Exploring learning techniques for edge AI taking advantage of NVMs

Michele Martemucci, François Rummens, Tifenn Hirtzlin, Adrien F. Vincent, Sylvain Saighi, Elisa Vianello

Exploring Learning Techniques for Edge AI Taking Advantage of NVMs

M. Martemucci[1*], F. Rummens[1], T. Hirtzlin[2], A.F. Vincent[3], S. Saighi[3] and E. Vianello[2]

[1]*CEA-LIST, Univ. Grenoble-Alpes, Grenoble, France*

[2]*CEA-LETI, Univ. Grenoble-Alpes, Grenoble, France*

[3]*Laboratoire de l'Intégration du Matériau au Système, Univ. Bordeaux, Bordeaux INP, CNRS, France*

**Corresponding author Email: michele.martemucci@cea.fr*

The relatively recent development and remarkable results of Artificial Neural Networks (ANNs) are due to the construction of gigantic databases and algorithmic innovations requiring large hardware resources, which results in equally substantial energy consumptions. As Artificial Intelligence (AI) is now being embedded more and more into various connected objects, ranging from medical implants to autonomous cars, it is clear that the algorithmic and hardware solutions available in data centres will not be able to cover all the AI integration needs. The field of microelectronics has been working for several years now on the development of emerging memory technologies with the aim of integrating Non-Volatile Memory (NVM) within computing units. In a conventional processor architecture, such co-integration between the computation units and the memory would simplify the memory hierarchy, but also increase the bandwidth between computation and data access.

In this study, we explore the potential of two non-volatile memory technologies, HfO_2-based FeRAM [1] and OxRAM [2], for enabling on-chip learning systems. Notably, the quasi-infinite reading endurance of OxRAM devices and their poor writing endurance makes them suitable for inference-only applications, whereas the reported large writing endurance of FeRAM device would effectively allow moving training on-chip as well. Eventually, the migration of inference and learning from data centres to edge devices will allow them to adapt to the evolution of input data, to specialize each device to its user, to retain private data and offer faster service.

To validate the feasibility of this approach, we designed a test chip in the 22nm FDSOI technology node. The primary objective of this chip is to demonstrate the implementation of a hybrid FeRAM/OxRAM memory circuit capable of storing the synaptic weights of a Neural Network (NN) during learning/inference phases, while accelerating NN training at the edge. Eventually, by incorporating synaptic metaplasticity in Binarized Neural Networks [3], the chip addresses the issue of catastrophic forgetting. The chip consists of two sub-cores, each comprising four 16kbit FeRAM arrays and one 16kbit OxRAM array. One FeRAM array and the OxRAM array can be operated simultaneously. The circuit leverages the OxRAM array to build a near-memory computing inference engine to accelerate the inference/feedforward pass of training, while FeRAM arrays store an 8-bit quantized version of the floating-point weights optimized during training.

Keywords: Artificial Neural Networks, Non-Volatile Memory, FeRAM, OxRAM, Edge Devices, On-Chip Learning, Inference, Synaptic Metaplasticity, Binarized Neural Networks.

References

[1] Francois, T. *et al*. 16kbit HfO2:Si-based 1T-1C FeRAM Arrays Demonstrating High Performance Operation and Solder Reflow Compatibility. *2021 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2021, pp. 33.1.1-33.1.4.

[2] Grenouillet, L *et al*. 16kbit 1T1R OxRAM arrays embedded in 28nm FDSOI technology demonstrating low BER, high endurance, and compatibility with core logic transistors. *2021 IEEE International Memory Workshop (IMW)*, Dresden, Germany, 2021, pp. 1-4.

[3] Laborieux, A., Ernoult, M., Hirtzlin, T. *et al.* Synaptic metaplasticity in binarized neural networks. *Nat Commun* 12, 2549 (2021).