

Multi-Partner Project: Scalable, Ferroelectric-based Accelerators for Energy Efficient Edge AI (Ferro4EdgeAI)

Theofilos Spyrou¹ Yashvardhan Biyani¹ Konstantinos Stavrakakis¹ Rajendra Bishnoi¹ Said Hamdioui¹
 Joel Minguet Lopez² Louise Dumas² Jean Coignus² Denys Ly² Hugo Chazot-Ranquet² Laurent Grenouillet²
 Fabien Grimaud² Simon Martin² Olivier Billoint² Francois Andrieu² Ruben Alcalá³ Stefan Slesazek³
 Athira Sunil³ Chong Peng³ Antoine Cauquil⁴ Rosario Pronsato⁴ Damien Deleruyelle⁴ Cédric Marchand⁴
 Alberto Bosio⁴ Ian O'Connor⁴ Giulio Urlini⁵ Simon Jeannot⁶ Mohammad Sajedi Alvar⁷
 Nima Akbari Moghaddam⁷ Thilo Werner⁷ Tony Schenk⁷ Bojun Cheng⁸ Mina Khoei⁸
 Lucía Pérez Ramírez⁹ EunJin Koh⁹ Somnath Kale⁹ Nicholas Barrett⁹

¹CE-Lab, Delft University of Technology, Delft, The Netherlands; ²CEA-Leti, Univ. Grenoble Alpes, Grenoble, France;

³NaMLab GmbH, Dresden, Germany; ⁴Lyon Institute of Nanotechnology (UMR CNRS 5270), École Centrale de Lyon, Lyon, France;

⁵STMicroelectronics SRL, Agrate, Italy; ⁶STMicroelectronics, Crolles, France; ⁷Ferroelectric Memory GmbH, Dresden, Germany;

⁸SynSense AG, Zurich, Switzerland; ⁹SPEC, CEA, CNRS, Université Paris-Saclay, F-91190 Gif-sur-Yvette, France

Abstract—The Computing-In-Memory (CIM) paradigm offers a promising solution to the memory-wall bottleneck that limits conventional Von Neumann architectures. By performing data processing at the same physical location where the data are stored, CIM-based architectures minimize costly data movement and drastically improve energy efficiency. When implemented with Ferroelectric Field Effect Transistors (FeFETs), additional advantages from the non-volatility, fast switching, and low operating voltage of FeFETs are added. However, the widespread adoption of FeFETs is limited by their poor endurance, which is overcome by a Back End of the Line (BEoL) integration of FeFET-2, where a ferroelectric capacitor (FeCAP) is wired to the gate of a CMOS transistor providing high endurance compatible with low-power edge applications. These properties enable dense, low-power, and high-speed matrix operations essential for AI workloads. As a result, FeFET-2-based CIM accelerators offer a promising solution for energy-efficient, high-performance AI at the edge. The Ferro4EdgeAI project aims to develop an ultra low-power, scalable edge accelerator for AI, targeting a significant gain in energy efficiency with respect to state-of-the-art AI hardware accelerators. To attain this, our project focuses on innovation all along the value chain from materials, physic concepts, device architecture, integration technologies, and accelerators in a holistic design space exploration approach.

Index Terms—Ferroelectric Hafnia, FeFET-2, Computing-In-Memory, Edge Computing, Artificial Intelligence

I. INTRODUCTION

Artificial Intelligence (AI) has gained a vital role in modern society and life. Deep Learning (DL) models offer solutions to a variety of tasks, such as visual, audio, and natural language processing (NLP), to serve a plethora of applications in domains ranging from healthcare to autonomous driving vehicles [1], [2]. To do so, Deep Neural Networks (DNNs) and

recent Transformer models [3] comprise a multitude of layers with billions or even trillions of parameters [4]. To satisfy the ever-growing needs of such complex models, clusters of power-hungry GPUs accessible via the cloud are employed for the training and inference processes. Remarkably, training GPT-3, which consists of 175 billion parameters, is estimated to have the same carbon footprint as driving to the Moon and back [5].

To this end, processing data close to their source has become a vital requirement for Internet of Things (IoT) applications powered by hand-held and battery-operated devices. The rise of IoT has given a significant push to the deployment of AI at the edge [6]–[9], particularly thanks to advantages such as low latency, enhanced security and privacy, energy savings and reduced bandwidth. To cope with the requirements posed by edge AI, new hardware accelerators that store and process data locally in an energy efficient manner are a necessity [10].

In light of new emerging technologies, the Computing-In-Memory (CIM) paradigm offers significant improvements in energy efficiency by merging data storage and processing in the same physical location [11]–[14]. This approach eliminates frequent data transfers that dominate traditional Von Neumann architectures. Further enhancements in energy efficiency, memory density, and performance can be achieved with memristor-based CIM accelerators [15]–[21] that employ emerging memristive devices such as Ferroelectric FETs (FeFETs) [22], [23].

The principal obstacle to the widespread adoption of FeFETs is their poor endurance [24], [25]. This is overcome by a Back End of the Line (BEoL) integration of FeFET-2, where a ferroelectric capacitor (FeCAP) is wired to the gate of a CMOS transistor. This structure enables high endurance, compatibility with ultra low-power edge applications, and non-volatility. The additional benefits of FeFET-2 devices (e.g., non-volatility, fast switching, and low operating voltage) make FeFET-2-based CIM AI engines well suited for edge computing [26].

