first such application comes from a recent Fermilab test beam experiment, where a VLSI neural network chip was interfaced to the data acquisition system of a prototype drift chamber. Drift time information from the sense wires, encoded as voltages, was passed to the neural network, which calculated the slope and intercept of the track traversing the chamber and sent this information back to the mother readout board to be read out with the rest of the event, without any dead time.

Neural network hardware is also finding its way into other trigger systems. The CDF experiment has



three neural network triggers in place for its 1992 run: an isolated endplug electron trigger, an isolated central photon trigger, and a semileptonic B particle trigger.

Also at Fermilab's Tevatron collider, a group in the D0 experiment is studying the use of neural networks in the muon trigger for the D0 Muon Upgrade. A neural network trigger for H1 at DESY has been under development for some time and will be tested in the current run. Several R&D projects at CERN are looking at the feasibility of neural networks for LHC experiment trigger systems.

Another application of neural networks under study is in adaptive control systems for accelerators. A group at SLAC recently simulated how a neural network control system could be trained both to emulate and control a section of beamline.

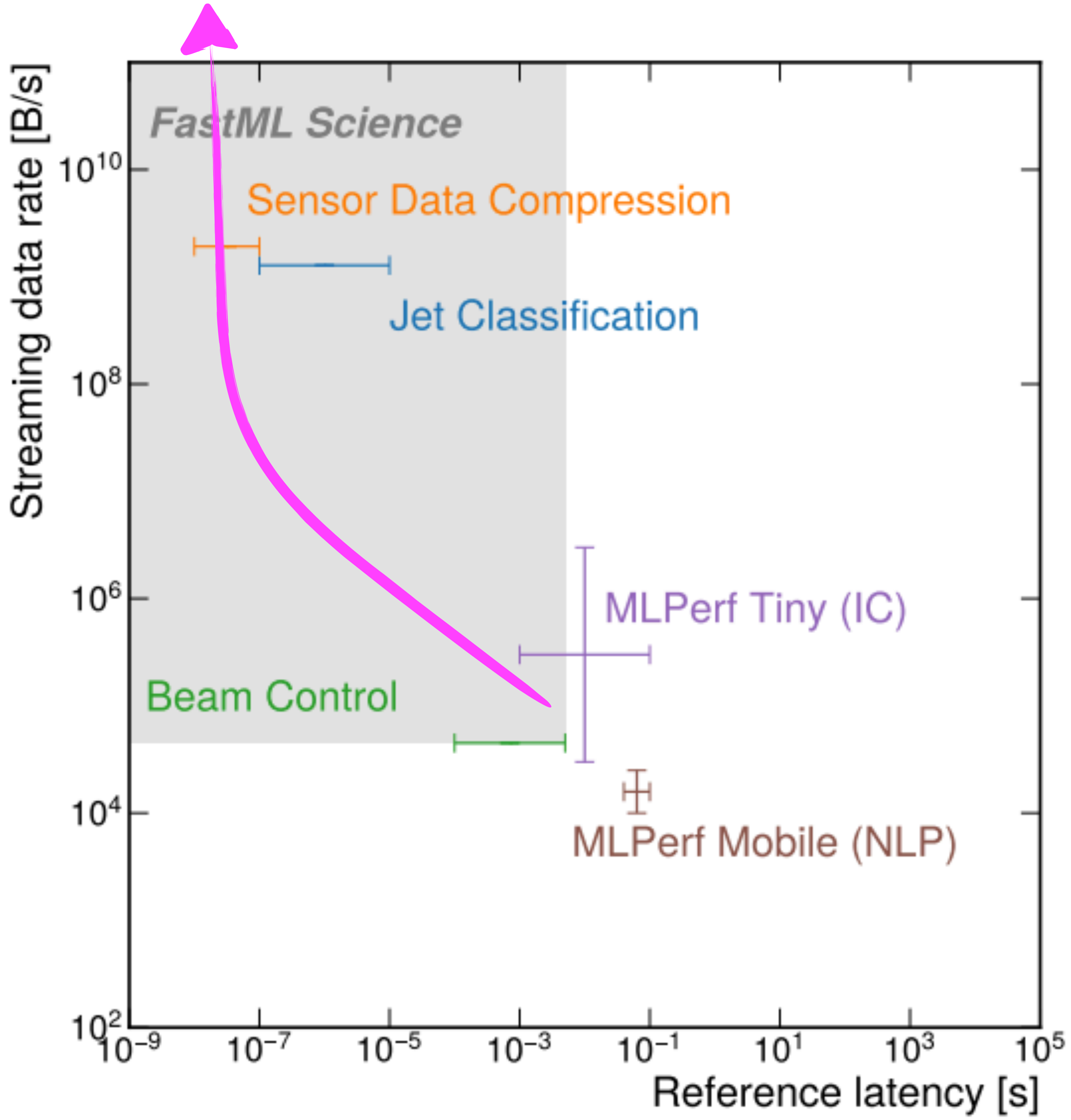
These new artificial intelligence techniques could go on to play an important role in the acquisition and analysis of experimental data for the coming generation of proton colliders.

From Bruce Denby and Clark Lindsey (Fermilab) and Louis Lyons (Oxford)

3. Community needs

https://cds.cern.ch/record/1732048

- This is not a new idea :)
- What's changed?
  - **Broader necessity**
    - Moore's Law has stalled - can't just rely on more datacenter compute
    - Internet-of-Things is growing rapidly
  - **Advances in hardware**
  - **Advances in ML**
  - **Advances in codesign tools**
- **But**, we have even harder problems than industry and other scientific applications - stimulates innovation!



4. Existing work

- application: CMS HGCal ECON data encoder
- tools: hls4ml for ML codesign of ICs
- application: NNs for waveform processing

