Thesis title:

FPGA-based Neuromorphic Hardware Platform

Institution:

Sorbonne Université
French National Centre for Scientific Research (CNRS)
Computer Science Laboratory of Sorbonne University's Faculty of Science and Engineering (LIP6)
Sorbonne Center for Artificial Intelligence (SCAI)

Location:

Paris, France

When:

Starting September or October 2023

Funding:

3-year PhD grant, ~2000€ monthly gross salary, social security benefits included.

Advisors:

Dr. Haralampos-G. Stratigopoulos, Sorbonne Université, CNRS, LIP6
Dr. Luis A. Camuñas-Mesa, Instituto de Microelectrónica de Sevilla (IMSE-CNM), CSIC and Universidad de Sevilla

Context:

Artificial Intelligence (AI) and Machine Learning (ML) algorithms have been a subject of interest for several decades now. Although AI and ML have gone through hype cycles of disappointment and enthusiasm, recent algorithmic advancements, in particular Deep Neural Networks (DNNs), as well as the availability of big data and the rapid growth of computing power, have renewed interest leading nowadays to applications in numerous distinct fields, i.e., robotics, medicine, autonomous vehicles, computer vision, speech recognition, natural language processing, gaming, etc.

DNN models are computationally intensive and from a hardware perspective this poses severe challenges of data storage, movement, and processing speed on conventional Central Processing Units (CPUs) with a traditional Von Neumann computer architecture. To this end, there are intense and on-going efforts nowadays towards designing dedicated and customized processors for AI [1-2], referred to as AI hardware accelerators, which belong to the larger family of domain-specific computing paradigms. Widely used AI hardware accelerators today are Graphics Processing Unit (GPUs) and Field-Programmable Gate Arrays (FPGAs), but orders of magnitude of energy-speed improvement can be achieved with Application-Specific Integrated Circuits (ASICs).

Another high incentive for designing AI hardware accelerators is to push the execution of AI algorithms from the cloud closer to the sources of data onto edge devices [3]. This is driven by energy, bandwidth, speed, availability, and privacy requirements. More specifically, edge computing reduces the data transfer requirement thus saving energy and bandwidth. Saving bandwidth is important given the forecast that several tens of billions of edge devices will be connected to the internet in the near future. Several applications, e.g., autonomous vehicles, require low-latency real-time computation which is slowed down due to the communication with the cloud. Also, several applications require availability, thus they need to be less dependent on the communication with the cloud. Finally, handling data locally offers privacy as opposed to transmitting sensitive data over the cloud. Edge AI is a challenging objective since edge devices have limited resources and are often battery-operated. Typically, AI hardware accelerators embedded on edge devices perform only inference with the DNN model trained in software and uploaded upfront.

This PhD will focus on the design of a hardware accelerator for Spiking Neural Networks (SNNs) that form the basis of neuromorphic computing. SNNs mimic brain-like functionality and are considered to be the third generation of neural networks bridging the gap between ML and the biological brain in terms of speed and energy consumption [4].

Nowadays, there exists large-scale neuromorphic research platforms, such as Intel's Loihi [5] and IBM's TrueNorth [6], and the SpiNNaker from the University of Manchester. These platforms allow mapping an arbitrary SNN on a fixed hardware layer but are still highly proprietary. For example, Intel upon request may grant access to use Loihi in the cloud for research purposes. On the other hand, in recent years they have been many neuromorphic chips [8] and FPGA implementations [9-10] demonstrated by academic groups.

This PhD will target an end-to-end architecture-to-Hardware Description Language (HDL) framework for SNN implementation. More specifically, the goal will be to develop a tool that allows users to automatically synthesize any arbitrary SNN in HDL (i.e., VHDL) starting from a high-level architectural description, i.e., number of layers, number of feature maps per layer, and neuron hyper-parameters. From a design perspective, the goal is a SNN hardware accelerator with small form factor and high energy efficiency, such that they it can be used in resource-constrained IoT nodes for near-sensor computation and near-sensor intelligence. Additional attributes will be on-line learning and fault-tolerance. The tool will provide a design ready to be flashed onto an FPGA for fast prototyping and experimentation. The design will be fully synthesizable so that the user can also proceed to an ASIC fabrication. Our group participates actively in the open hardware movement, so we are planning to make the framework open-source. The PhD may culminate in an ASIC fabrication using the open-source tool suite Coriolis [11] for chip design.

Short Bibliography:

[1] S. Bavikadi et al., "A survey on machine learning accelerators and evolutionary hardware platforms," IEEE Des. Test, vol. 39, no. 3, pp. 91–116, 2022.
[2] M. Bouvier et al., "Spiking neural networks hardware implementations and challenges: A survey," ACM J. Emerg. Technol. Comput. Syst., vol. 15, no. 2, 2019.
[3] M. Shafique et al., "An overview of next-generation architectures for machine learning: Roadmap, opportunities and challenges in the IoT era," in Proc. Design Autom. Test Europe Conf. (DATE), 2018, pp. 827–832.
[4] K. Roy, A. Jaiswal A, and P. Panda P, "Towards spike-based machine intelligence with neuromorphic computing," Nature, vol. 575, no. 7784, pp. 607-617, 2019.
[5] M. Davies et al., "Loihi: A neuromorphic manycore processor with on-chip learning," IEEE Micro, vol. 38, no. 1, pp. 82–99, 2018.
[6] P. A. Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," Science, vol. 345, no. 6197, pp. 668–673, 2014.
[7] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker Project," Proc. IEEE, vol. 102, no. 5, pp. 652–665, 2014.
[8] A. Basu, L. Deng, C. Frenkel and X. Zhang, "Spiking Neural Network Integrated Circuits: A Review of Trends and Future Directions," in Proc. IEEE Cust. Integr. Circuits Conf. (CICC), 2022.
[9] L. A. Camuñas-Mesa, Y. L. Domínguez-Cordero, A. Linares-Barranco, T. Serrano-Gotarredona, and B. Linares-Barranco, "A configurable event-driven convolutional node with rate saturation mechanism for modular convnet systems implementation," Front. Neurosci., vol. 12, 2018, Article 63.
[10] D. Gerlinghoff, Z. Wang, X. Gu, R. Goh and T. Luo, "E3NE: An End-to-End Framework for Accelerating Spiking Neural Networks With Emerging Neural Encoding on FPGAs," IEEE Transactions on Parallel & Distributed Systems, vol. 33, no. 11, pp. 3207-3219, 2022
[11] http://coriolis.lip6.fr/

Expected skills:

We seek a highly motivated talent with a M.Sc. degree or equivalent in Electrical Engineering or Computer Engineering and with background knowledge on circuit design, computer-aided design tools (e.g. Cadence, Synopsis, Mentor), and technical computing languages (e.g. MATLAB). Knowledge on AI algorithms and applications is a plus.

How to apply:

Send by e-mail a detailed CV to Haralampos-G. Stratigopoulos (e-mail: haralampos.stratigopoulos@lip6.fr).