This work received funding from Chips Joint Undertaking (Chips JU) under grant agreement N° 101135656 (project Ferro4EdgeAI). Chips JU receives support from the European Union's Horizon Europe research and innovation program. More information is available at www.ferro4edgeai.eu.

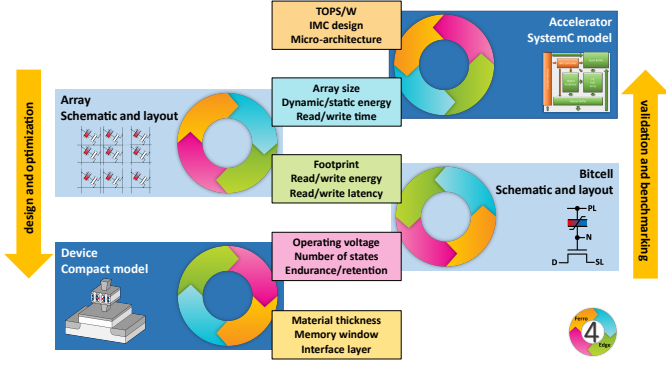


Fig. 1. Overview of the holistic DSE methodology [27] used in Ferro4EdgeAI.

Ferro4EdgeAI aims to develop intelligent edge processors based on ferroelectric (FE) technology and the CIM paradigm that will demonstrate significant energy efficiency gains of at least three orders of magnitude with respect to cloud-based CMOS. Achieving energy efficient smart edge computing is only possible using a holistic approach [27], where different mechanisms are integrated in order to collaborate across the design stack. The focus of the project is set on innovation throughout the value chain from materials, physical concepts, device architecture, integration technologies, and accelerators in a holistic Design Space Exploration (DSE) approach, as demonstrated in Fig. 1.

The Ferro4EdgeAI project will provide an ultra-low power, scalable edge accelerator for AI incorporating a memory-augmented neural network, based on low cost, high density, multi-level, BEoL integrated FE technology. With the gain in energy efficiency, the POPS/W barrier is expected to be exceeded with respect to state-of-the-art CMOS accelerators and predictions for other emerging technology AI hardware. Finally, the unique features and characteristics of FE technology are fully explored within the project in light of targeted applications on (i) image recognition and (ii) automatic speech recognition, to realize System-Technology Co-Optimization (STCO).

The remainder of the paper is structured as follows. Section II focuses on the optimization of FE material for high-stability multi-level operation. Section III explores the design and optimization of FeFET-2 devices targeting low-operating voltage. Section IV presents the design of FeFET-2-based crossbar arrays suitable for scalable systems integration. Section V completes the chain with the accelerator design and simulation. Finally, Section VI concludes the paper.

II. FERROELECTRIC MATERIAL OPTIMIZATION

The material investigated for ferroelectric optimization is hafnium zirconium oxide (HZO), a hafnia-based compound. To avoid CMOS compatibility issues, only pure hafnium zirconium oxide is evaluated as the ferroelectric material, with titanium nitride (TiN) used for both top and bottom electrodes. The study focuses on mapping a broad HZO process window. HZO film thicknesses ranging from 9 nm down to 4 nm and variations in hafnium (Hf) content from 44% to 66% have been extensively explored for this project.

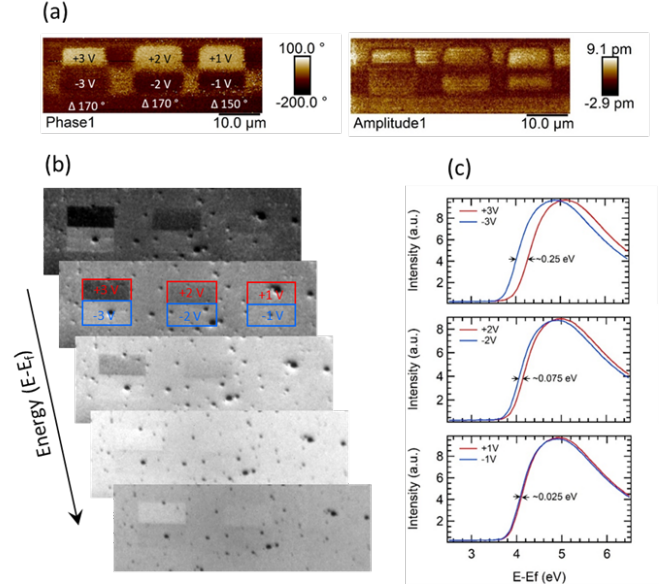


Fig. 2. Example of (a) piezoresponse phase and amplitude images of domains written with different tip biases. (b) Energy-filtered PEEM image stack acquired by varying the photoelectron kinetic energy. (c) Threshold spectra extracted from the rectangular region in (b), showing shifts in photoemission threshold between up and down polarized domains.

For fabrication, the electrode layers are deposited at room temperature using direct current (DC) sputtering and the ferroelectric HZO layer is deposited at 280 °C using Atomic Layer Deposition (ALD) with metal-organic precursors for Hf and Zr and ozone (O_3) as the oxidizing agent. The HZO composition (i.e., $Hf_xZr_{1-x}O_2$) is tuned by adjusting the ALD cycle ratios. Post-deposition, Rapid Thermal Annealing (RTA) is also required to crystallize the HZO film and, ideally, stabilize the ferroelectric phase.

