# The Promises of Near-Memory Signal Processing for In-Memory-Computing

C. Grewing[(1)], M. Schiek[(2)], A. Ashok[(1)], A. Firdauzi[(1)], S. Kusuma[(1)],

K. Winterberg[(1)], A. Zambanini[(1)], and S. van Waasen[(1,3)]

(1) Peter Grünberg Institute, Integrated Computing Architectures ( ICA / PGI-4), Forschungszentrum Jülich, Germany
(2) Peter Grünberg Institute, Neuromorphic compute Nodes, (NCN / PGI-14), Forschungszentrum Jülich, Germany
(3) Faculty of Engineering, Communication Systems (NTS), University of Duisburg-Essen, Germany

Neuromorphic computing based on crossbar circuits is an implementation of the In-Memory-Computing principle, which enables executing a variety of algorithms directly on a memory block formed by weights at the crosspoints of the crossbar [1] [2]. Computing arrays [CA] of significant size operate with high data throughput, with Network on Chips [NoC] integrating the computing arrays in a larger fabric. However, data transport between the CAs requires power and area consumption. For many algorithms, additional data processing is required e.g., to evaluate multiplication results, prepare the signals or to feedback data samples [3]. Spike based data and asynchronous networks promise low latency and low power consumption in application specific solutions. However, for flexible use of the crossbar array, various processing alternatives must be implemented and versatile systems rely on digital signal transfer to and from the crossbar arrays [3]. To maintain the energy efficiency of in-memory computing based on crossbar circuits at the system level, low-power solutions are needed. Within the BMBF funded project NEUROTEC II we developed the system on chip ORCA with multiple CAs featuring an energy and area efficient Near-Memory signal processing solution. The design is intended to offer maximum flexibility for various neuromorphic paradigms. To ensure low latency, low power consumption and flexible use of the crossbar array data, transfer to and from the computing array is minimized. The signal processing near the crossbar array receives and transmits data from the NoC and makes it available at the interface to the crossbar array or reads it from there. The processing block in the feedback path consists of three identical blocks in a row, enabling data filtering and bit shifting, another block for amplification, bias or clipping, and a counter block, which is used for time encoding. To minimize latency, the signal feeds back to the input of the matrix, using bias or RELU functions. As an example, a Leaky Integrate-and-Fire function simulated on a Simulink-based model is shown in Fig. 1. Area and power consumption estimates from the HDL implementation are presented, along with a comparison with current analogue or digital implementations.

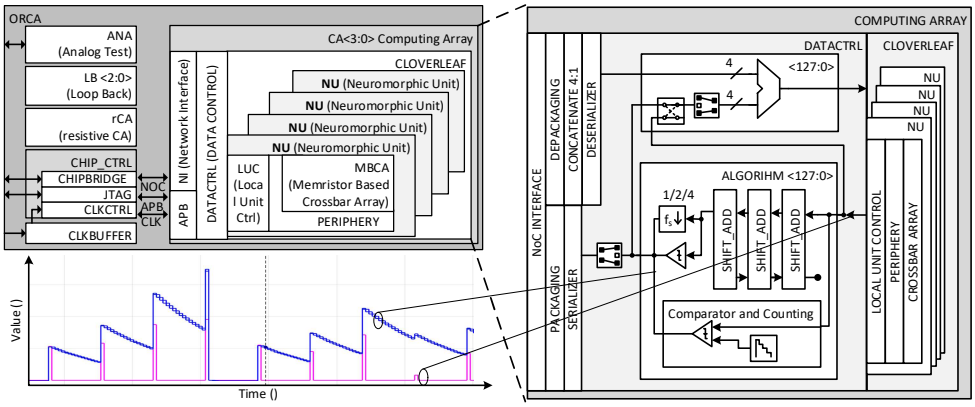


Figure 1: Orca SoC (upper left), Computing Array (right), ALGORITHM Simulation (lower left)

[1] K.Gao et al., Advanced Functional Materials, e28309, 2025

[2] C. Bengel et al., ISCAS, [Proceedings], 2022

[3] X.Duan et al., Adv Mater, 36, 14, 2024