

FETs, a highly CMOS-compatible and scalable eNVM device that has become particularly competitive and widely adopted in recent research [15]. To quantify the energy, gas, and material impact of HZO fabrication, we gather process descriptions for atomic layer deposition (ALD) steps together with tool power measurements directly from a semiconductor fab [16]. In this context, recipe specifically refers to the real fab-based data sets. These data enable quantitative estimation of embodied carbon associated with FeFET-specific fabrication steps.

We introduce a carbon model that includes embodied carbon and operational carbon, COFFEE, for HZO FeFETs (Figure 1). The modeling flow separates the total embodied carbon into two parts: a CMOS baseline computed with ACT [4], which relies on iMec characterization [17], and a FeFET-specific component. For the latter, the model provides a step-by-step analysis of fabrication stages and corresponding carbon footprint impact from recipes and quantifies process-specific parameters, thereby offering clear insights for device and technology designers. Furthermore, we collect recent publications on HZO-based FeFETs to extract parameters for lifetime and embodied carbon analysis [18]–[24]. Finally, we employ NVMEexplorer [10] to evaluate operational parameters such as memory array latency and energy, enabling a systematic discussion of the trade-off between operational carbon and embodied carbon across diverse architectural configurations and application workloads. The main contributions of this work are:

- 1) We propose a framework (COFFEE) for evaluating the embodied carbon emissions of HZO-based FeFET eNVMs. COFFEE models the fabrication process, with detailed step-level analysis and configurable parameters according to emerging device characteristics.
- 2) We present a comparative analysis of embodied carbon emissions between HZO-based FeFETs vs. conventional SRAM at a fixed storage capacity of 2 MB. Under the same optimal-target configuration. The embodied carbon overhead of HZO-FeFETs per unit area can be as high as 11% compared to CMOS baseline. We also compare representative HZO-FeFETs and SRAM across capacities from 2 MB to 32 MB, the embodied carbon per MB remains consistently around $4.3\times$ better than SRAM, highlighting the significant sustainability advantage achievable through high memory density.
- 3) Using COFFEE, we study the carbon impact of implementing edge AI accelerators using HZO-FeFETs. The case study shows 42.3% embodied carbon reduction and up to 70% operational carbon reduction per inference.

COFFEE is available in the below GitHub repository: <https://github.com/S4AI-CornellTech/COFFEE>

II. BACKGROUND AND RELATED WORKS

In this section, Section II-A overviews carbon modeling methodologies and related work, Section II-B summarizes the advantages and challenges of FeFET technologies with an emphasis on HZO-based FeFET, and Section II-C discusses the limitations of existing FeFET eNVM models.

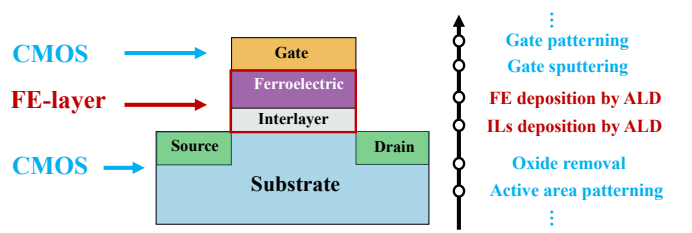


Fig. 2. Take HZO FeFET as an example, the manufacturing steps can be divided into two parts. One is the traditional CMOS baseline, marked by blue color, the other is FEOL special steps to deposit a ferroelectric layer and (optionally) an interfacial layer, bounded by red outline. [22].

A. Carbon modeling

The total carbon footprint (CF) of a computing system is defined as the sum of its operational (OCF) and embodied carbon footprints (ECF), with the embodied carbon proportion according to the fraction of application run-time (T) over the chip expected lifetime (LT).

$$CF = OCF + \frac{T}{LT} ECF \quad (1)$$

Embodied carbon can be calculated by quantifying each component of manufacturing emissions [4]. More concretely, the embodied footprint (ECF) is computed by the carbon per unit area (CPA) multiplied by the die area (A), which in turn is dependent on the fab yield (Y), the electrical energy consumed per unit area (EPA), the carbon intensity of the electrical energy (CI), the emissions per unit area from process gases (GPA), and the emissions per unit area associated with raw-material procurement (MPA), as expressed below:

$$\begin{aligned} ECF_{SoC} &= CPA \times A \\ &= \frac{1}{Y} ((CI_{fab} \times EPA + GPA + MPA) \times A) \end{aligned} \quad (2)$$

Recent studies show that the carbon footprint of mobile systems has shifted from operational to embodied emissions, directing the focus of sustainable computing research toward the manufacturing stage [3]. Existing carbon estimation frameworks, including ACT [4], 3D-Carbon [25] and ECO-CHIP [26] for 3D and 2.5D ICs, EPiCarbon [27], and nano-3D system [28], focus on CMOS or other non-eNVM, existing process flows. Their manufacturing energy models rely on CMOS data from iMec [17]. As a result, they give no coverage to the extra and modified stages used in eNVM fabrication without flexibility to incorporate device-specific parameters.