As the main requirement for the material beyond ferroelectric behavior is multilevel operability, it is necessary to understand the engineering of HZO ferroelectric film reliability for such operation. The retention characteristics of switched domains using Piezo response Force Microscopy (PFM) [28] and Photoemission Electron Microscopy (PEEM) [29], [30] have been investigated for this purpose. PFM allows precise writing and mapping of polarization with nanoscale resolution, while PEEM provides complementary information about surface potential and electronic structure linked to ferroelectric domains. Preliminary PFM and PEEM project results on 10 nm HZO films are presented in Fig. 2.

Aside from the localized analysis that the different microscopy techniques can provide, it is also important to evaluate the HZO material in a device context. Stand-alone capacitors (FeCAPs) are the ideal isolated structure for such initial electrical measurements. Two reliability aspects can be highlighted for multilevel capabilities, which are wake-up and imprint. The first, shown in Fig. 3, focuses on the stability of the ferroelectric response during the lifetime of the material, i.e., how large is the remanent polarization variation between the pristine state and a cycled stage. This aspect is important because it is

necessary to have a ferroelectric material with a stable operation window for array-level implementation. Here it can be observed how the behavior can be tuned with the composition of the HZO. As an additional fabrication parameter, one can also include thickness (excluded from the current figure) to further control the performance.

For the objectives of the Ferro4EdgeAI project, the continuation in optimization for multilevel capabilities is planned, particularly for FeFET-2 devices. It is still necessary to not only understand the behavior of the ferroelectric but also minimize the impact that this can have on the subsequent levels of the project discussed in the later sections.

III. FEFET-2 DEVICE DESIGN AND OPTIMIZATION

In parallel to material optimization efforts on small silicon coupons for pre-development in academic labs, we started enabling the transfer of these learnings to full 200 mm wafer level on a production-like route.

FeFET-2 single devices and arrays are being designed, produced and optimized in close interaction between the consortium partners to meet the requirements of a CIM-based accelerator array. Single devices are important for tuning fundamental device characteristics, whereas arrays allow the identification of potential reliability threats such as cross-talk, which results in disturbs during write or read operations.

The first fabricated batch of 200 mm wafers varied the HZO thickness between 6 nm and 10 nm and the capacitance ratio (CR) of FeCAP and the transistor gate between 0.07 and 0.14. To ensure that the fabrication route works properly, Scanning Electron Microscopy (SEM), shown in Fig. 4, was used together with electrical testing of both FeCAP and FeFET-2 devices, as presented in Fig. 5.

One key challenge in the novel FeFET-2 design is the influence and control of the floating node between the FeCAP and the CMOS transistor. The floating node potential will directly influence the FeCAP device characteristics, and the effect of leakage to the transistor gate must be addressed.

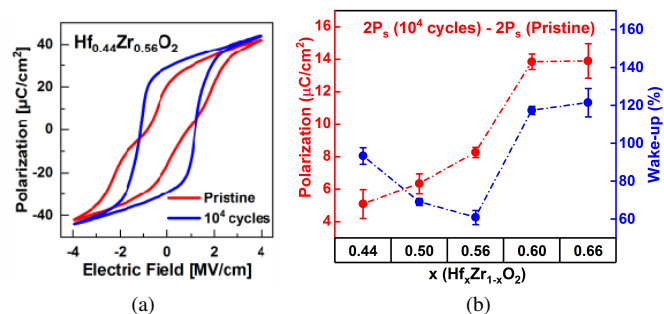


Fig. 3. Example of electrical results for 8 nm thick HZO: (a) Polarization vs. applied electric field at pristine state (red) and after 10^4 bipolar 100kHz 4MV/cm wake-up cycles (blue). (b) Remanent polarization after 10^4 bipolar 100kHz 4MV/cm wake-up cycles (red) and the percentage change from the pristine state, i.e. wake-up for different HZO compositions (blue).

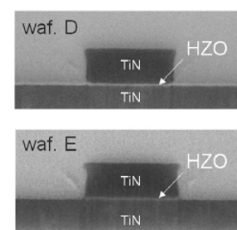
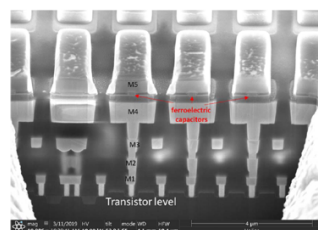
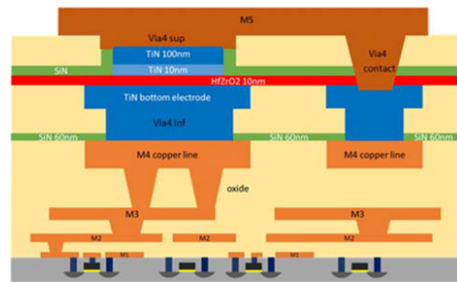


Fig. 4. (a) Schematic cross section of MAD200 wafer showing foundry front-end-of-line (FEOL) fabrication up to M4 and BEOL integration of FeCAPs up to M5. (b) SEM observation showing sub- μm -scaled FeCAPs integrated in the BEOL. (c) SEM observations of non-etched HZO 8nm and etched HZO 8nm.

