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

# D4.12 – DNN with NVM: Power management concept

| | |
|---|---|
| **Work Package:** | WP4 (Design and architecture) |
| **Dissemination level:** | Confidential |
| **Official due date:** | 31 December 2021 |
| **Document editor:** | Johannes  Partzsch (TUD) |
| **Contributing partners:** | TUD |
| **Internal reviewers:** | Ilja Ocket (IMEC), Björn Debaillie (IMEC) |
| **Document version:** | V1 |

# 1. Publishable summary

For processing deep neural networks (DNNs) at the edge, one typically deploys conventional processors (microcontrollers, CPUs or GPUs) implemented in CMOS technology, often in combination with an external DRAM. With more and more non-volatile memory (NVM) technologies becoming available that can be integrated into the same processor chip during the CMOS manufacturing process, new opportunities arise for DNN processing at the edge. NVM is especially suited for storing the weights of neural networks, as they are fixed after training and only read accesses are needed during inference. This, for example, enables low-power edge applications where a DNN is executed only rarely at large time intervals so that the DNN chip can be disabled in between to save power.

For DNN processing, NVM can be either used for in-memory computing or as replacement for digital memories like SRAM or DRAM. This report focuses on the 2nd case and explores the integration of ferro-electric RAM into the SpiNNaker2 neuromorphic system for energy efficient DNN processing. The report contains a brief introduction to the SpiNNaker2 chip which contains 152 processing elements (PEs) each with an ARM M4F core, local SRAM and a hardware accelerator for DNN layers. A unique feature of SpiNNaker2 is that PEs can switch their performance level within a few nano seconds to either run at highest speed but higher power consumption or at lower speed but better energy efficiency. This allows to perform dynamic power management during DNN inference at lowest energy cost while fulfilling timing or maximum power limits.

To study different power management concepts, detailed performance models for DNN processing on SpiNNaker2 are developed based on power measurements of a prototype chip. The performance models in turn are used to showcase power management approaches, e.g., when the time budget is limited. By means of dynamically switching between performance levels, more than 10% of energy can be saved in contrast to the case when the faster performance level is used.

In a second step, we adapt the performance models to study the integration of ferro-electric RAM into the SpiNNaker2 chip in 22nm FDX technology for DNN processing. We consider the NVM to replace or complement the on-chip SRAM in different architecture variants. While the read energy and leakage power of the NVM are less than for the SRAM, there is no significant benefit for DNN inference as the energy of other components dominates. The main conclusion to be drawn from this study is that a simple replacement of SRAM by NVM does not improve the system-level efficiency. Instead, all parts of a DNN chip like the accelerators, the memory, and the communication system need to be co-optimized for maximum energy efficiency. The models and methods developed in this document provide a basis for such system-level optimizations and leverage the design of NVM-based DNN inference chips in the future.