B. FeFET Technology

As shown in Figure 2, FeFETs incorporate a ferroelectric layer into the gate stack, enabling threshold modulation through ferroelectric polarization to encode a stored value [8]. This mechanism provides non-volatility and compact single-transistor bitcells, making FeFETs attractive for high-density on-chip memory with low leakage. HZO-based FeFETs have emerged as a highly CMOS-compatible technology, enabling large-scale integration at the advances node [12]. Recent studies have reported HZO-based FeFETs scaled to sub-5 nm ferroelectric thickness while still exhibiting strong ferroelectricity [15]. This scaling enables polarization switching at relatively low

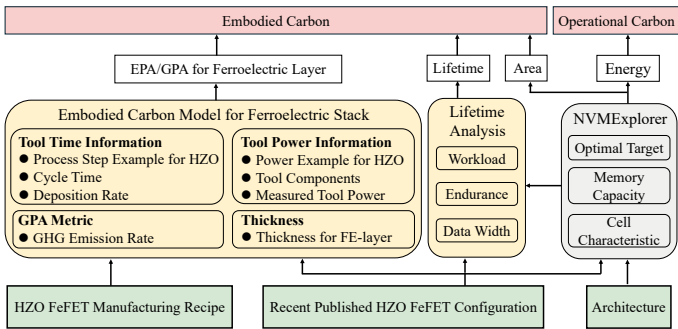


Fig. 3. Diagram of carbon calculation in the proposed COFFEE framework. Source inputs (green) include manufacturing recipe and recent published device configurations [18]–[24]. Pink blocks represent carbon categories. Yellow blocks denote the integrated embodied carbon tools and lifetime analysis, while gray boxes [10] indicate carbon-related attributes integrated in existing tools.

operating voltages, thereby reducing overall write energy and also lowering the demand for peripheral circuits, such as charge pumps, that supply the write voltage.

However, HZO-based FeFETs face critical challenges, notably limited endurance caused by charge trapping [23]. Excessive write operations exacerbate these effects, further degrading device reliability and reducing operational lifetime.

C. Performance modeling tools of eNVMs

Several tools are capable of evaluating and optimizing the energy efficiency of eNVM-based systems, with applicability to FeFET-based designs such as HZO-FeFET. NVSim [14] characterizes eNVM array power and area at the circuit level for a particular cell configuration and device type. NVM-Explorer [10] offers a cross-stack simulation platform that combines top-conference device databases and application characteristics with underlying characterization via NVSim. These tools quantify FeFET power, performance, and area (PPA) and demonstrate its runtime energy advantages as on-chip memory. However, existing tools do not address life cycle sustainability, including the impact of manufacturing.

III. PROPOSED FRAMEWORK

As shown in Figure 3, COFFEE is a cross-stack framework that integrates two key new components: (i) an embodied carbon model that estimates the fabrication-stage emissions of HZO-based FeFETs, and (ii) an operational carbon model that derives operational energy eNVM architectural design space exploration tools (i.e., NVMExplorer [10]). Inputs to COFFEE include both manufacturing recipe and HZO FeFET cell characteristics from recent device publications. The recipes provide detailed characterization of process steps, including examples of tool power and tool time, while publications supply process thicknesses for embodied carbon evaluation and endurance data for lifetime analysis. The HZO FeFET cell characteristics are combined with architecture parameters such as memory capacity to form NVMExplorer configurations. NVMExplorer explores design space under user-defined optimal target constraint (e.g., area, read/write latency, memory leakage power, and read/write energy delay product) to identify feasible design points. The overall EPA/CPA of HZO-FeFETs

