Hardware Acceleration of Complex Machine Learning Models through Modern High-Level Synthesis

Serena Curzel [1, 2], Antonino Tumeo [2], Fabrizio Ferrandi [1]
[1] Politecnico di Milano, [2] Pacific Northwest National Laboratory

CHALLENGE

- ❑ Complex scientific experiments generate large amounts of data, machine learning (ML) methods are often used to process them near the instruments (e.g., electron microscopes, particle detectors, environmental sensors).
- ❑ Heterogeneous systems containing specialized FPGA/ASIC accelerators are needed (GPUs are good for training, but they consume too much power and cannot meet strict latency requirements of inference).

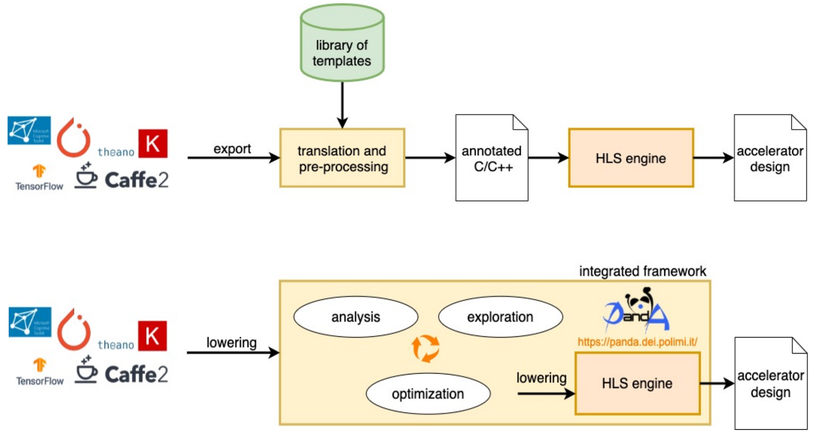
Need for specialization to achieve efficiency

Need for flexibility to keep up with fast evolution

Fast and optimized algorithm-to-chip design flow

APPROACH

- ❑ High-Level Synthesis (HLS) is an established method to simplify and speed up the design of hardware accelerators through automatic translation of high-level (C/C++) code.
- ❑ Existing HLS-based design flows for ML (e.g., FINN [1], hls4ml [3]) have limited flexibility.
- ❑ We propose SODA [5], an open-source, compiler-based framework that supports multiple FPGA/ASIC targets and can easily adapt to new types of algorithms.



"CLASSIC" HLS FLOW [1, 3]

- ✓ Highly optimized for specific applications;
- ✗ Limited support for new types of models (usually focused on Multi-Layer Perceptrons and Convolutional Neural Networks);
- ✗ Choice of target board is limited by the selected HLS tool.

"MODERN" HLS FLOW [5]

- ✓ Embraces multi-level MLIR infrastructure [4] to enable both high-level algorithmic optimizations and low-level hardware-oriented ones;
- ✓ Uses PandA - Bambu [2] as HLS tool, supporting multiple FPGA and ASIC targets with no changes in the input code;
- ✓ Future research will allow to tune the design process to the specific needs of sparse and graph-based algorithms.

FPGA results (Xilinx Zynq-7000)

	Clock (MHz)	Registers	LUTs	Latency (s)
LeNet	146.75	44171	44325	0.689
ResNet-50	217.82	20861	20522	284.640
GCN 1 (dense input)	204.80	14812	8965	1302.990
GCN 2 (sparse input)	210.00	14639	8685	6.414

ASIC results (45 nm)

	Clock (MHz)	Size (mm^2)	Cells	Latency (s)
LeNet	200	0.262	187415	0.362
ResNet-50	200	0.305	220156	212.190
GCN 1 (dense input)	200	0.062	44864	572.738
GCN 2 (sparse input)	200	0.062	45028	4.138

CONCLUSION

- ❑ SODA provides a novel approach that combines MLIR and High-Level Synthesis to bridge the gap between ML algorithmic frameworks and hardware design.
- ❑ Preliminary results (no optimizations included) show that SODA can tackle small and large CNNs, and simple Graph Convolutional Networks for dense and sparse inputs.
- ❑ The design flow is available to any high-level framework with a lowering to basic MLIR dialects, allowing to accelerate parts of complex scientific experiments beyond ML algorithms.

References
[1] Michaela Blott et al. 2018. FINN-R: An end-to-end deep-learning framework for fast exploration of quantized neural networks. ACM Transactions on Reconfigurable Technology and Systems (TRETS).
[2] Politecnico di Milano. Bambu: A Free Framework for the High-Level Synthesis of Complex Applications. https://panda.dei.polimi.it/?page_id=31
[3] Javier Duarte et al. 2018. Fast inference of deep neural networks in FPGAs for particle physics. Journal of Instrumentation.
[4] MLIR project. Multi-Level IR Compiler Framework. https://mlir.llvm.org/
[5] Jeff Zhang et al.2021. Towards Automatic and Agile AI/ML Accelerator Design with End-to-End Synthesis. In IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP).

U.S. DEPARTMENT OF ENERGY | BATTELLE
PNNL is operated by Battelle for the U.S. Department of Energy

For additional information, contact:
Serena Curzel | serena.curzel@pnnl.gov

10/6/21 | PNNL-SA-165702

POLITECNICO MILANO 1863

Pacific Northwest NATIONAL LABORATORY
www.pnnl.gov