Exploiting FPGAs and Spiking Neural Networks at the micro-edge: the EdgeAI approach *

Paolo Meloni, Paola Busia, Gianluca Leone, Luca Martis, and Matteo A. Scrugli

Università degli Studi di Cagliari, Cagliari, Sardinia, Italy name.surname@unica.it

Abstract. This paper outlines the initial FPGA-centric endeavors within the EdgeAI project, targeting scenarios where extremely constrained power-energy parameters intersect with the demand for high performance and accuracy in executing Artificial Intelligence (AI) algorithms. Our discussion, after presenting the generalities of the EdgeAI project, revolves around the project objective of leveraging simultaneously event-based spiking neural networks and low-end FPGA chips for very-low-power near-sensor AI inference. We present the hardware/software implementation of this approach and the early results on the project use cases.

Keywords: FPGAs · Spiking neural networks · edge computing.

1 At-the-edge AI and the EdgeAI project

At-the-edge Artificial Intelligence (AI) empowers Machine Learning (ML) and Deep Learning (DL) at the network's periphery, closer to sensors and actuators, for localized data collection and processing, reducing latency, enhancing data privacy and security, and diminishing the need for cloud connectivity.

However, this poses challenges related to the execution of a complex computing workload on resource-constrained platforms. Thus, it requires dealing with diverse technologies and optimizing energy usage.

The EdgeAI project [1], a joint effort of 42 partners, part of the Key Digital Technologies (KDT) Joint Undertaking (JU), aims to face such challenges, to play a pivotal role in Europe's digital evolution towards smarter processing solutions at the edge. It focuses on creating fresh electronic parts and systems, refining processing setups, improving connectivity, and developing software, algorithms, and middle-layer technologies.

1.1 EdgeAI applications

The main aim of EdgeAI is to advance solutions across various layers of AI technology, culminating in the creation of real-time performing multimodal edge AI implementations for diverse industrial sectors.

* EdgeAI -*Edge AI Technologies for Optimised Performance Embedded Processing*- project is funded by Key Digital Technologies Joint Undertaking (KDT JU) - grant agreement No 101097300.

DOI: https://doi.org/10.1007/978-3-031-55673-9_21

The EdgeAI project partners work to demonstrate the applicability of the developed approaches in 20 demonstrators across five industrial value chains:

- digital industry,
- energy,
- agri-food and beverage,
- mobility, and
- digital society,

considering performance, security, trust, and energy efficiency demands inherently in each of these demonstrators. EdgeAI is designed to provide benefits across industrial sectors, to significantly contribute to the ubiquitous adoption of at-the-edge AI in society.

1.2 Reconfigurable computing in EdgeAI

The EdgeAI vision spans across the the whole computing continuum [2], as represented in Figure 1, comprising:

- the micro-edge (processing units in embedded microcontrollers, sensors, and actuators, etc.)
- the deep-edge (processing units providing extended processing power, in gateways, mobile phones, programmable logic controllers, etc.)
- the meta-edge (on-premises high-performance edge processing microservers combining different microcontrollers and processors for specific operations)

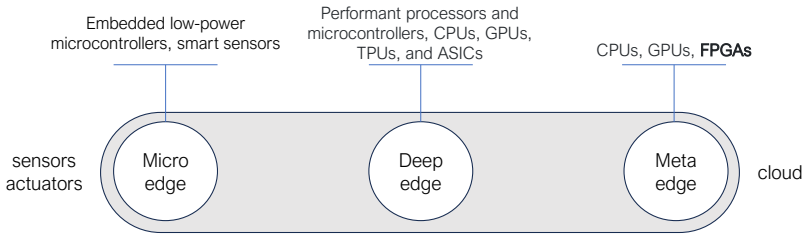

Fig. 1: EdgeAI vision of the computing continuum.

To this aim, different research tasks focus on a wide scope of heterogeneous processing platforms, aiming to amalgamate different computing solutions such as central processing units (CPUs), graphics processing units (GPUs), tensor processing units (TPUs), application-specific integrated circuits (ASICs), neuromorphic processing units (NPUs), system-on-chip (SoC) [2].

Moreover, since adaptability is one of the key requirements, and the possibility of reconfiguring the processing nodes and easily adapting them to different use cases and usage modes is of foremost importance, field-programmable gate arrays (FPGAs) are a main key enabling technology in the project and in the AI domain in general. Thus, numerous research efforts are spent by the community to optimize their use for the inference of AI algorithms in an edge computing

protocol. To date, however, most approaches are focused on exploiting the parallelism of resources on FPGA fabric, for processing image streams with impressive data rates, tolerating power consumption figures compliant with the deep- or meta-edge, while how to use FPGAs for processing sensor data at the micro-edge is still an open question. In this area, in the initial phase of the project, two main research lines have emerged. On one hand, EdgeAI researchers have explored and confirmed the use of FPGAs at the meta-edge, focusing on tool-enabled workload reduction for Convolutional Neural Networks (CNNs), implementing techniques for hybrid quantization that have enabled mid-end FPGAs to outperform alternatives in energy efficiency [2].