TABLE I
PARAMETER RANGES AND SOURCES. THE *Source* COLUMN INDICATES ORIGIN, AS DESCRIBED IN SEC III-A

Parameter	Range	Source
ALD manufacturing process flow design related parameters		
t_{layer}	3 (Al_2O_3), 20 (HZO) nm	Recipe
T_{cycle}	30 (Al_2O_3), 100 (HZO) s	Recipe
R_{dep}	0.1 (Al_2O_3), 0.2 (HZO) nm/cycle	Recipe
R_{GHG}	26.9 $\mu\text{g}/(\text{nm}\cdot\text{cm}^2)$	[29]
Foundry related parameters		
Process node	28 nm	[17]
GPA_{CMOS}	0.1375 $\text{kg CO}_2/\text{cm}^2$	[17]
MPA_{CMOS}	0.5 $\text{kg CO}_2/\text{cm}^2$	[17]
EPA_{CMOS}	0.9 kWh/cm^2	[17]
$\text{GPA}_{\text{Fe-layer}}$	225.96 $\mu\text{g CO}_2/\text{cm}^2$	[29]
$\text{EPA}_{\text{Fe-layer}}$	0.26 kWh/cm^2	Recipe

(Section III-A) is derived from the output parameters EPA/CPA of the ferroelectric layers (Section III-B). The EPA/CPA value for FeFET, together with lifetime estimates (Section III-C) and runtime energy metrics, are then fed into the carbon models [4], to enable end-to-end carbon evaluation of HZO FeFET.

A. Method for FeFET carbon modeling

The embodied carbon of FeFET-based chips is determined by manufacturing energy consumption, GHG emissions, and facility overheads, which are computed using Equation (2). Table I provides the detailed parameter definitions and the corresponding ranges. The *Source* column lists the origin of each parameter: standard CMOS data from iMec [17], HZO process flow data from recipe, and GPA metrics from [29].

1) *EPA accounting for FeFET*: To compute the FeFET manufacturing energy, we separate the contributions from the conventional CMOS baseline and the FeFET-specific layers. An area-weighted formulation is adopted, where the ferroelectric layer is assumed to be precisely deposited (e.g., the ferroelectric ALD steps) only to the FeFET array region. The per-unit-area manufacturing energy of the CMOS baseline (EPA_{CMOS}) is based on SOTA architectural carbon modeling tools (i.e., ACT [4]). Summing the two components yields the overall per-unit-area manufacturing energy of FeFET ($\text{EPA}_{\text{FeFET}}$), where the ratio of the ferroelectric deposition area to the total on-chip memory area is obtained from the reported area efficiency (AE) of cell arrays vs. total array area including CMOS peripherals [10]:

$$\text{EPA}_{\text{FeFET}} = \text{EPA}_{\text{CMOS}} + \text{EPA}_{\text{Fe-layer}} \cdot \text{AE} \quad (3)$$

2) *GPA accounting for FeFET*: Similar to EPA accounting, an area-weighted formulation will be used to integrate the GHG emission for FeFET accounting for the complete on-chip memory:

$$\text{GPA}_{\text{FeFET}} = \text{GPA}_{\text{CMOS}} + \text{GPA}_{\text{Fe-layer}} \cdot \text{AE} \quad (4)$$

3) *MPA and Yield accounting for FeFET*: For MPA and Yield, the additional ALD steps for HZO-FeFET integration introduce no high-emission precursors and maintain compatibility with standard CMOS processes [12]. Therefore, both MPA and Yield are assumed to be identical to those of the baseline CMOS flow.