To link electrical test results to material development as well as to circuit simulation, three different FeCAP models were developed:

- 1) A high-speed simplified **empirical model**, that excludes time dependency of ferroelectric switching, making it suitable for fast circuit simulations. A set of input parameters are adjustable to achieve the desired ferroelectric properties and behavior.
- 2) A **nucleation-limited switching (NLS) model**, in which the switching rate depends on the applied electric field and the duration of the applied pulse according to the assumption of domain nucleation as the limiting mechanism for the switching rate [31].
- 3) A simplified **Preisach model** in Verilog-A using a mathematical representation of saturated and non-saturated loops [32], [33].

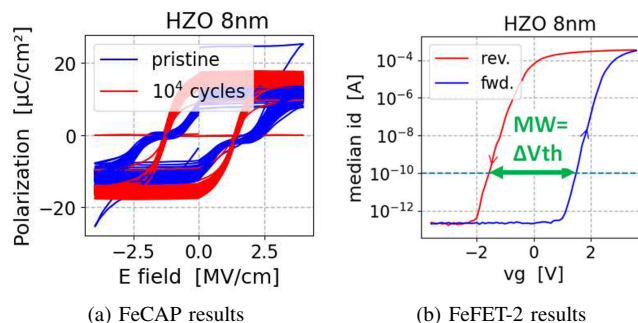


Fig. 5. Reliability of electrical results for 8 nm thick HZO: (a) Polarization vs. applied electric field at pristine state (blue) and after 10^4 bipolar 100kHz 4MV/cm wake-up cycles (red). Shorted device data are filtered from dataset. (b) Median forward (fwd.) and reverse (rev.) QS IV FeFET-2 characteristic measured at wafer scale regarding 75 devices per wafer. Horizontal dashed line represents the threshold current used for V_{th} extraction.

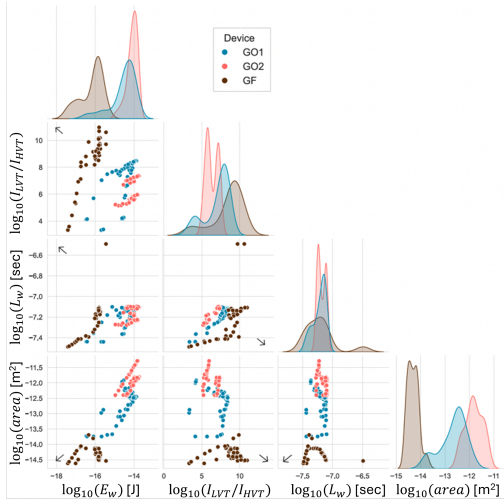


Fig. 6. Scatterplot matrix (pairplot) for 4-KPI optimization regarding write latency, write energy, current ratio, and area, using the DTCO framework.

In addition, the following structures will be further explored:

- Elementary 1 Transistor - 1 Capacitor (1T1C) FeFET-2 devices with dimensional variations for an in-depth investigation of FeFET-2 performance for Single Level Cell (SLC) and Multi-Level Cell (MLC) operation, and to refine bitcell dimensions for fabrication.
- Two types of 1T1C FeFET-2 arrays: one dedicated to low-bias SLC operation and one for higher-bias MLC operation.
- 1T2C FeFET-2 arrays to verify the promise of a novel 1T2C FeFET-2 bitcell structure.
- Exploratory concepts of FeFET-2 bitcells to assess their potential and understand the critical engineering aspects in favor of or against their future use: 1TnC, 2TnC for multi-mode / multi-context CIM, analog TCAM for analog search applications, non-volatile logic functions, and a 2x2 crossbar for (small-scale) matrix multiplications.

To assist in the design and exploration of bitcells and arrays, we implemented a powerful Design–Technology Co-Optimization (DTCO) framework [34] that links material, device, and circuit design choices to system-level performance in ferroelectric memory-based computing. This method enables designers to simultaneously tune competing goals such as speed, energy, and area, extracting Pareto-optimal trade-offs across technology generations (from 130nm to 28nm). Fig. 6 presents the bitcell optimization for four Key Performance Indicators (KPIs), namely, write latency L_w , write energy E_w , current ratio I_{LVT}/I_{HVT} , and area using the DTCO framework.

Overall, in FeFET-2 bitcells we were able to demonstrate energy reductions of up to 24x. This development is mandatory in such a vertical research project, as it provides a scalable, data-driven approach for optimizing emerging non-volatile CIM hardware; an essential step toward more energy-efficient and adaptive AI engines that exploit the physics of FE materials.

