

Master Thesis

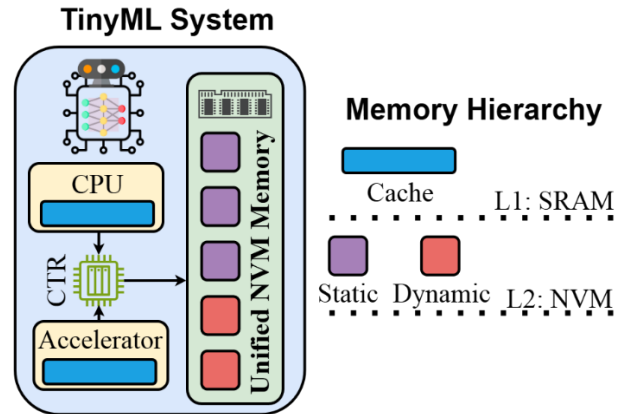
Enhancing TinyML Accuracy via NVM integration

From smart home applications to healthcare monitoring, the proliferation of Artificial Intelligence (AI) in everyday systems has driven an unprecedented demand for devices capable of executing complex Machine Learning (ML) algorithms, leading to increased cost and energy consumption. TinyML has emerged as a promising alternative to large-scale AI by bringing computation closer to the edge, enabling secure, private, and near-instant responses at low power.

At the same time, embedded IoT devices based on power-efficient hardware, such as low-cost microcontroller units (MCUs), have rapidly increased in popularity, with the market expected to reach one trillion devices by 2035. While these platforms are ideal for TinyML, they impose strict power and memory constraints. Deploying modern Deep Neural Networks (DNNs), which continue to grow in parameter count and complexity, is therefore challenging, often forcing the use of smaller models with degraded performance. This creates a pressing need to explore new memory technologies for TinyML.

To address these challenges, non-volatile memories (NVMs) are introduced, as they offer low static power consumption and high density, particularly through multi-level cell (MLC) technology. However, NVMs also present limitations, including high write latency and limited endurance, meaning only a finite number of write operations can be performed before memory degradation occurs.

In this thesis, we explore the integration of NVMs into TinyML systems. The objective is to leverage state-of-the-art neural architectures to achieve high accuracy while exploiting NVM technologies. To this end, this work investigates techniques to mitigate write latency and limited endurance, including software optimizations to reduce memory accesses and novel hardware solutions to minimize the impact of DNN inference on application latency and memory wear-out.



Skills required/beneficial for the thesis :

- Programming skills (C++, Python).
- Experience with digital design and computer architecture.
- Experience with machine learning frameworks is beneficial but not required.

Skills acquired within the thesis :

- Knowledge of emerging NVM technologies.
- Experience in co-designing hardware (NVM) and software (ML).
- Research and writing skills.

Language :

- The collaboration with the colleagues would be in English.

Contact :

- Georgios Mentzos - georgios.mentzos@kit.edu
- Miran Tobar – miran.tobar@kit.edu