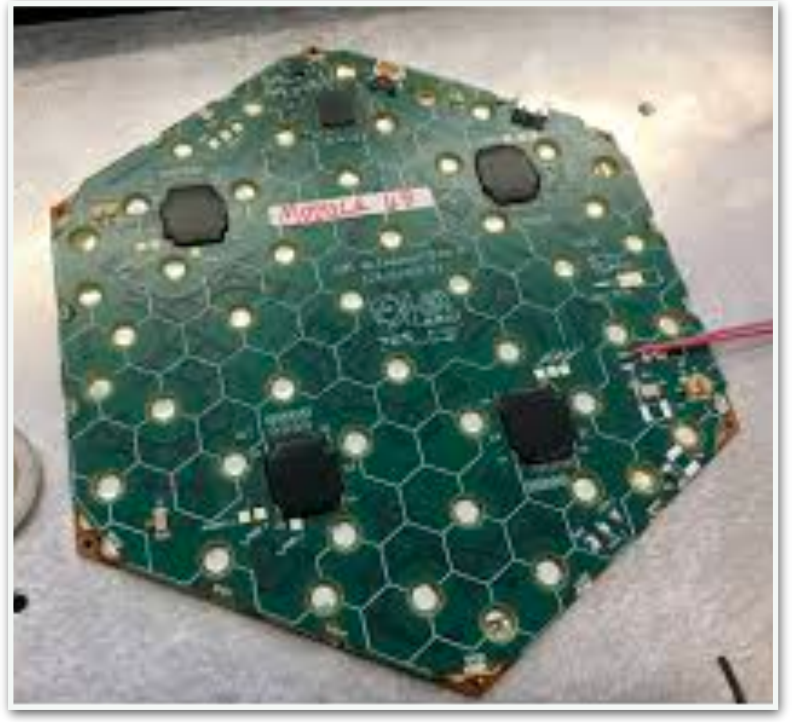
CMS HGCal data compression

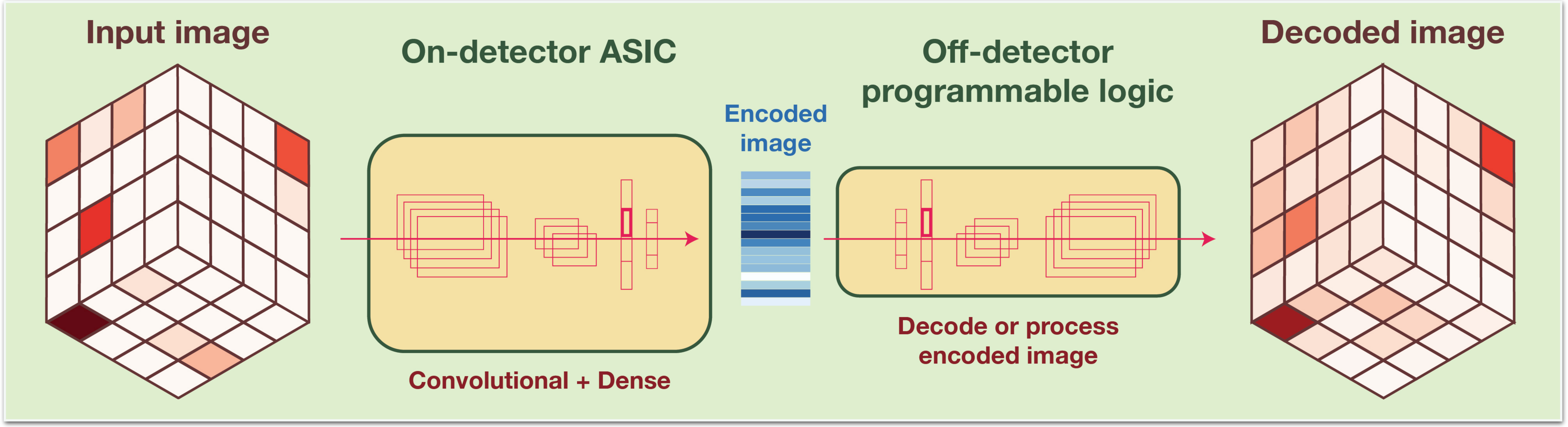
https://arxiv.org/abs/2105.01683

The task:

Trigger path stage	Number channels	bits/channel	Average Compression factor	Data rate*	# links* (10.24 Gbps)
Raw data	6M	20	1	5 Pb/s	1M
Hardware reduction	1M	7	1	300 Tb/s	60k
Threshold selection	1M	7	7	40 Tb/s	9k

The concept:

