Back-end-of-line integration of emerging memory technologies for neuromorphic edge computing

Luca Fehlings[a], Erika Covi[a,*]

[a] Zernike Institute for Advanced Materials & Groningen Cognitive Systems and Materials Center (CogniGron), University of Groningen, 9747 AG Groningen, The Netherlands

The transition from cloud-based data classification to edge computing has facilitated real-time data processing in closer proximity to sensors, thereby reducing latency and enhancing efficiency. However, this paradigm introduces stringent constraints on power consumption, compactness, and performance [1, 2]. Addressing these challenges necessitates unconventional hardware solutions capable of meeting these demanding requirements.

Brain-inspired architectures, notably spiking neural networks (SNNs), present a compelling solution to low-latency, stateful, and energy-efficient computation [3]. Nevertheless, existing implementations primarily depend on digital or mixed-signal Complementary Metal-Oxide-Semiconductor (CMOS) technologies, which are unable to satisfy the rigorous memory, area, and power constraints of edge computing. The integration of emerging memory technologies at the back-end-of-line (BEOL) of CMOS circuits od in 3D arrays offers a compelling opportunity to enhance neuromorphic hardware [1, 4]. Indeed, emerging non-volatile memory devices hold the promise to enable energy-efficient, massively parallel computing architectures due to their CMOS-compatible operating voltages and analogue behaviour [2, 4, 5]. These properties facilitate the hardware implementation of efficient neural dynamics and synaptic plasticity [4, 6], which are key attributes for hardware-based brain emulation. However, realising this potential requires overcoming critical challenges, including fabrication compatibility, device variability, reliability, scalability, and system integration [4, 5].

This presentation underlines the necessity of design-technology co-optimization (DTCO) to enable the seamless integration of emerging memory devices with CMOS circuits, providing a design foundation for future memory systems based on BEOL and 3D integration of emerging memory devices. It will address challenges and opportunities in the collaborative design of devices, circuits, and architectures, emphasising the necessity for a holistic approach to fully realise the potential of neuromorphic computing at the edge.

Acknowledgments

This work was supported by the European Research Council (ERC) through the European's Union Horizon Europe Research and Innovation Programme under Grant Agreement No 101042585. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The authors would like to acknowledge the financial support of the CogniGron research center and the Ubbo Emmius Funds (Univ. of Groningen).

References

1. E. Covi et al., Front. in Neurosci. **15**, 611300 (2021).
2. D. V. Christensen et al., Neuromorph. Comput. Eng. **2**, 022501 (2022).
3. E. Chicca et al., Proc. of the IEEE **102**, 1367 (2014).
4. D. Ielmini and S. Ambrogio, Nanotech. **31**, 092001 (2019).
5. A. Amirsoleimani et al., Adv. Intell. Sys., 2, 2000115 (2020).
6. E. Covi et al., Neuromorph. Comput. Eng. **2**, 012002 (2022).

[*] *corresponding author e-mail: e.covi@rug.nl*