Emerging Memory for Neuromorphic Edge Computing

Technical University of Munich, Germany; TUM School of Computation, Information and Technology

erika.covi@tum.de

The shift from cloud-based data classification toward edge computing has enabled real-time data processing closer to the source of data collection, cutting latency and improving overall efficiency. Yet this shift brings with it strict demands around power consumption, physical footprint, and computational performance. Meeting these demands calls for novel hardware approaches that can operate within such tight constraints.

Brain-inspired computing paradigms, particularly spiking neural networks (SNNs), offer a promising path toward low-latency, stateful, and energy-efficient processing. However, current implementations largely rely on digital or mixed-signal CMOS technologies, which fall short of the demanding memory, area, and power requirements typical of edge environments. Incorporating emerging memory technologies at the back-end-of-line (BEOL) of CMOS circuits, or within 3D array configurations, opens up exciting possibilities for advancing neuromorphic hardware.

Non-volatile memory devices, in particular, show strong potential for enabling energy-efficient, massively parallel computing due to their CMOS-compatible operating voltages and analogue behaviour. These characteristics make it more practical to implement efficient neural dynamics and synaptic plasticity in hardware, both of which are essential for brain-inspired emulation. Realizing this potential, however, means tackling a set of critical obstacles: fabrication compatibility, device variability, reliability, scalability, and system-level integration.

This presentation highlights the importance of design-technology co-optimization (DTCO) as a means of seamlessly combining emerging memory devices with CMOS circuits, laying a design foundation for next-generation memory systems built on BEOL and 3D integration. It will explore the challenges and opportunities that arise when co-designing devices, circuits, and architectures together, making the case for a holistic approach to enabling the full potential of neuromorphic computing at the edge.