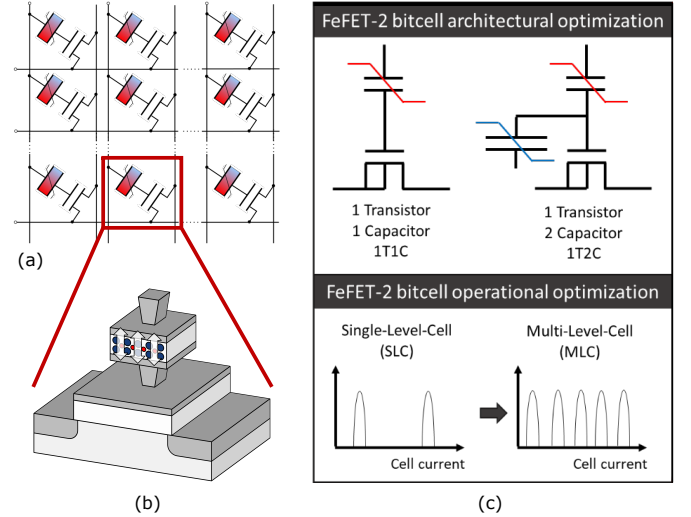


Fig. 7. (a) FeFET-2-based crossbar array. (b) FeFET-2 device: a FeCAP wired to the gate of a CMOS transistor. (c) FeFET-2 bitcell structures and programming protocols.

IV. FEFET-2 ARRAY DESIGN AND FABRICATION

While previous sections detailed the ongoing developments at both the FeFET-2 active material and bitcell levels, concretizing a neural network accelerator prototype based on this technology requires optimizing the FeFET-2 statistical reliability at the crossbar array scale. In particular, the Ferro4EdgeAI project aims to elucidate the performance of FeFET-2-based crossbar arrays in the context of CIM applications based on a 130nm CMOS technology, where the FeFET-2 cell aims to encode the synaptic weights of various types of neural networks in the analog domain. To do so, various architectural and operational characteristics of the FeFET-2-based crossbar arrays are engineered within the project, with the goal of optimizing the following metrics:

- **Crossbar array density:** Reliable FeFET-2 MLC programming protocols are developed, as shown in the bottom of Fig. 7c. This is key to maximize FeFET-2 array density, leading to lower hardware resource requirements while maintaining high accuracy with low silicon overhead.
- **Crossbar array capacity:** The FeFET-2-based crossbar arrays and the CIM accelerator are co-designed with regard to implementing a suitable array partitioning, with the aim of reducing the periphery design complexity at the accelerator scale.
- **Reliable and energy efficient programming and reading of FeFET-2 devices in crossbar arrays:** Innovative crossbar biasing schemes are developed, targeting reliable FeFET-2 bitcell operation in the array environment. In parallel, the employment of two different FeFET-2 cell structures within the array is considered, including both 1T1C and 1T2C, as shown on the top of Fig. 7c. Remarkably, this enables the evaluation of the impact of the capacitive ratio (and therefore the associated voltage divider) between the FeCAP and the MOSFET on the FeFET-2 operating reliability and energy efficiency.

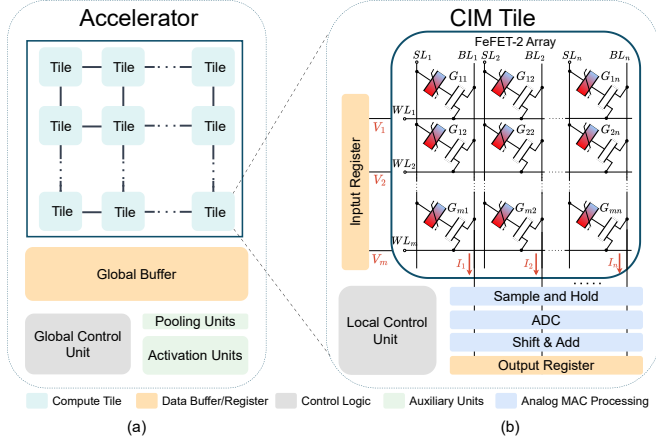


Fig. 8. Accelerator design overview. (a) Multi-tile structure with digital units composing the accelerator. (b) Micro-architecture of the CIM tile consisting of FeFET-2 crossbar array and the periphery circuitry.

- **Parallelizable reading feasibility on FeFET-2-based crossbar arrays:** Achieving fast and massively parallel computation at the crossbar scale becomes crucial to reduce the overall accelerator latency and energy consumption. Leveraging an innovative digital single/multiple cell selection mechanism at the crossbar array scale, Ferro4EdgeAI aims at evaluating the ability to parallelize FeFET-2 bitcell reading operations.

V. ACCELERATOR DESIGN AND SIMULATION

Illustrated in Fig. 8, the FE-based AI accelerator to be delivered by Ferro4EdgeAI is built upon the FeFET-2 devices and arrays presented in previous sections and follows the CIM paradigm to achieve ultra low-power execution of AI applications at the edge. The building block of the accelerator is the CIM tile, as presented in Fig. 8b. The CIM tile is composed of a micro-architecture design that uses a FeFET-2 array described in Section IV combined with the necessary digital periphery circuitry. Being the heart of the accelerator, the CIM tile frames its functionality to ensure the execution of the backbone of AI workloads, that is, the Matrix-Vector Multiplication (MVM) operation.

To be capable of supporting the execution of a complete neural network inference, the accelerator is equipped with multiple CIM tiles, as illustrated in Fig. 8a. The tiles are connected to each other, while the global control unit handles all inter-communication among tiles, loads the data from the input buffer, and streams (partial) results to next layer blocks or to the output. Since the size of crossbar array is kept relatively small (typically in the range of a few kilobits), a multi-tile structure is a necessity for fitting all the network’s synaptic weights and parameters. Finally, additional digital units dedicated to the calculation of AI-related kernels that cannot be processed by the CIM tiles, such as pooling or activation functions, complete the overall accelerator design.