On the other hand, we have studied the usage of FPGAs at the micro-edge, for the inference of more power-efficient event-based algorithms such as Spiking Neural Networks (SNNs), combining light-weight topologies with low-power reconfigurable devices, to enable inference in an envelope of few milliwatts.

In the rest of the paper, we focus on this latter effort, providing an overview of the hardware/software developed IPs and a glimpse on achieved results.

2 Background

SNNs emerge as a promising solution with energy-efficient, event-driven processing. However, exploiting the benefits of event-driven processing often requires specialized computational architectures.

FPGAs, due to their highly customizable hardware design, present themselves as suitable candidates for these computational tasks. Their design allows for the exploitation of sparse neuron firing patterns. The core Digital Signal Processor (DSP) slices are engineered to adeptly handle a suite of arithmetic operations including addition, multiplication, and multiply-and-accumulate. On another front, BRAM (Block Random Access Memory) units, due to their adaptable design and size, are especially suitable for integrating SNN models and facilitating data access and management. Although most effective pure-SNN processors are neuromorphic ASICs like Truenorth, Spinnaker, and Loihi, in related literature, some efforts concentrate on leveraging FPGAs to enhance flexibility and integration [3]. Some aim to simulate bio-realistic neural tissue [4], while others focus on processing event-encoded images via SNNs, relying on mid-to-high-end FPGA devices [5] [6]. Other studies [7][8] target smaller devices, emphasizing spiking convolution layers for two-dimensional inputs, like images from DVS cameras. However, these implementations might not be suitable for low-power smart sensors due to their higher power/energy/cost requirements.

Within EdgeAI, we have explored SNN execution on low-power FPGAs, to demonstrate the practicality of such an approach within feasible power budgets for battery-operated and cost-effective sensor nodes.

3 Light-weight SSNs for the micro-edge

We have created an architectural template, that we named `SYNtzulu`, and tested it across various near-sensor data processing scenarios. Figure 2 illustrates its

system-level architecture. It is implemented on a Lattice iCE40UP5k FPGA and comprehends as main processing elements:

- a RISC-V core handling input/output data flow and configuring `SYNtzulu`;
- an 8-way SIMD SNN engine executing the inference of feed-forward SNNs composed of dense layers of LIF neurons.

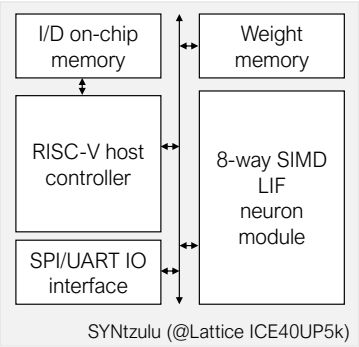

Fig. 2: Architecture overview of the low-power SNN engine developed within EdgeAI

The SNN engine comprises a configurable encoding slot preprocessing the sensor data to obtain spike trains, and a dual decoding slot translating output spike trains into classification outcomes. Moreover, it is tailored for executing SNN inference exploiting spike sparsity, and fostering low power consumption.

3.1 Capabilities and Constraints in Processing

The inference time of a specific SNN in `SYNtzulu` is directly influenced by the number of synapses and the time needed for computing the synaptic current. All other operations happen concurrently and overlap during execution. Since the SNN consumes 8 synapses per cycle, the inference throughput is thus given by the number of active synapses divided by 8, multiplied by the clock frequency, settable to up to 45 MHz.

The active synapse count is determined by the SNN topology and by the sparsity, since, thanks to a dynamic *spike stack* inside the SNN engine, inactive upstream neurons can be skipped during the computation of the LIF dynamics.

The primary limitation of `SYNtzulu` is the size of the weight memory, imposing a cap on the number of synapses in the executing SNN. While the memory composition can vary using on-FPGA storage macros, the default ICE40UP5k configuration can store 131,072 8-bit weights.

We have demonstrated that lightweight SNNs fitting within this limit can achieve top-tier accuracy in processing online sensor data, more specifically in ECG, EEG, and EMG [9, 10, 11]. Some example results are reported in Figure 3 and Figure 4.

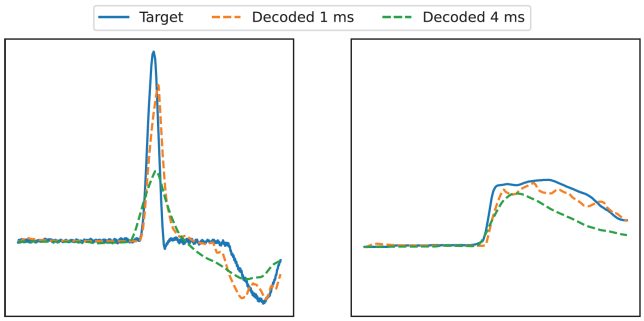

Fig. 3: View of example results on the EEG decoding dataset. The Figure shows two differently timed decoding experiments of the velocity of a macaque hand on the basis of the recorded intracranial EEG.

