

PEPR IA - AdaptING project - thesis proposal



Reconfigurable Computing in Memory Banks

Context

Lab-STICC is a research unit historically recognized in the field of information and communication technologies covering a broad scientific spectrum around digital sciences, and in particular with ability to address various disciplinary fields (Information Theory, Waves & Materials, Embedded Electronics and Computing, Data Sciences, Communication and Signal Detection..) following multiple themes/application sectors. The ARCAD Team, mainly located at Université de Bretagne-Sud (UBS), started in 2011 to work in the field of Coarse Grained Reconfigurable Architectures (CGRA) and in the field of edge AI through national and international projects.

The previous work in ARCAD team include:

1. Different AI accelerators targeting Content Adressable Memories (CAM) designed and prototyped on FPGA [1]. Efficient generation of FPGA-based CNN Accelerators from High-Level Descriptions with the help of complete DSE tool in collaboration with CEA [2].
2. CGRAs and associated compilation tool for Ultra-Low Power (ULP) design, transprecision and embedded AI also investigated in this context [3][4][5][8].
3. Dataflow and near-memory computing [6][7].

The AdaptING project proposes a new architectural paradigm called adaptive architecture, which aims to make hardware adaptable to any given AI application and its constraints in terms of accuracy, energy, latency, and reliability. This approach goes beyond the current state-of-the-art hardware architectures and targets the next generation of AI by investigating and designing flexible, efficient, sustainable, and reliable embedded AI on adaptive architectures.

Scientific challenge

Memory is a vital part of any digital system. It is used to store the programs as well as the data which these programs operate on. The memory must therefore be (very) fast, but ideally also (very) large, and inexpensive. The impossible meeting between these three requirements through a unique technology has led to proposing hybrid solutions based on a memory hierarchy, thus relying on various memory technologies. Cache memory, embedded as close as possible to the processor, on the same chip, is a key element of this hierarchy, giving the programmer the illusion of both fast and large memory.

Through the numerous hardware solutions implemented from generation to generation, more and more transistors in a circuit are allocated for the sole purpose of improving memory access. In many cases, more than 80% of the area of a chip is dedicated to caches, memories and memory controllers, interconnects, etc., whose sole purpose is to store/transfer data or control the storage/transfer of data. Memory accesses are more expensive than an arithmetic operation [9]. As a result, the total energy spent for moving data has

reached excessive proportions. In a mobile system, memory aspects alone can consume up to 62% of its energy [10].

AI applications, and especially machine learning ones, further exacerbate the pressure on the memory subsystem by continuously accessing an ever increasing volume of data. Most of the kernels of AI applications present a low arithmetic intensity, with very simple operations, like addition of a bias for instance, or data clipping for a ReLU. However, they require many data movements for simple computations, hurting their energy efficiency. To tackle this issue, reconfigurable devices have been heavily studied from the computation angle, by offering customisable computation capabilities to relevantly balance flexibility and performance. The computing in-memory or computing near-memory techniques have also been proven interesting approaches to avoid incessant data movements.

The idea is to join these two trends and consider the reconfigurability feature from the other main angle: the memory.

Goals

The goal of this thesis is to explore a reconfigurable memory architecture that embeds some lightweight and elementary computing capabilities, like simple additions or comparisons. This kind of approach is expected to save some data movements between the local memory and the main computing components of the chip.

The objectives are: (i) to update on the bibliography related to computing in-memory, reconfigurable devices, embedded AI, (ii) to propose a new hardware/software solution for reconfigurable computing inside the memory banks, (iii) to run experiments on applications from the embedded AI domain.

Keywords

'CGRA', 'Memory', 'Compilation', 'AI', 'Energy Efficiency'

Skills

- hardware architectures, VHDL, Verilog, SystemVerilog
- LLVM
- C/C++
- AI applications

Location

- Research team: ARCAD, Lab-STICC
- Address: Lab-STICC rue Saint-Maudé 56100 Lorient France

Dates

- 2025-2028

Salary

- 2100 € per month before tax
- possibilities to teach

Supervisors

- Kevin Martin (Lab-STICC, ARCAD team, Lorient) - kevin.martin@univ-ubs.fr
- Camille Moniere (Lab-STICC, ARCAD team, Lorient) - camille.moniere@univ-ubs.fr

Application

Send resume and application letter to kevin.martin@univ-ubs.fr

References

- [1] Hugues Wouafo, Cyrille Chavet, Philippe Coussy: Clone-Based Encoded Neural Networks to Design Efficient Associative Memories. *IEEE Trans. Neural Networks Learn. Syst.* 30(10): 3186-3199 (2019)
- [2] Nermine Ali, Jean-Marc Philippe, Benoît Tain, Philippe Coussy: Generating Efficient FPGA-based CNN Accelerators from High-Level Descriptions. *J. Signal Process. Syst.* 94(10): 945-960 (2022)
- [3] Satyajit Das, Kevin J. M. Martin, Thomas Peyret, Philippe Coussy: An Efficient and Flexible Stochastic CGRA Mapping Approach. *ACM Trans. Embed. Comput. Syst.* 22(1): 8:1-8:24 (2023)
- [4] Chilankamol Sunny, Satyajit Das, Kevin J. M. Martin, Philippe Coussy: Energy Efficient Hardware Loop Based Optimization for CGRAs. *J. Signal Process. Syst.* 94(9): 895-912 (2022)
- [5] Prasad, R.; Das, S.; Martin, K.; Tagliavini, G.; Coussy, P.; Benini, L.; Rossi, D. TRANSPIRE: An energy-efficient TRANSprecision floating-point Programmable archItectuRE. *Design, Automation Test in Europe Conference Exhibition (DATE)*, 2020, 1067-1072
- [6] Rodrigo Cataldo, Ramon Fernandes, Kevin J. M. Martin, Jarbas Silveira, Gustavo Sanchez, Johanna Sepúlveda, César A. M. Marcon, Jean-Philippe Diguët: Subutai: Speeding Up Legacy Parallel Applications Through Data Synchronization. *IEEE Trans. Parallel Distributed Syst.* 32(5): 1102-1116 (2021)
- [7] Notifying Memories: a case-study on Data-Flow Applications with NoC Interfaces Implementation. Kevin Martin, Mostafa Rizk, Martha Johanna Sepulveda Florez, Jean-Philippe Diguët. *Design Automation Conference*, Jun 2016, Austin, United States. (10.1145/2897937.2898051)
- [8] Satyajit Das, Kevin J. M. Martin, Davide Rossi, Philippe Coussy, Luca Benini: An Energy-Efficient Integrated Programmable Array Accelerator and Compilation Flow for Near-Sensor Ultralow Power Processing. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* 38(6): 1095-1108 (2019)
- [9] Mark Horowitz. *Computing's energy problem (and what we can do about it)*. In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014.
- [10] Amirali Boroumand, Saugata Ghose, Youngsok Kim, Rachata Ausavarungnirun, Eric Shiu, Rahul Thakur, Daehyun Kim, Aki Kuusela, Allan Knies, Parthasarathy Ranganathan, and Onur Mutlu. *Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks*. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS '18*, 2018. ACM.