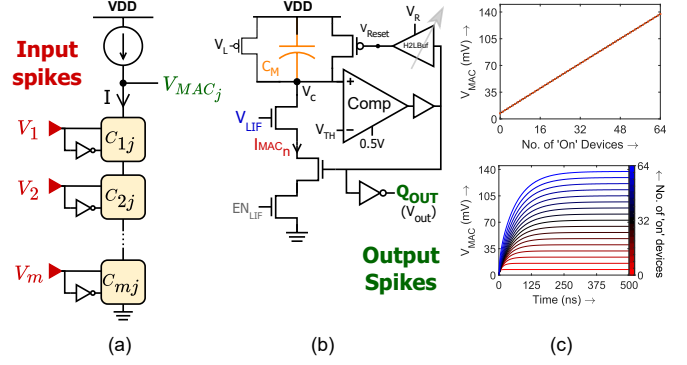


Fig. 9. C3CIM micro-architecture [35]. (a) A column of the memristive CIM architecture. (b) LIF circuit. (c) I/O characteristics.

A. CIM micro-architecture design and implementation

A memristive CIM tile consists of two main parts, namely, the *crossbar array* and the *periphery circuitry*, as illustrated in Fig. 8b. The periphery circuitry is essential to provide a digital interface for the use of the analog crossbar array, where the local control unit handles the overall operation. At the input, the input register drives each wordline row i with input voltage $V_i \forall i \in [1, m]$, applied to all devices in that row. At the output, Analog-to-Digital Converters (ADCs) convert the bitline current $I_j \forall j \in [1, n]$ to digital values $D_j \forall j \in [1, n]$. The crossbar array is an $m \times n$ matrix of memory *bitcells*. Fig. 8b showcases the typical FeFET-2 1T1C bitcells, but any other bitcell topology presented in Section IV can be utilized. In the case of Spiking Neural Network (SNN) applications, special periphery circuitry is employed to implement the spiking neuron, such as the Leaky Integrate and Fire (LIF) circuit shown in Fig. 9b. The LIF component integrates I_j directly in the analog domain, updating the membrane potential V_C of the spiking neuron, which fires a spike at the output when threshold V_{TH} is exceeded.

The FeFET-2 bitcells store the data in terms of gate threshold voltage states, i.e., High Threshold Voltage (HTV) representing logic ‘0’, and Low Threshold Voltage (LTV) representing logic ‘1’. When input voltage V_i is applied on the i^{th} row, the individual current of cell C_{ij} is equal to $I_{ij} = V_i \cdot G_{ij}$ according to Ohm’s law, where G_{ij} represents the transistor’s effective conductance. The bitline output current I_j at the j^{th} column is then equal to the accumulation of the output current of the bitcells in that column in accordance with Kirchoff’s current law, as shown in Eq. 1:

$$I_j = \sum_{i=1}^m V_i \cdot G_{ij}, \quad \forall j \in [1, n] \quad (1)$$

In other words, a Multiply-Accumulate (MAC) operation of the input voltages and memristor states is realized in a single column. Since all columns operate in parallel, the memristive crossbar can perform an MVM operation with complexity $O(1)$, demonstrating superior performance and potential to efficiently accelerate AI applications [36].

As the size of the array increases, the accumulated MAC currents can become orders of magnitude higher than the currents of individual cells, resulting in increased power consumption. In combination with circuit-level non-idealities, such as IR drop, this causes non-linearity in the digital conversion and, therefore, erroneous outputs, which degrade computational accuracy.

To address these challenges, a Constant Column Current CIM (C3CIM) [35] micro-architecture is proposed. Fig. 9a illustrates the generic schematic of a single column. Output is generated in the form of an analog voltage V_{MAC_j} , which is then passed to a trans-conductance amplifier to produce the corresponding current I_{MAC_j} . In the end, an LIF circuit integrates the MAC currents to implement spiking functionality. The four unique features of this micro-architecture are listed below:

- A constant current readout scheme maintains a very low and fixed column current regardless of the number of activated rows. The output V_{MAC_j} scales *linearly* with the number of active bitcells storing logic ‘1’ within the column, as shown in Fig. 9c.
- A 2T1C bitcell structure with complementary inputs allows the bitcell resistance to be a combination of input and weight. Their serial arrangement enables the outcome of the MAC operation to be encoded as column resistance.
- Because of the complementary inputs, C3CIM inherently compensates for the non-zero resistance of the access transistor by maintaining a fixed number of access transistors in the current path, which can be treated as a constant offset during post-processing.
- Natural compatibility with SNNs is supported as a result of the binary input operation.

Preliminary evaluations demonstrate an energy efficiency of 30 fJ/MAC (≈ 0.5 fJ for 1-bit \times 1-bit MAC). It consumes an average of 14.6 nJ per inference for a Binary Neural Network (BNN) trained on the MNIST dataset with a topology of $784 \times 3136 \times 10$ neurons, and 11.2 nJ per inference for an SNN trained on a reduced-resolution version of MNIST with a topology of $324 \times 81 \times 10$ neurons for 60 time steps. Given the truncated topologies, both networks maintain a balanced trade-off between energy consumption and performance, achieving a classification accuracy of 88.4% and 87.2%, respectively.