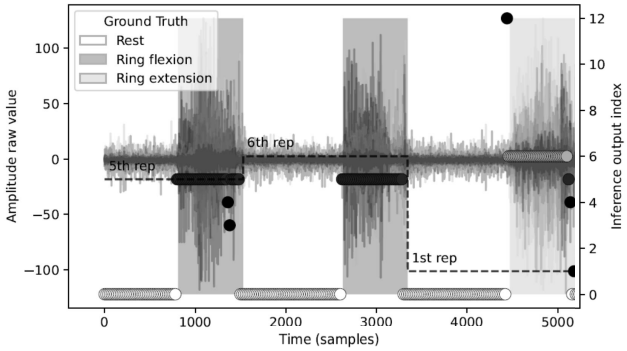

Fig. 4: Example result on the EMG classification datasets. Identification of rest and three gestures (two ring flexions and one ring extension from EMG)

Power consumption measured on the breakout board corresponds to around 12 mW during inference and around 1.2 mW during idle, when the system is acquiring input data and no events need to be processed. Considering the sparsity in the datasets mentioned above and the corresponding sampling times, average power consumption goes down to around 1.4 mW. Summarizing, in this use case, `SYNtzulu` consumes just 0.5 mW/MHz and shows energy efficiency corresponding to 136 pJ per synapse[1]. These measurements highlight that, when the task's nature (and consequently, the SNN algorithm) permits a lightweight approach, `SYNtzulu` can be utilized to develop highly efficient near-sensor processing nodes, resulting in substantial energy savings.

4 Conclusion and future work

The EdgeAI project is entering the second year of activities. FPGA-related activities will continue to improve accuracy and power efficiency in the project use

[1] Based on worst-case power usage and minimal performance due to worst-case pipeline imbalance

case, focusing on improved training and quantization strategies, as well as on testing novel architectural template features and alternative FPGA devices for implementation.

5 Disclosure of interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] *EdgeAI website*. https://edge-ai-tech.eu/. Accessed on January 10, 2024.

[2] Ovidiu Vermesan and Dave Marples. *Advancing Edge Artificial Intelligence: Systems Contexts*. River Publishers, 2023.

[3] Murat Isik. "A Survey of Spiking Neural Network Accelerator on FPGA". In: *arXiv preprint arXiv:2307.03910* (2023).

[4] Gianluca Leone, Luigi Raffo, and Paolo Meloni. "A Bandwidth-Efficient Emulator of Biologically-Relevant Spiking Neural Networks on FPGA". In: *IEEE Access* 10 (2022), pp. 76780–76793.

[5] Sixu Li et al. "A Fast and Energy-Efficient SNN Processor With Adaptive Clock/Event-Driven Computation Scheme and Online Learning". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 68.4 (2021), pp. 1543–1552. DOI: 10.1109/TCSI.2021.3052885.

[6] Sathish Panchapakesan, Zhenman Fang, and Jian Li. "SyncNN: Evaluating and accelerating spiking neural networks on FPGAs". In: *ACM Transactions on Reconfigurable Technology and Systems* 15.4 (2022), pp. 1–27.

[7] Hanwen Liu et al. "A Low Power and Low Latency FPGA-Based Spiking Neural Network Accelerator". In: *2023 International Joint Conference on Neural Networks (IJCNN)*. 2023, pp. 1–8. DOI: 10.1109/IJCNN54540.2023.10191153.

[8] Alessio Carpegna, Alessandro Savino, and Stefano Di Carlo. "Spiker: an FPGA-optimized Hardware accelerator for Spiking Neural Networks". In: *2022 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. 2022, pp. 14–19. DOI: 10.1109/ISVLSI54635.2022.00016.

[9] Gianluca Leone et al. "On-FPGA Spiking Neural Networks fornbsp;Multi-Variable End-to-End Neural Decoding". In: *Applied Reconfigurable Computing. Architectures, Tools, and Applications: 19th International Symposium, ARC 2023, Cottbus, Germany, September 27–29, 2023, Proceedings*. Cottbus, Germany: Springer-Verlag, 2023, 185–199. ISBN: 978-3-031-42920-0. DOI: 10.1007/978-3-031-42921-7_13. URL: https://doi.org/10.1007/978-3-031-42921-7_13.

[10] M. A. Scrugli et al. "On-FPGA Spiking Neural Networks for Integrated Near-Sensor ECG Analysis." In: *2024 Design, Automation Test in Europe Conference Exhibition (DATE)*. 2024.

[11] M. A. Scrugli et al. "sEMG-based Gesture Recognition with Spiking Neural Networks on Low-power FPGA." In: *Design and Architecture for Signal and Image Processing*. Springer International Publishing, 2024.