This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826655. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, Netherlands, Switzerland

Addressing the call/topic: H2020 ECSEL-2018-2-RIA
Research and Innovation Action

TEMPO

Technologies and hardware for neuromorphic computing

Deliverable

D4.11 – DNN PPA: From Module to Task

Work Package:	WP4 (Design and architecture)
Dissemination level:	Confidential
Official due date:	31.08.2022
Document editor:	Thomas Kämpfe (FhG)
Contributing partners:	FhG, TUD
Internal reviewers:	Björn Debaillie (IMEC), Ilja Ocket (IMEC)
Document version:	V1.0

1. Publishable summary

This document gives an overview of a power-performance-area (PPA) study about the integration of non-volatile memory (NVM) into the SpiNNaker 2 hardware architecture for inference of deep neural networks (DNN). Starting point of the study is the existing architecture and implementation of the SpiNNaker 2 neuromorphic hardware. SpiNNaker 2 is a many-core chip in GF 22nm FDx technology with 153 low-power ARM-based processing elements for the energy-efficient processing of both biologically inspired and deep neural networks. Each processing element has an integrated machine learning accelerator for the efficient processing of convolutional and fully-connected layers. While the existing SpiNNaker2 chip is heavily based on SRAM, NVM technology is becoming more and more mature allowing a streamlined integration into CMOS manufacturing processes in the near future. As NVM, and especially the FeFET (ferro-electric field-effect transistor) technology considered in this study, offers high potential for reducing the power consumption (both for leakage and memory access), it is of great interest to quantitatively study the impact of NVM integration for DNN inference on power, performance and silicon area on the chip-level.

The approach in this study is as follows: we first choose an operating point (clock frequency) of the SpiNNaker 2 processing element that is compatible with the maximum speed of the considered NVM macros. Then, for a given task (processing a convolutional layer from VGG-16), we perform a power simulation and extract the power contribution of the SRAM and other components. Next, we hypothetically replace a varying number of SRAM macros by NVM macros for the storage of weights of the convolutional layer. Due to the lower energy per read of the NVM, this reduces the overall energy for processing the convolutional layer. At the same time, this leads to an increase of silicon area due to the lower memory density of the NVM compared to SRAM. We show and discuss the energy-speed-area trade-off between SRAM-only and SRAM+NVM setups in detail.

The results show only a small energy-saving potential when replacing SRAM by NVM for DNN weight storage in the SpiNNaker2 architecture at the price of significantly increased silicon area. Ultimately, the existing SpiNNaker2 implementation shows the best energy-speed-area trade-off compared to the new considered variants at a lower clock frequency with and without NVM.

The main conclusion to be drawn from this study is that a plain replacement of SRAM by NVM does not necessarily improve the system-level energy-efficiency. For the SpiNNaker2 architecture, which is optimized for low energy-per-operation, a setup with SRAM only provides the optimal operating point. Of course, NVM might play off its strengths for other DNN systems operating at lower speed or requiring a power-off mode.