B. Accelerator simulator

Before fabricating a prototype of the FE accelerator, a simulation framework is essential to iteratively optimize the design, where KPIs such as energy consumption and area are re-evaluated to guide design refinements. The framework spans multiple abstraction levels, namely, *device level*, *CIM tile level*, *accelerator level*, and *system level*.

At the device level, a set of candidate bitcell structures is analyzed and optimized for feasible precision and reliability using the DTCO framework described in Section III. For each bitcell structure, a behavioral model is derived to serve as input to higher-level simulations. At the CIM tile level, the simulator performs DSE over parameters such as array size, precision, and resource sharing policies. At the accelerator level, it estimates latency and energy, including the impact of inter-tile

communication, while considering a small set of commands from the global controller’s Instruction Set Architecture (ISA). Finally, the full-system simulator integrates the accelerator with a processing unit and external memory to execute end-to-end neural network inference and report accuracy, throughput, energy, and memory traffic.

After bitcell optimization using the DTCO framework, the CIM tile simulation focuses on speed, accuracy, parameterization, and a clear vertical interface, as described below:

- **Speed:** Tile models must run fast to support broad DSE without becoming a bottleneck to DNN inference.
- **Accuracy:** The tile must reflect the effects of periphery and digital conversion of a real-world, non-ideal crossbar array, so as to produce credible cycle counts and energy numbers to feed higher abstraction levels.
- **Parameterization:** A configurable tile description can represent many hardware points and align with technology swaps through a component library, preserving consistency over implementation changes.
- **Clear interface:** A callable Application Programming Interface (API) provides support for a cross-layer co-optimization, where the various abstraction levels can receive/provide input from/to lower/higher levels.

For the accelerator simulation and full-system integration and validation, realistic workloads are essential. We therefore use DNN applications as workloads, where both fully connected and convolutional layers are mapped to the same MAC primitive to target edge tasks such as anomaly detection, audio denoising, eye-tracking, and object detection. The application kernels are compiled into the accelerator’s ISA and simulated across diverse hardware configurations. This systematic benchmarking will provide simulation results to quantify performance, energy efficiency, and throughput tradeoffs across the design space, providing data-driven guidance for hardware instantiation decisions that will shape the final accelerator design before prototyping.

VI. CONCLUSIONS

The Ferro4EdgeAI holistic approach across the entire value chain aims to position FE technologies for ultra low-power edge AI applications. First, our project targets multi-level functionality in hafnia-based thin films. The FeFET-2-based bitcells will aim for low operating voltage and robust multi-level operation for high density logic operations and data storage. BEoL design, fabrication, and characterization of multi-level, low voltage, FeFET-2 arrays is underway, allowing to define, design, build and demonstrate a low-power, scalable FE AI accelerator suitable for scalable systems integration. The final objective entails the development of a hierarchical simulation framework that integrates the calibrated models developed. Simulations from device to system level enable us to perform DSE across the complete technological stack for design and performance optimizations. Moreover, Ferro4EdgeAI dedicates resources to the study of the ecological impact of (i) the constituting FE materials and their fabrication processes and (ii) the FeFET-2-based accelerator prototype during its active lifetime, quantified for the targeted applications.