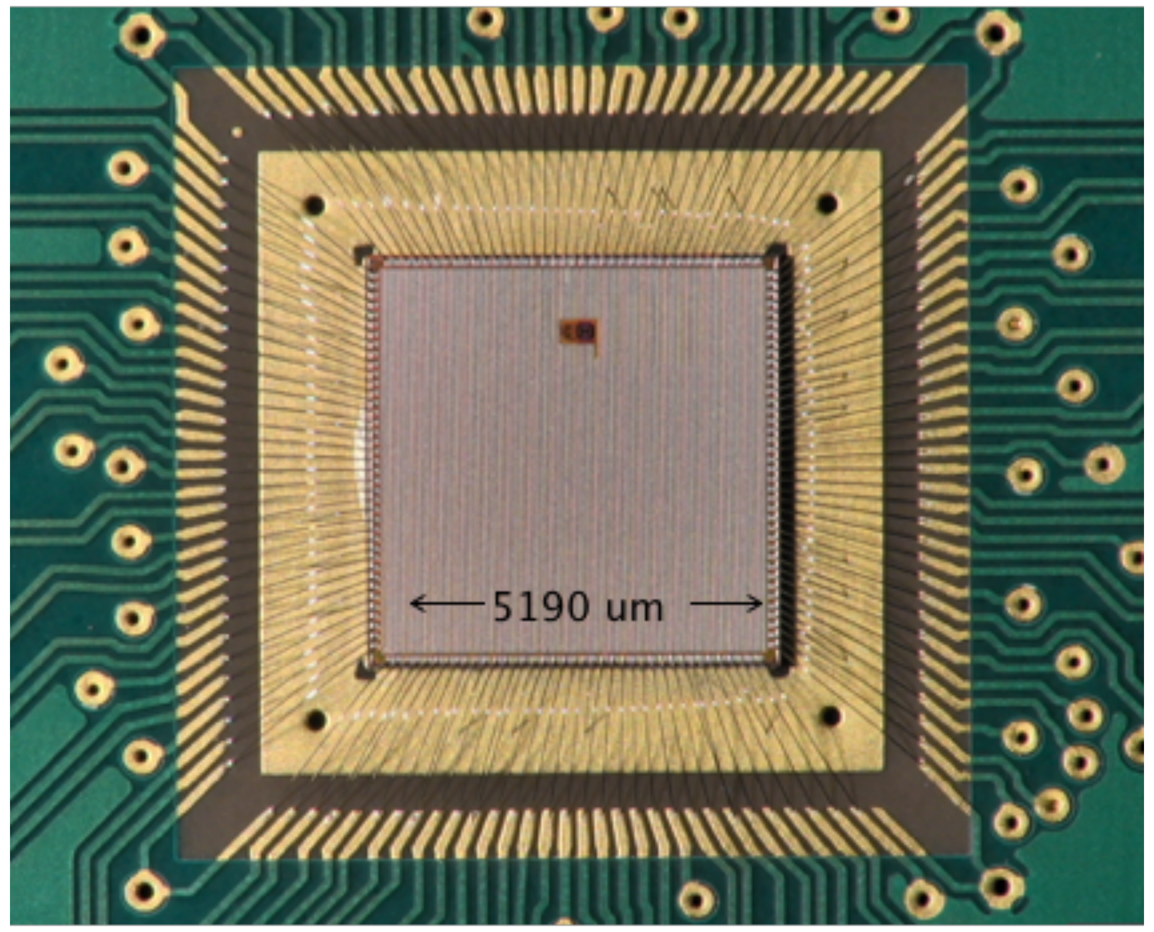




CMS HGCal data compression

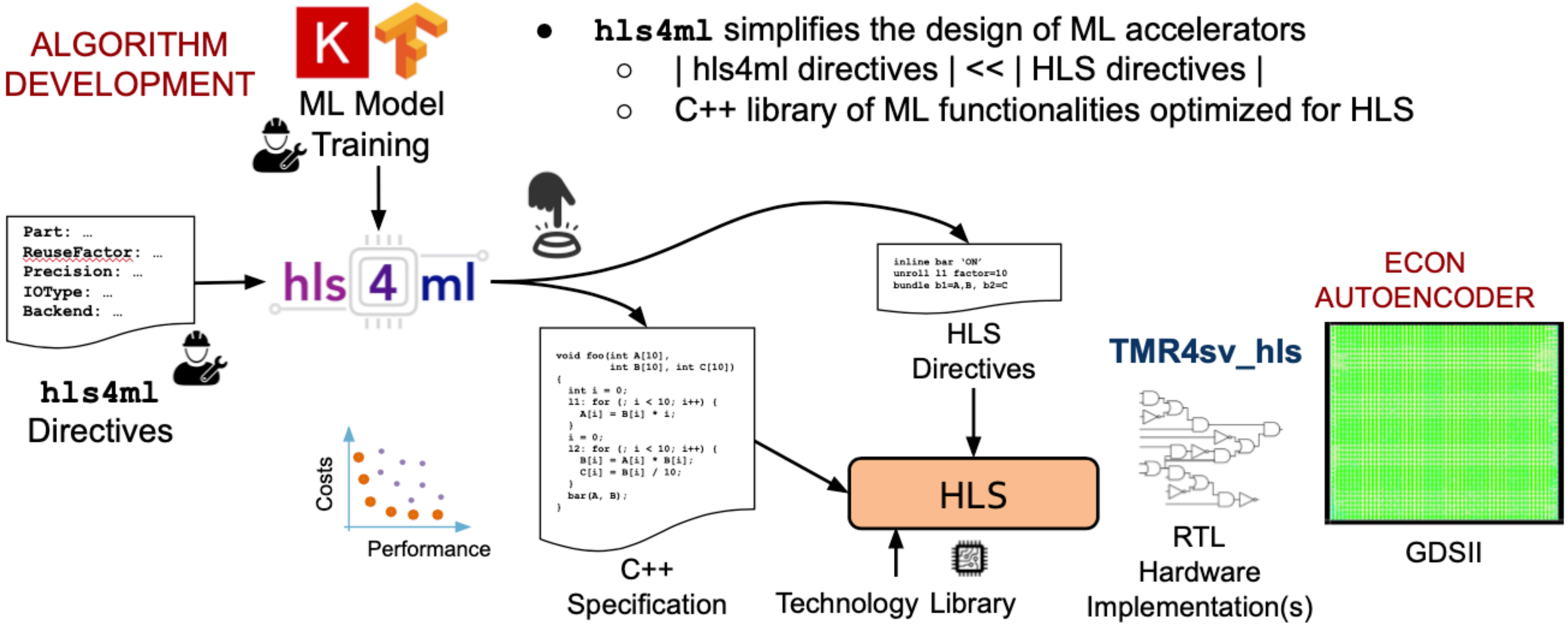
- **QKeras used for quantization-aware training**
  - Weights at 6b, but accumulations padded with 3b to be sure no saturation
  - More lower-precision outputs is better
    - for both high- and low-bandwidth scenarios, for full range of module occupancy
  - Adding weights to I2C ~doubles the area, but important for reconfigurability
- **Chip Fabricated! Functionality and SEE tests complete, look out for papers/talks!**

Metric	Simulation	Target
Power	48 mW	<100 mW
Energy / inference	1.2 nJ	N/A
Area	2.88 mm^2	<4 mm^2
Gates	780k	N/A
Latency	50 ns	<100 ns



hls4ml

- **`hls4ml`** simplifies the design of ML accelerators
  - | hls4ml directives | << | HLS directives |
  - C++ library of ML functionalities optimized for HLS



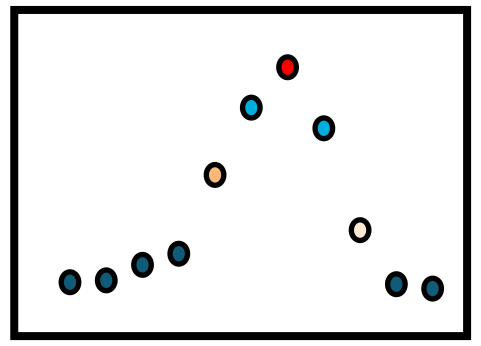
https://github.com/fastmachinelearning/hls4ml
https://github.com/fastmachinelearning/hls4ml-tutorial

Waveform Processing Using Neural Networks on Front End Electronics

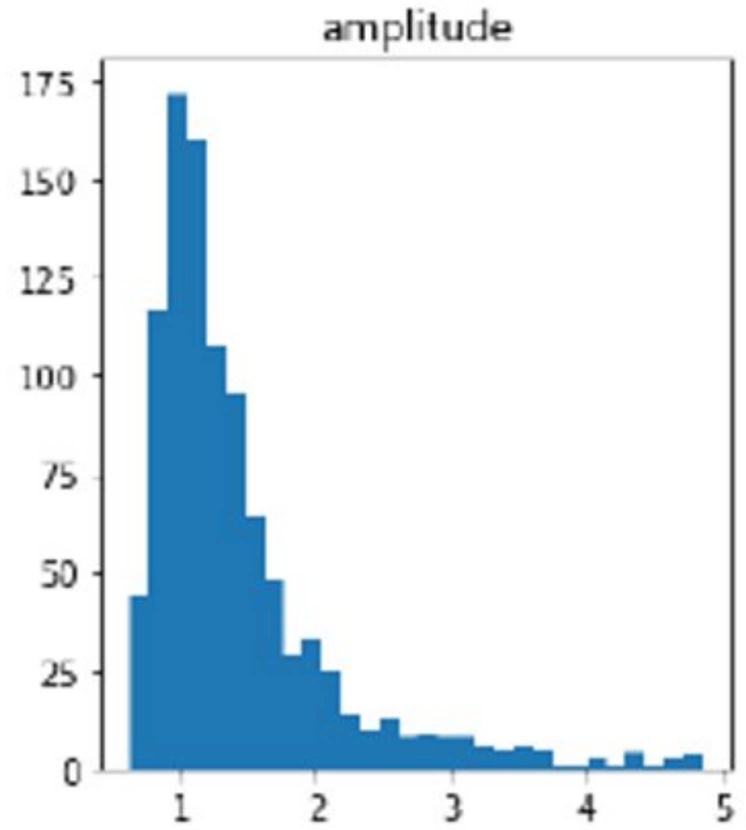
Neural network design methodology

Sensor Response → FEE(Amplifier + Shaper + ADC) → Sampled Waveform → Neural Network

