Exploration of high speed in-memory computing operators in advanced CMOS technology node

Contact Persons: Prof. David Atienza (david.atienza@epfl.ch)

Dr. Alexandre Levisse (alexandre.levisse@epfl.ch)

Project Description

With the incoming innovations in *Artificial Intelligence* (AI) such as self-driving cars or natural language recognition, *Convolutional Neural Networks* (CNN) have been gaining popularity. Their main asset lays in their natural capability to process data with spatial or temporal interrelationships. However, nowadays CNNs implementations are extremely energy hungry, forbidding their use in embedded systems.

In this context, the ESL laboratory proposed an innovative solution featuring *In-Memory Processing* (IMP) to execute CNNs inside an embedded system. The solution developed in ESL relies on a breakthrough trendy approach named *Bitline Computing* and consists in performing logic operations directly in the cache periphery. It enables fast, low power and low area overhead computation, but opens new questions at all the design levels.
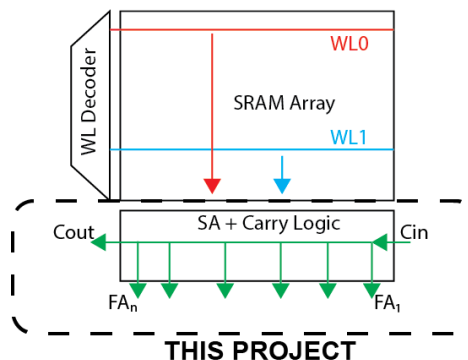

Figure 1: operation of an addition operation performed in an in-memory processing array with highlighted project objectives.

In this project, we propose to focus on the design of the IMP memory under development in ESL. And particularly on the in-memory computation block. This block, developed in ESL enables standard *read*, *write*, *and/nor* and *shift* operations at high frequencies. However, more complex operations such as *addition* are slower due to the static carry ripple used.

This project proposes to explore the opportunities opened by dynamic logic to improve the *addition* operation (cf Figure 1) in order to enhance its speed in nominal supply conditions but also at low voltage. For that purpose, the student will use an advanced CMOS technology PDK

(28nm HPM technology) and industrial tools such as Cadence Virtuoso tool for circuit and layout edition, Synopsys (Calibre, Hspice) and Mentor Graphic (Eldo) tools for electrical simulations and post layout verification.

The project will be carried out at the Embedded Systems Laboratory of EPFL under the supervision of Prof. David Atienza, Dr. Alexandre Levisse and PhD student William Simon.

Project objectives:

1. Understanding of the memory operation and features.
2. Simulation of the currently used compute cache logic
3. Optimization of the cache logic using a Manchester carry adder.
4. Benchmark the solutions in terms of area and performances.
5. If the previous objectives are filled, this work will be part of a submission in a peer-reviewed conference or journal.

Required knowledge and skills:

- Good understanding of MOS transistors behavior
- Advanced knowledge on transistor-level circuit design
- Good analytical skills
- Good background on computer architecture

Appreciated skills:

- Scientific curiosity
- Good communication skills
- Advanced English
- Autonomous work ability

Type of work: 20% theory analysis, 80% hardware design and exploration