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [2] S. Diware *et al.*, "Reliable and energy-efficient diabetic retinopathy screening using memristor-based neural networks," *IEEE Access*, vol. 12, pp. 47 469–47 482, 2024.
- [3] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [4] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, no. 120, pp. 1–39, 2022.
- [5] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 09, 2020, pp. 13 693–13 696.
- [6] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.
- [7] N. D. Lane, S. Bhattacharya, A. Mathur, P. Georgiev, C. Forlivesi, and F. Kawars, "Squeezing deep learning into mobile and embedded devices," *IEEE Pervasive Comput.*, vol. 16, no. 3, pp. 82–88, 2017.
- [8] A. Singh *et al.*, "Low-power memristor-based computing for edge-ai applications," in *ISCAS*, 2021, pp. 1–5.
- [9] S. Diware *et al.*, "Adaptive multi-threshold encoding for energy-efficient eeg classification architecture using spiking neural network," in *DATE*, 2025, pp. 1–7.
- [10] A. Gebregiorgis, A. Singh, A. Yousefzadeh, D. Wouters, R. Bishnoi, F. Catthoor, and S. Hamdioui, "Tutorial on memristor-based computing for smart edge applications," *Mem. - Mater. Devices Circuits Syst.*, vol. 4, p. 100025, 2023.
- [11] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature nanotechnology*, vol. 15, no. 7, pp. 529–544, 2020.
- [12] Y. Biyani *et al.*, "SABCIM: Self-Adaptive Biasing Scheme for Accurate and Efficient Analog Compute-in-Memory," in *IEEE ASP-DAC*, 2026, pp. 1–7.
- [13] L. Huijbregts *et al.*, "Energy-efficient snn architecture using 3nm finfet multiport sram-based cim with online learning," in *DAC*, 2024, pp. 1–6.
- [14] L. Huijbregts, M. D. Gomony, A. Gebregiorgis, F. Catthoor, M. Taouil, R. Joshi, S. Hamdioui *et al.*, "Dream-cim: A digital sram-based cim accelerator for energy-and area-efficient edge ai," *IEEE Transactions on Circuits and Systems for Artificial Intelligence*, 2025.
- [15] Y. Long, E. Lee, D. Kim, and S. Mukhopadhyay, "Flex-PIM: A ferroelectric FET based vector matrix multiplication engine with dynamical bitwidth and floating point precision," in *Proc. Int. Jt. Conf. Neural Netw. (IJCNN)*, 2020, pp. 1–8.
- [16] T. Soliman *et al.*, "Felix: A ferroelectric FET based low power mixed-signal in-memory architecture for DNN acceleration," *Trans. Embed. Comput. Syst.*, vol. 21, no. 6, pp. 1–25, 2022.
- [17] R. Bishnoi *et al.*, "Energy-efficient computation-in-memory architecture using emerging technologies," in *IEEE ICM*, 2023, pp. 325–334.
- [18] S. Diware *et al.*, "Severity-based hierarchical eeg classification using neural networks," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 17, no. 1, pp. 77–91, 2023.
- [19] A. Singh *et al.*, "A 115.1 TOPS/W, 12.1 TOPS/mm² Computation-in-Memory using Ring-Oscillator based ADC for Edge AI," in *IEEE AICAS*, 2023, pp. 1–5.
- [20] S. Diware *et al.*, "Mapping-aware biased training for accurate memristor-based neural networks," in *IEEE AICAS*, 2023, pp. 1–5.
- [21] —, "Accurate and energy-efficient bit-slicing for rram-based neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 1, pp. 164–177, 2022.
- [22] C. Marchand, I. O'Connor, M. Cantan, E. T. Breyer, S. Slesazek, and T. Mikolajick, "FeFET based logic-in-memory: an overview," in *Int. Conf. Des. Technol. Integr. Syst. (DTIS)*, 2021, pp. 1–6.
- [23] A. Gebregiorgis, H. A. Du Nguyen, J. Yu, R. Bishnoi, M. Taouil, F. Catthoor, and S. Hamdioui, "A survey on memory-centric computer architectures," *ACM J. Emerg. Technol. Comput. Syst. (JETC)*, vol. 18, no. 4, pp. 1–50, 2022.
- [24] M. Trentzsch *et al.*, "A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs," in *Int. Electron Devices Meet. (IEDM)*, 2016, pp. 11–5.
- [25] A. I. Khan, A. Keshavarzi, and S. Datta, "The future of ferroelectric field-effect transistor technology," *Nature Electronics*, vol. 3, no. 10, pp. 588–597, 2020.
- [26] A. Keshavarzi, K. Ni, W. Van Den Hoek, S. Datta, and A. Raychowdhury, "Ferroelectronics for edge intelligence," *Micro*, vol. 40, no. 6, pp. 33–48, 2020.
- [27] M. Niemier *et al.*, "Cross layer design for the predictive assessment of technology-enabled architectures," in *Des. Autom. Test Eur. (DATE)*, 2023, pp. 1–10.
- [28] A. Gruverman, M. Alexe, and D. Meier, "Piezoresponse force microscopy and nanoferroic phenomena," *Nature communications*, vol. 10, no. 1, p. 1661, 2019.
- [29] N. Barrett *et al.*, "Full field electron spectromicroscopy applied to ferroelectric materials," *J. Appl. Phys.*, vol. 113, no. 18, 2013.
- [30] W. Hamouda, F. Mehmood, T. Mikolajick, U. Schroeder, T. O. Mentis, A. Locatelli, and N. Barrett, "Oxygen vacancy concentration as a function of cycling and polarization state in TiN/Hf_{0.5}Zr_{0.5}O₂/TiN ferroelectric capacitors studied by x-ray photoemission electron microscopy," *Appl. Phys. Lett.*, vol. 120, no. 20, 2022.
- [31] R. Guido, X. Wang, B. Xu, R. Alcala, T. Mikolajick, U. Schroeder, and P. D. Lomenzo, "Ferroelectric Al_{0.85}Sc_{0.15}N and Hf_{0.5}Zr_{0.5}O₂ domain switching dynamics," *ACS Appl. Mater. Interfaces*, vol. 16, no. 32, pp. 42 415–42 425, 2024.
- [32] S. Miller, J. Schwank, R. Nasby, and M. Rodgers, "Modeling ferroelectric capacitor switching with asymmetric nonperiodic input signals and arbitrary initial conditions," *J. Appl. Phys.*, vol. 70, no. 5, pp. 2849–2860, 1991.
- [33] B. Jiang *et al.*, "Computationally efficient ferroelectric capacitor model for circuit simulation," in *Symp. on VLSI Technol.*, 1997, pp. 141–142.
- [34] R. Pronsato, A. Cauquil, P. Vivet, J. Coignus, D. Deleruyelle, C. Marchand, and I. O'Connor, "Exploring enhancements of 1T1C FeMFET bitcell with a versatile DTCO methodology," in *Int. Conf. on Very Large Scale Integration (VLSI-SoC)*, 2025.
- [35] Y. Biyani, R. Bishnoi, T. Spyrou, and S. Hamdioui, "C3CIM: Constant column current memristor-based computation-in-memory micro-architecture," in *DATE*, 2025, pp. 1–7.
- [36] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. of the IEEE*, vol. 105, no. 12, pp. 2295–2329, Sep. 2017.