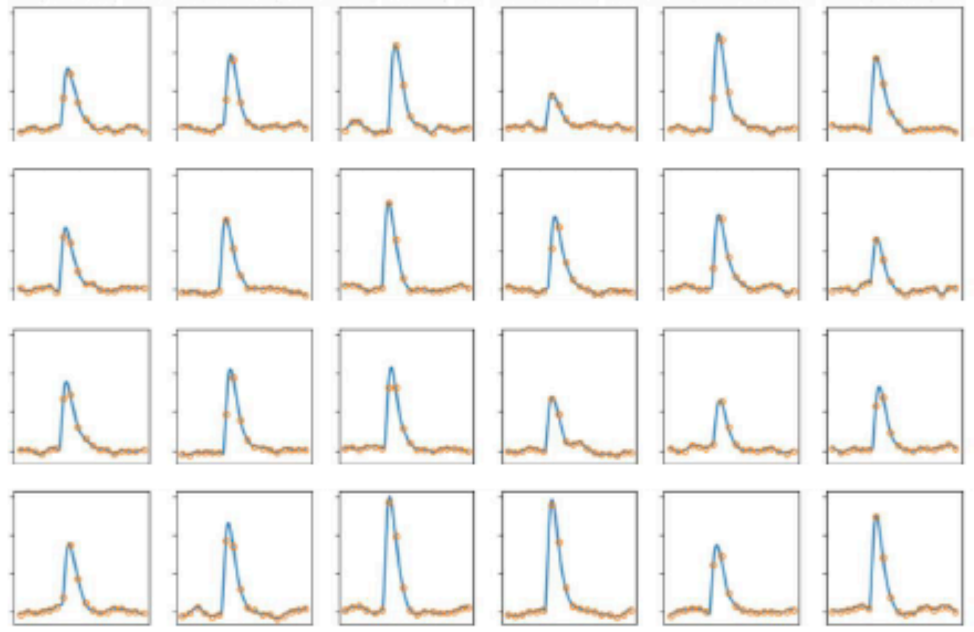
Modelled in Mathematica framework



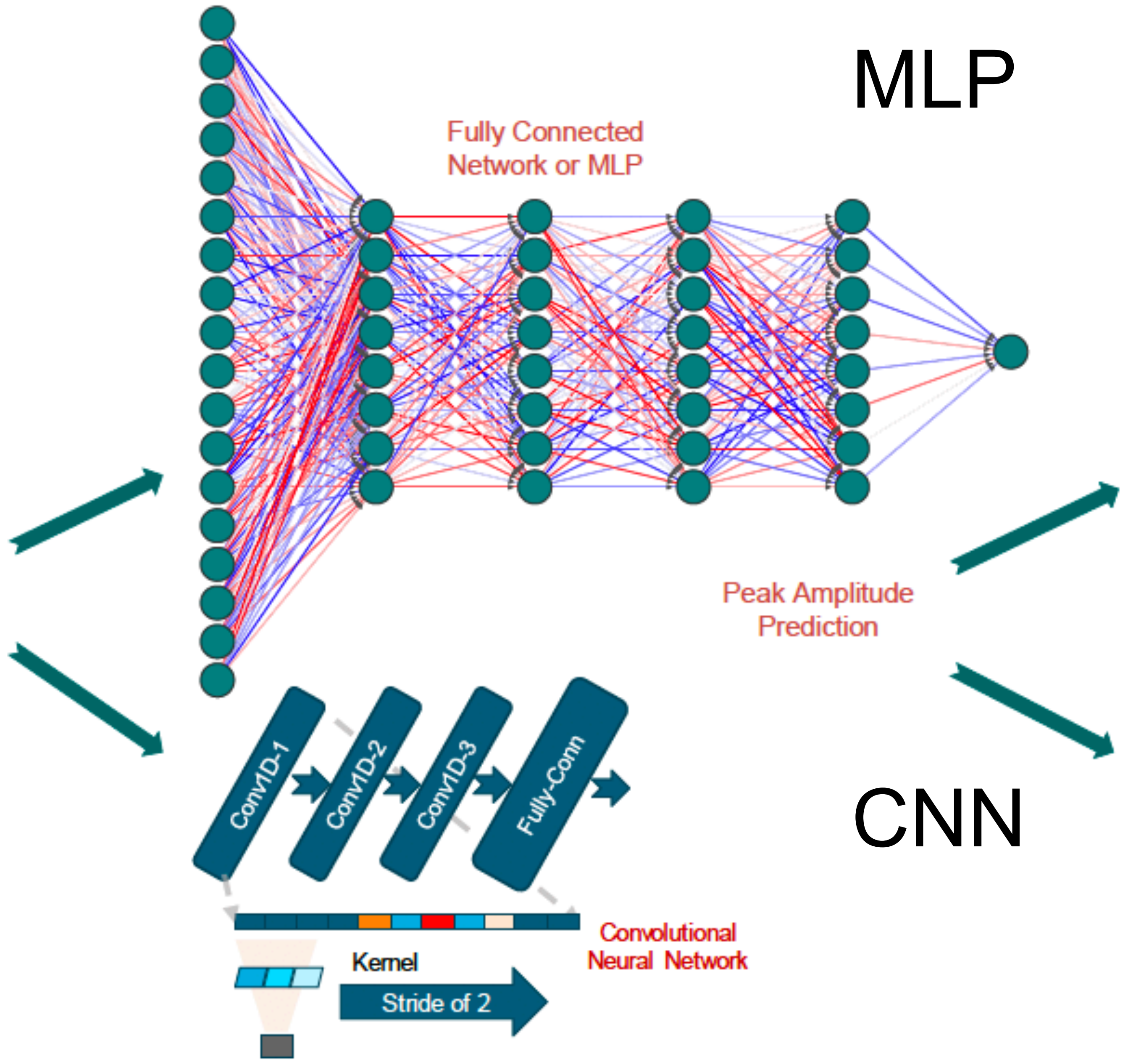
Sampled Waveform

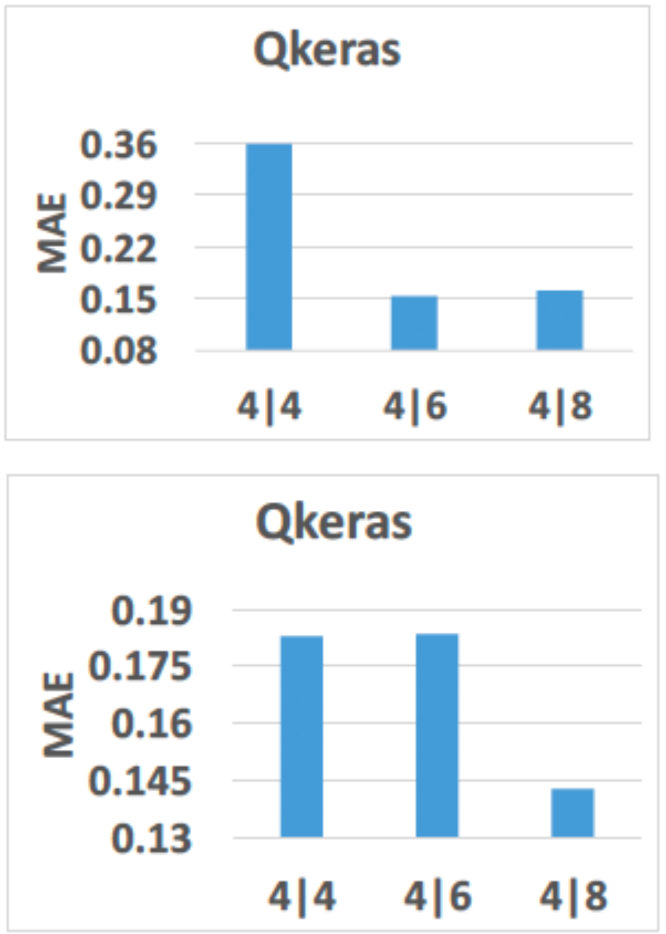


- ❑ Estimating peak amplitude for energy deposited on sensor



Sampled Waveforms



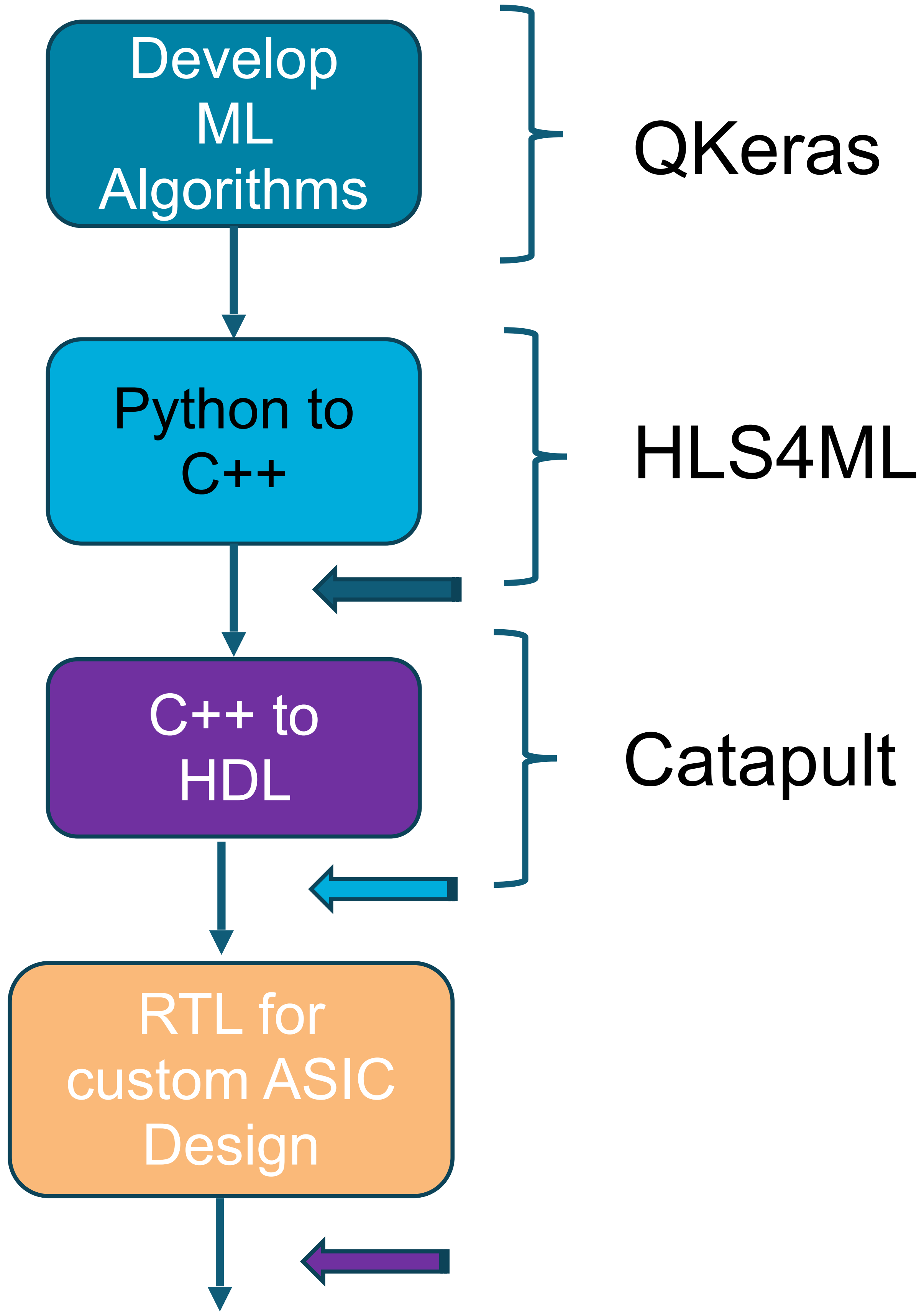


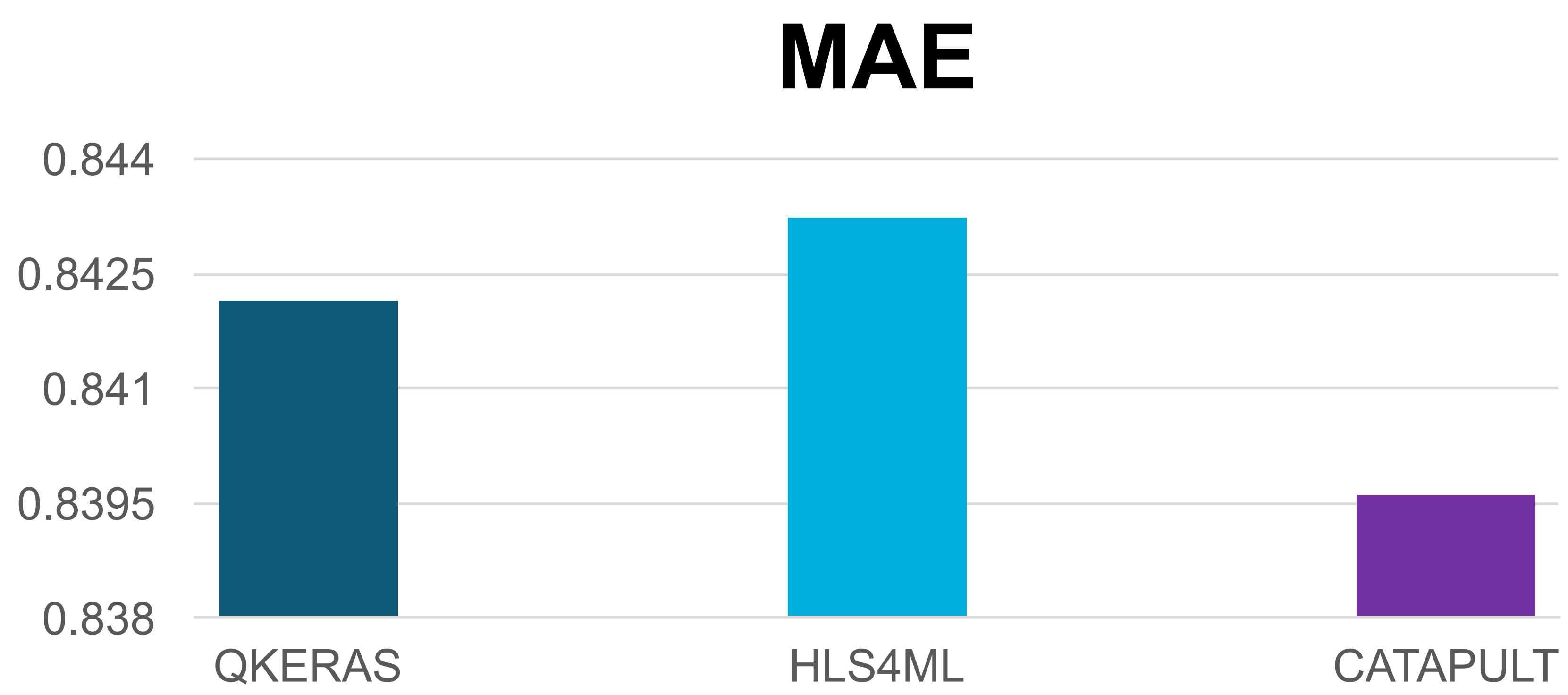
- ❖ Optimized number of layers and neurons on hidden layers