REFERENCES

- [1] C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. S. Blair, and A. Friday, "The real climate and transformative impact of ict: A critique of estimates, trends, and regulations," *Patterns*, vol. 2, no. 9, 2021.
- [2] A. S. Andrae and T. Edler, "On global electricity usage of communication technology: trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117–157, 2015.
- [3] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks, and C.-J. Wu, "Chasing carbon: The elusive environmental footprint of computing," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 854–867.
- [4] U. Gupta, M. Elgamal, G. Hills, G.-Y. Wei, H.-H. S. Lee, D. Brooks, and C.-J. Wu, "Act: designing sustainable computer systems with an architectural carbon modeling tool," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, 2022, pp. 784–799.
- [5] S. Salahuddin, K. Ni, and S. Datta, "The era of hyper-scaling in electronics," *Nature electronics*, vol. 1, no. 8, pp. 442–450, 2018.
- [6] Y. S. Shao, J. Clemons, R. Venkatesan, B. Zimmer, M. Fojtik, N. Jiang, B. Keller, A. Klinefelter, N. Pinckney, P. Raina *et al.*, "Simba: Scaling deep-learning inference with multi-chip-module-based architecture," in *Proceedings of the 52nd annual IEEE/ACM international symposium on microarchitecture*, 2019, pp. 14–27.
- [7] B. Zimmer, R. Venkatesan, Y. S. Shao, J. Clemons, M. Fojtik, N. Jiang, B. Keller, A. Klinefelter, N. Pinckney, P. Raina *et al.*, "A 0.32–128 tops, scalable multi-chip-module-based deep neural network inference accelerator with ground-referenced signaling in 16 nm," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 4, pp. 920–932, 2020.
- [8] 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.
- [9] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang *et al.*, "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks," *Nature communications*, vol. 9, no. 1, p. 2385, 2018.
- [10] L. Pentecost, A. Hankin, M. Donato, M. Hempstead, G.-Y. Wei, and D. Brooks, "Nvmexplorer: A framework for cross-stack comparisons of embedded non-volatile memories," in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2022, pp. 938–956.
- [11] M. Hellenbrand, I. Teck, and J. L. MacManus-Driscoll, "Progress of emerging non-volatile memory technologies in industry," *MRS communications*, vol. 14, no. 6, pp. 1099–1112, 2024.
- [12] A. Aziz, E. T. Breyer, A. Chen, X. Chen, S. Datta, S. K. Gupta, M. Hoffmann, X. S. Hu, A. Ionescu, M. Jerry *et al.*, "Computing with ferroelectric fets: Devices, models, systems, and applications," in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2018, pp. 1289–1298.
- [13] A. Chen, "A review of emerging non-volatile memory (nvm) technologies and applications," *Solid-State Electronics*, vol. 125, pp. 25–38, 2016.
- [14] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 7, pp. 994–1007, 2012.
- [15] S. S. Cheema, D. Kwon, N. Shanker, R. Dos Reis, S.-L. Hsu, J. Xiao, H. Zhang, R. Wagner, A. Datar, M. R. McCarter *et al.*, "Enhanced ferroelectricity in ultrathin films grown directly on silicon," *Nature*, vol. 580, no. 7804, pp. 478–482, 2020.
- [16] S. Jadhav, V. Gund, B. Davaji, D. Jena, H. G. Xing, and A. Lal, "Hzo-based ferronems mac for in-memory computing," *Applied Physics Letters*, vol. 121, no. 19, 2022.
- [17] M. G. Bardon, P. Wuytens, L.-Å. Ragnarsson, G. Mirabelli, D. Jang, G. Willems, A. Mallik, A. Spessot, J. Ryckaert, and B. Parvais, "Dtco including sustainability: Power-performance-area-cost-environmental score (ppace) analysis for logic technologies," in *2020 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2020, pp. 41–4.
- [18] E. Yu, G. K. K. U. Saxena, and K. Roy, "Ferroelectric capacitors and field-effect transistors as in-memory computing elements for machine learning workloads," *Scientific Reports*, vol. 14, no. 1, p. 9426, 2024.
- [19] T. Hu, X. Sun, M. Bai, X. Jia, S. Dai, T. Li, R. Han, Y. Ding, H. Fan, Y. Zhao *et al.*, "Enlargement of memory window of si channel FeFET by inserting Al₂O₃ interlayer on ferroelectric Hf_{0.5}Zr_{0.5}O₂," *IEEE Electron Device Letters*, vol. 45, no. 5, pp. 825–828, 2024.
- [20] C.-Y. Liao, K.-Y. Hsiang, F.-C. Hsieh, S.-H. Chiang, S.-H. Chang, J.-H. Liu, C.-F. Lou, C.-Y. Lin, T.-C. Chen, C.-S. Chang *et al.*, "Multibit ferroelectric FET based on nonidentical double HfZrO₂ for high-density nonvolatile memory," *IEEE Electron Device Letters*, vol. 42, no. 4, pp. 617–620, 2021.
