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
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.

Contents

1 Executive Summary 3
2 Science Drivers 4
3 Community Needs 5
4 Existing Work 7
5 Applications, Design, and Technology 11
   5.1 System-level use-cases ................................. 11
   5.2 Efficient machine learning training and implementation 12
   5.3 AI co-design tools and methodology ...................... 13
   5.4 Emerging microelectronics technologies .................... 14
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!**
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

- This is not a new idea :)
3. Community needs

- 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
The task:

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

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!**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Simulation</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>48 mW</td>
<td>&lt;100 mW</td>
</tr>
<tr>
<td>Energy / inference</td>
<td>12 nJ</td>
<td>N/A</td>
</tr>
<tr>
<td>Area</td>
<td>2.88 mm²</td>
<td>&lt;4 mm²</td>
</tr>
<tr>
<td>Gates</td>
<td>780k</td>
<td>N/A</td>
</tr>
<tr>
<td>Latency</td>
<td>50 ns</td>
<td>&lt;100 ns</td>
</tr>
</tbody>
</table>
**hls4ml**

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

[GitHub](https://github.com/fastmachinelearning/hls4ml)
[GitHub Tutorial](https://github.com/fastmachinelearning/hls4ml-tutorial)
Waveform Processing Using Neural Networks on Front End Electronics

Neural network design methodology

- 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
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

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

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

• 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
Advanced, sub-10-nm CMOS processes and holistic optimization of circuits, architectures, and algorithms. It includes, for example, taking advantage of aggressive voltage supply scaling [77], very deep pipelines and extensive data reuse in architectures [78], and lowering the precision of weights and activations of the algorithms [79]. As a result, very compact state-of-the-art neural networks, such as MobileNet based on 3.4M parameters and 300M multiply-and-add operations per inference [80], can now be fitted entirely on a single chip. However, on all these fronts, advances are saturating and cannot rely on the faltering Moore's law. Advanced geometry nodes such as 28 nm and below are currently being investigated for Phase III upgrades of HL LHC. A community-driven e-port for modeling radiation effects led by INFN and CERN is currently underway. Similarly, for the fully depleted 22 nm FDSOI process, Fermilab is developing cryogenic models at 4K with EPFL and Synopsys. The back gate control available in 22 FDX allows digital operation at ultra-low supply voltages of 400 mV and below.

Beyond CMOS

For a more detailed review of beyond CMOS technologies, please see Ref. [29] and references therein written by Dmitri Strukov – what follows here is a very reduced summary.

- TensorFlow / TF Keras / PyTorch / ONNX
- scikit-learn / XGBoost / TMVA
- hls 4 ml
- Conifer
- Mentor
- Cadence
- OpenROAD
- Google + SKywater
- MLIR
- CIRCT

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
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