- ❖ Investigated effect of weight quantization on inferencing accuracy
- ❖ Preliminary results are encouraging with acceptable inferencing accuracy

S. Miryala *et al* 2022 *JINST* **17** C01039

Snowmass 2021

Neural Network to ASIC Design





- ❖ Mean Absolute Error is calculated at each stage
- ❖ Good match (< 1%), no loss of accuracy

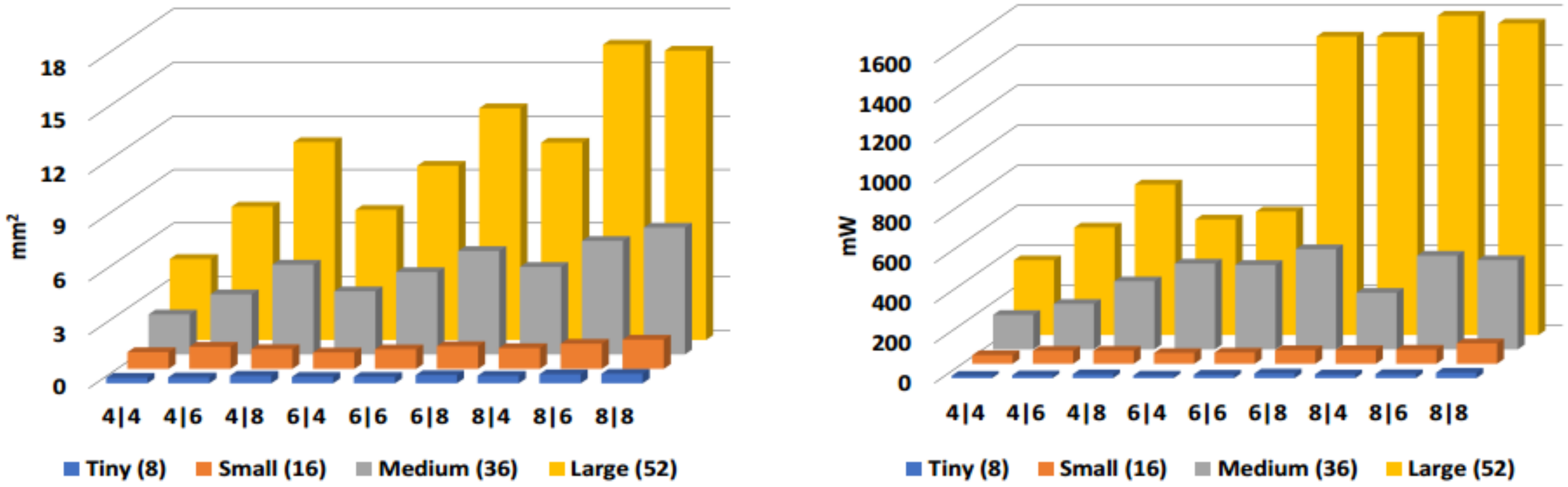


Fig. : Area and power comparison for different precision configuration

- ❑ Neural networks are synthesized in a commercial 65nm process
- ❑ Bigger networks → more area, increased power consumption
- ❑ The networks has a latency of 3-5 clock cycles and throughput of 1 clock cycle

S. Miryala *et al.*, "Peak Prediction Using Multi Layer Perceptron (MLP) for Edge Computing ASICs Targeting Scientific Applications," *2022 23rd International Symposium on Quality Electronic Design (ISQED)*, 2022

