System-level simulation and design space exploration of non-volatile neuromorphic architectures

Contact: FranÃ§ois DUHEM DRF//INAC/SPINTEC Francois.DUHEM@cea.fr 04 38 78 52 98

PhD may follow: No

Summary :

Hardware neural network implementation is a hot topic in research and is now considered as strategic for several international companies. Leading projects in neuromorphic engineering have led to powerful brain-inspired chips such as SyNAPSE, TrueNorth and SpiNNaker. Most of these technologies work well in centralized computing farms but will not fit embedded systems or Internet-of-Things (IoT) requirements, due to their energy consumption. Heterogeneous integration between CMOS and emergent technologies is seen as an opportunity to go past this limitation. In particular, Magnetoresistive Random-Access Memory (MRAM) is considered one of the most promising Non-Volatile Memory (NVM) technology expected to mitigate energy consumption when integrated in computing architectures. However, we still miss a high-level perspective on how NVM actually benefits energy efficiency and how it can be improved any further. In the frame of a collaboration between Spintec and the LEAT lab in Sophia Antipolis, a spiking neural network simulator has been developed in SystemC to estimate metrics such as energy consumption, silicon area and performance in various architectures, laying ground for system-level exploration of non-volatile neuromorphic architectures.

In this context, the aim of the internship is to carry on this work and add new features to the simulator. In particular, the intern will have to refine NVM modeling and compare with actual hardware implementations to assess the simulator accuracy. The intern will eventually demonstrate the simulator functionalities by defining the fittest architecture for a vision-based cognitive task, showcasing the benefits of this non-volatile architecture compared to its volatile counterpart.

Full description :

Hardware neural network implementation is a hot topic in research and is now considered as strategic for several companies. Indeed, the recent interest around deep neural networks for pattern recognition has put a new spotlight on neuromorphic engineering and a few industrial giants now dominate the deep learning sector, mainly American (Nvidia, Google, IBM, Intel...). They usually rely on General-Purpose Graphics Processing Units (GPGPUs) for the learning process, and dedicated hardware for inference on embedded targets, which is known to be energy-efficient.

Leading projects in neuromorphic engineering have led to powerful brain-inspired chips able to simulate numerous spiking neurons to investigate a new kind of computer architecture (SyNAPSE, TrueNorth), or to help neuroscientists through international projects such as the Human Brain Project in Europe (SpiNNaker). Most of these technologies work well in centralized computing farms but will not fit embedded systems or Internet-of-Things (IoT) requirements, due to their energy consumption. Heterogeneous integration between CMOS and emergent technologies is seen as an opportunity to go past this limitation.

Non-Volatile Memories (NVMs) have gained traction in the last few years as they are expected to help mitigating the ever growing energy consumption due to leakage in advanced technology nodes. Among emerging NVM technologies, Magnetoresistive Random-Access Memory (MRAM) is considered to be one of the most promising as it reaches performance levels close to those of Static RAM (SRAM) with very high endurance and good downsize scalability.

One promising use of MRAM is non-volatile processors, where non-volatile storage elements are integrated in the memory hierarchy in order to reduce energy consumption. Many studies showed magnetic and hybrid general-purpose processor architectures, with the resulting trade-off between performance, area and energy consumption. These studies do not consider domain-specific accelerators, even though they are usually necessary to achieve higher energy efficiency compatible with embedded systems requirements. In this context, a collaboration between Spintec and the LEAT lab

(Laboratoire d'Electronique, Antennes et TÃ©lÃ©communications) in Sophia Antipolis led to the development of a SystemC simulator that is able to evaluate non-volatile spiking neural network implementations. It infers important metrics such as silicon area, energy consumption and performance while ensuring functional validation. The simulator is expected to lay the ground for wide adoption of such architectures in the growing market of embedded neuromorphic architectures, with potential applications in smart cities, wearables, IoT, mobile robotics, connected and autonomous vehicles, and so on.

The goal of the internship is to continue the development of the simulator and extend it with new features to provide finer-grained comparison between volatile and non-volatile architectures. The intern's missions will be the following:
- Bibliography on NVM and non-volatile architecture modeling
- Refinement of the NVM models
- Development of new features for the simulator
- Assessment of the simulator accuracy
- Demonstration on an application to be defined

Requested skills :
Applicants should have background in embedded systems, system architecture, electronics and programming language such as C/C++ (SystemC appreciated). Knowledge in RTL development is a plus.