- [21] Y. Zhou, W. Huang, R. Zhu, R. Huang, and K. Tang, "A reliable 2-bit MLC FeFET with high uniformity and 10⁹ endurance by gate stack and write pulse co-optimization," in *2024 IEEE European Solid-State Electronics Research Conference (ESSERC)*. IEEE, 2024, pp. 657–660.
- [22] Y. Zhou, H. Shao, W. Huang, R. Zhu, Y. Zhang, R. Huang, and K. Tang, "A compact writing scheme for the reliability challenges in 1t multi-level fefet array: Variation, endurance and write disturb," *IEEE Electron Device Letters*, 2024.
- [23] Y. Zhou, Z. Liang, W. Luo, M. Yu, R. Zhu, X. Lv, J. Li, Q. Huang, F. Liu, K. Tang *et al.*, "Ferroelectric and interlayer co-optimization with in-depth analysis for high endurance fefet," in *2022 International Electron Devices Meeting (IEDM)*. IEEE, 2022, pp. 6–2.
- [24] M. Liao, X. Shao, J. Chai, X. Sun, X. Ke, H. Xu, J. Xiang, X. Wang, and W. Wang, "Investigation of reliability characteristics of Hf_xZr_{1-x}O₂-based FeFET and AFeFET non-volatile memory," in *2024 IEEE 17th International Conference on Solid-State & Integrated Circuit Technology (ICSICT)*. IEEE, 2024, pp. 1–3.
- [25] Y. Zhao, Y. Zhao, C. Wan, and Y. Lin, "3d-carbon: An analytical carbon modeling tool for 3d and 2.5 d integrated circuits," in *Proceedings of the 61st ACM/IEEE Design Automation Conference*, 2024, pp. 1–6.
- [26] C. C. Sudarshan, N. Matkar, S. Vrudhula, S. S. Sapatnekar, and V. A. Chhabria, "Eco-chip: Estimation of carbon footprint of chiplet-based architectures for sustainable vlsi," in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2024, pp. 671–685.
- [27] F. Fayza, C. Demirkiran, S. P. Rao, D. Bunandar, U. Gupta, and A. Joshi, "Epicarbon: A carbon modeling tool for electro-photonics accelerators," in *2025 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*. IEEE, 2025, pp. 1–9.
- [28] D. Grey-Stewart, D. Kong, M. Elgamal, G. Kyriazidis, J. Morris, and G. Hills, "Quantifying trade-offs in power, performance, area, and total carbon footprint of future three-dimensional integrated computing systems," in *2025 Design, Automation & Test in Europe Conference (DATE)*. IEEE, 2025, pp. 1–7.
- [29] D. Pan, D. Guan, T.-C. Jen, and C. Yuan, "Atomic layer deposition process modeling and experimental investigation for sustainable manufacturing of nano thin films," *Journal of Manufacturing Science and Engineering*, vol. 138, no. 10, p. 101010, 2016.
- [30] X. Chen, L. Han, A. Bhagavathula, and U. Gupta, "Carbonclarity: Understanding and addressing uncertainty in embodied carbon for sustainable computing," 10 2025, pp. 1–9.
- [31] C. Y. Yuan and D. Dornfeld, "Environmental performance characterization of atomic layer deposition," in *2008 IEEE International Symposium on Electronics and the Environment*. IEEE, 2008, pp. 1–6.
- [32] R. Pachauri, M. Allen, V. Barros, J. Broome, W. Cramer, R. Christ, J. Church, L. Clarke, Q. Dahe, P. Dasgupta *et al.*, "Fifth assessment report of the intergovernmental panel on climate change," *IPCC Climate Change*, 2014.
- [33] I. P. on Climate Change (IPCC), "2019 refinement to the 2006 ipcc guidelines for national greenhouse gas inventories," p. 824, 2019.
- [34] J. Meza, Q. Wu, S. Kumar, and O. Mutlu, "A large-scale study of flash memory failures in the field," *ACM SIGMETRICS Performance Evaluation Review*, vol. 43, no. 1, pp. 177–190, 2015.
- [35] K. Seshadri, B. Akin, J. Laudon, R. Narayanaswami, and A. Yazdanbakhsh, "An evaluation of edge tpu accelerators for convolutional neural networks," 2022. [Online]. Available: <https://arxiv.org/abs/2102.10423>
- [36] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: <https://arxiv.org/abs/1704.04861>
- [37] A. Samajdar, J. M. Joseph, Y. Zhu, P. Whatmough, M. Mattina, and T. Krishna, "A systematic methodology for characterizing scalability of dnn accelerators using scale-sim," in *2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2020, pp. 58–68.
- [38] H. J. Byun, U. Gupta, and J.-s. Seo, "3d ic architecture evaluation and optimization with digital compute-in-memory designs," in *Proceedings of the 29th ACM/IEEE International Symposium on Low Power Electronics and Design*, ser. ISLPED '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1–6. [Online]. Available: <https://doi.org/10.1145/3665314.3670838>