5. Applications, design, technology

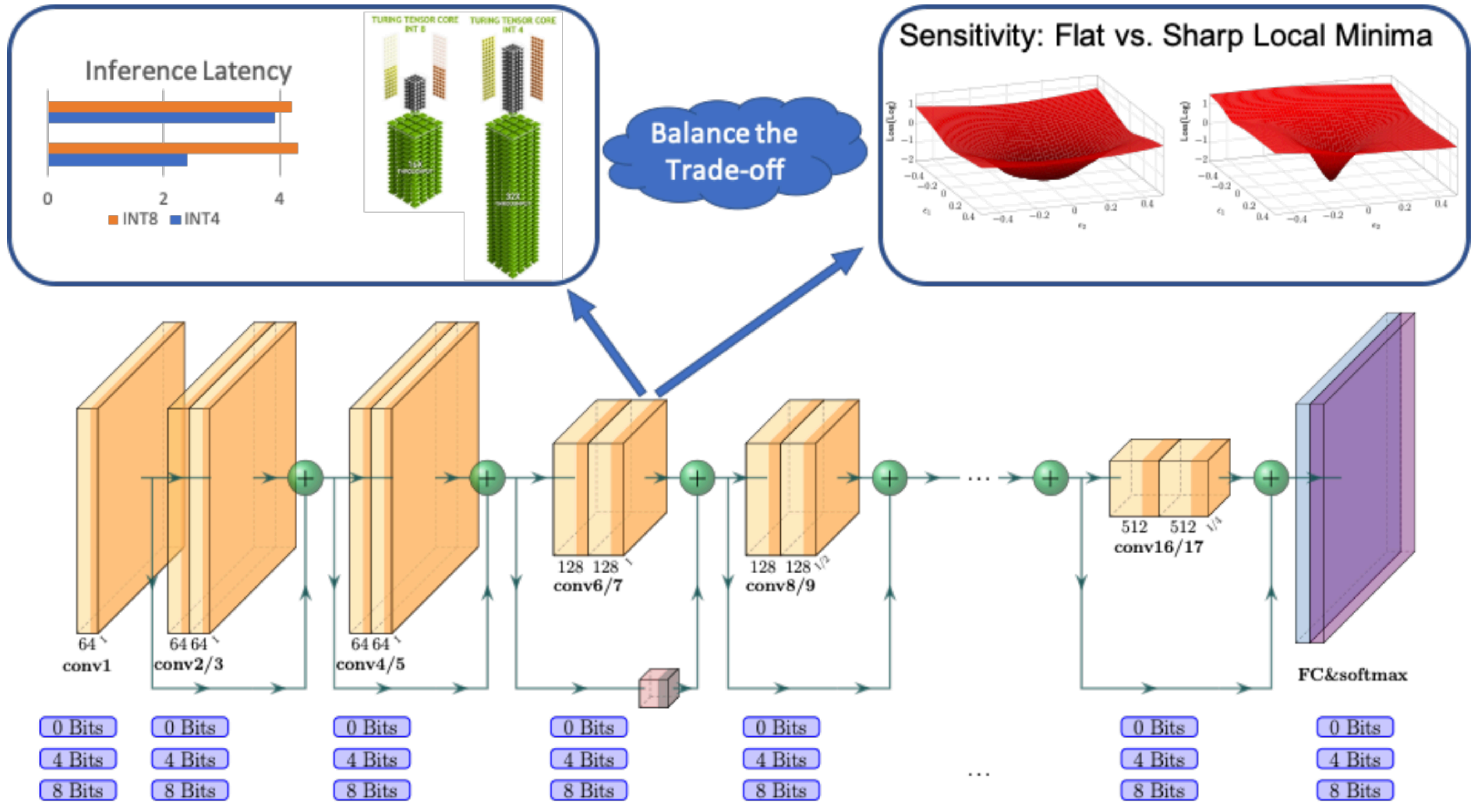
Snowmass 2021

- System-level use-cases
  - **Sensor-integrated AI**
    - Readout electronics integrated directly with sensor (e.g. bump-bonded, TSVs, etc.)
    - Typically for ADC, but AI could be before or after analog-to-digital
  - **On-detector data compression/concentration**
    - Digitized data needs to be further compressed or aggregated to satisfy data transmission constraints

Efficient ML

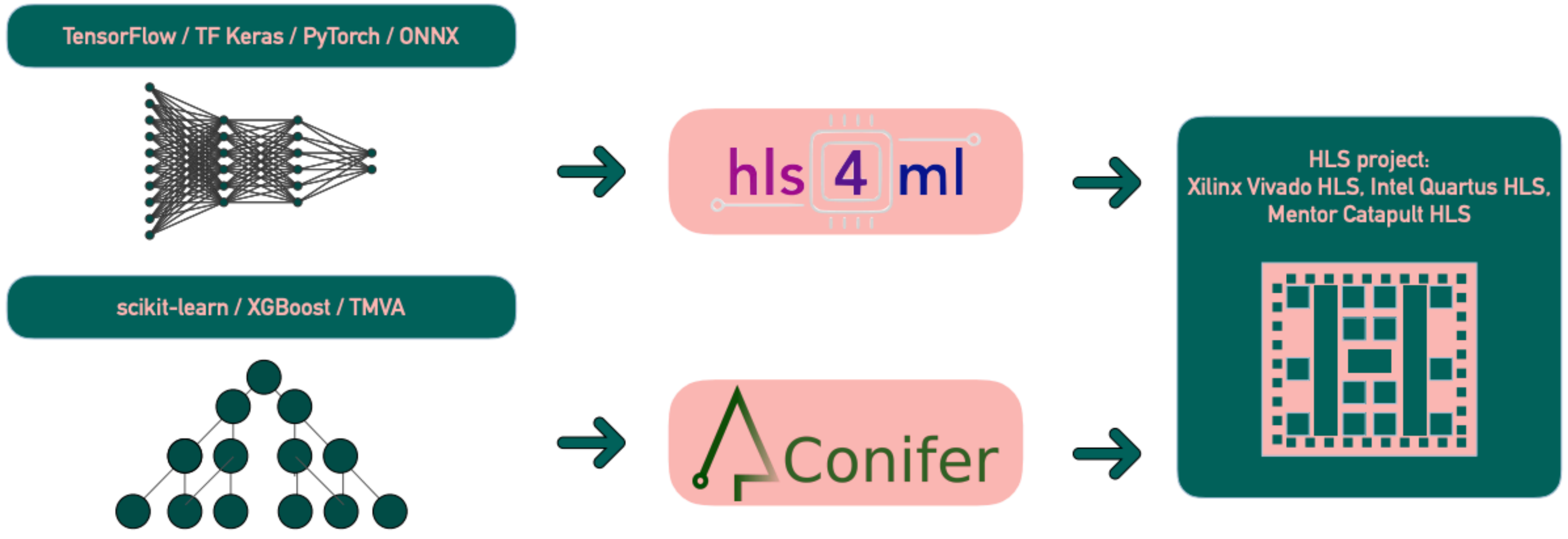
- **A discussion of strategies for improving ML efficiency to enable lower latency**
  - Designing new efficient ML architectures
  - NN & hardware co-design
  - Quantization
  - Pruning and sparse inference
  - Knowledge distillation

- Other important ML topics for front-ends
  - **Fault-tolerant, reliable ML**
  - **Domain adaptation & transfer learning**
    - Reconfigurable architectures



Codesign and validation tools

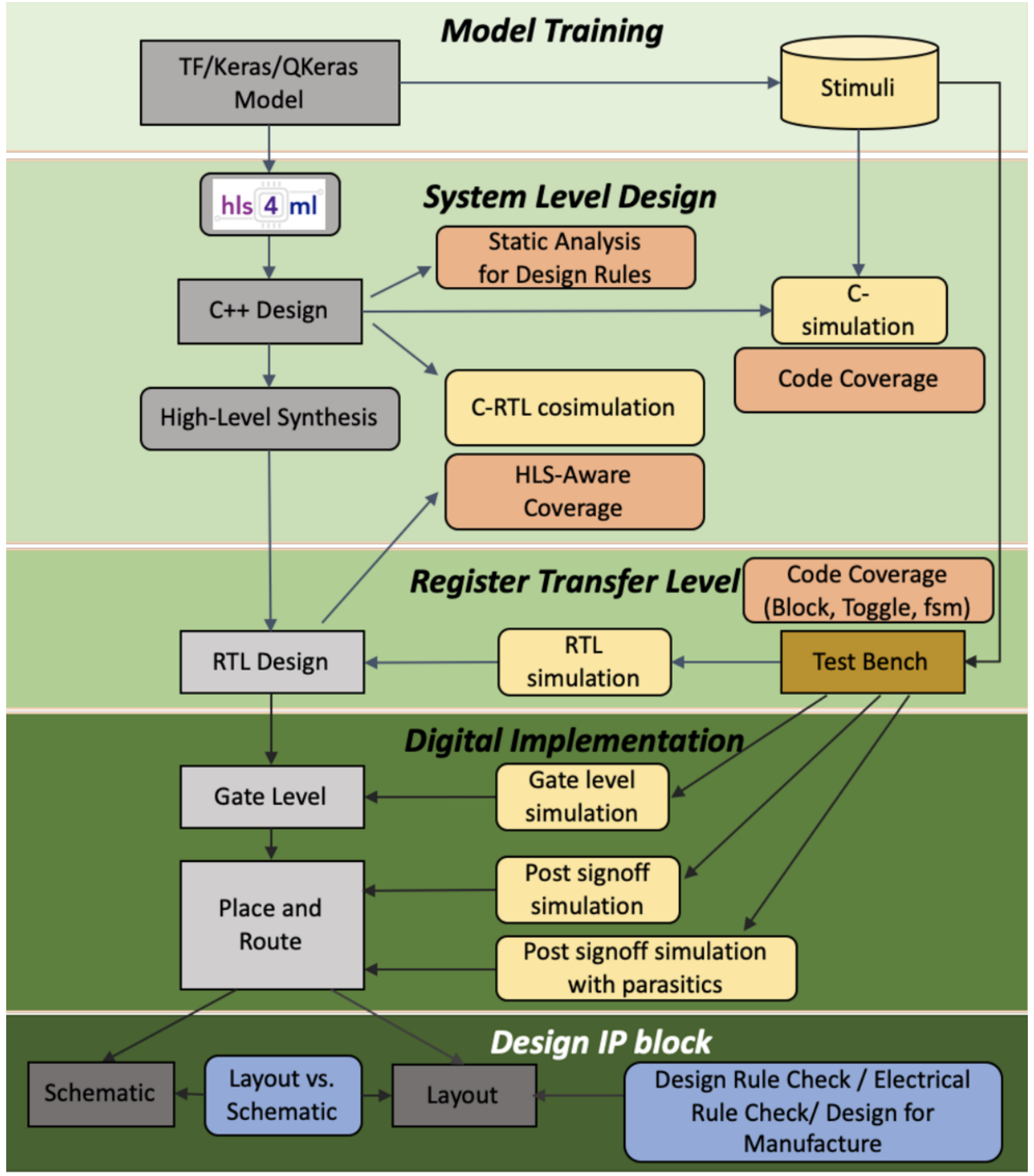
https://arxiv.org/abs/2207.07958



A Siemens Business

Multi-Level IR Compiler Framework

Circuit IR Compilers and Tools



Emerging technologies

- Advanced technology nodes
  - 28nm → 22nm FDSOI/FDX → sub-10nm
- Promising beyond-CMOS 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

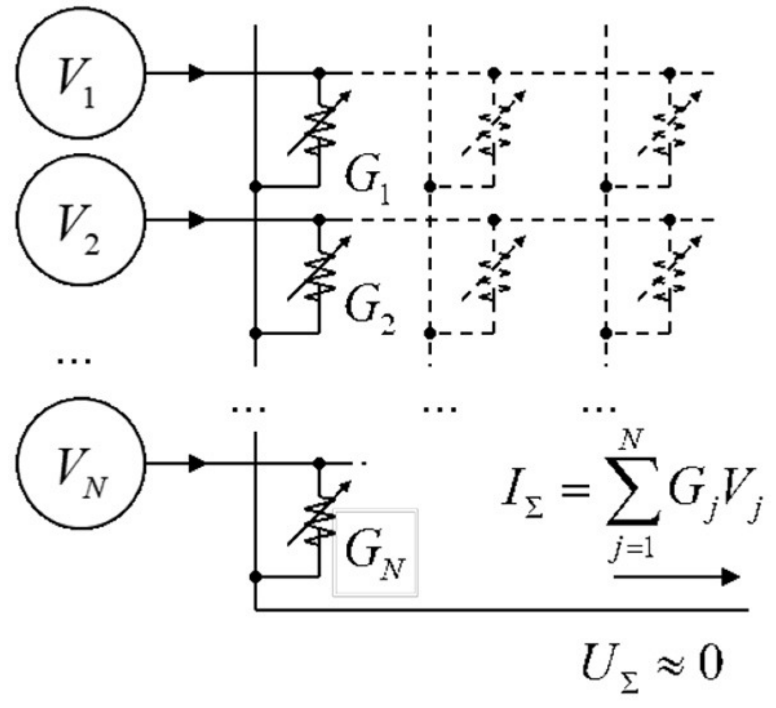


FIGURE 10 | Analog vector-by-matrix multiplication (VMM) in a crossbar circuit with adjustable crosspoint devices. For clarity, the output signal is shown for just one column of the array, while sense amplifier circuitry is not shown. Note that other VMM designs, e.g., utilizing duration of applied voltage pulses, rather than their amplitudes, for encoding inputs/outputs, are now being actively explored see, e.g., their brief review in Bavandpour et al. (2018).

Parting thoughts

Promote interdisciplinary collaborations
physicists, computer scientists, electrical and computer engineers, software engineers, and **industry**

Build open-source, multi-technology codesign workflows

Novel ML research concepts: efficient, fault-tolerant, reliable, domain adaptation

Explore novel microelectronics technologies

Open data, task-based, and data-based benchmarks

Support ecosystem integration and operation

***Strong connections with IF04, CompF3, CompF4 can help amplify the messages within Snowmass*

Extra