Thermal Effects on Monolithic 3D Ferroelectric Transistors for Deep Neural Networks Performance

*Shubham Kumar, Yogesh Singh Chauhan, and Hussam Amrouch**

Monolithic three-dimensional (M3D) integration advances integrated circuits by enhancing density and energy efficiency. Ferroelectric thin-film transistors (Fe-TFTs) attract attention for neuromorphic computing and back-end-of-the-line (BEOL) compatibility. However, M3D faces challenges like increased runtime temperatures due to limited heat dissipation, impacting system reliability. This work demonstrates the effect of temperature impact on single-gate (SG) Fe-TFT reliability. SG Fe-TFTs have limitations such as read-disturbance and small memory windows, constraining their use. To mitigate these, dual-gate (DG) Fe-TFTs are modeled using technology computer-aided design, comparing their performance. Compute-in-memory (CIM) architectures with SG and DG Fe-TFTs are investigated for deep neural networks (DNN) accelerators, revealing heat's detrimental effect on reliability and inference accuracy. DG Fe-TFTs exhibit about 4.6x higher throughput than SG Fe-TFTs. Additionally, thermal effects within the simulated M3D architecture are analyzed, noting reduced DNN accuracy to 81.11% and 67.85% for SG and DG Fe-TFTs, respectively. Furthermore, various cooling methods and their impact on CIM system temperature are demonstrated, offering insights for efficient thermal management strategies.

1. Introduction

The growing need for applications with high memory requirements, especially in the field of neuromorphic computing, has highlighted the necessity for more advanced compute-in-memory (CIM) architectures.[1] These architectures aim to minimize data transfer between processing elements and memory blocks by using non-volatile memories for CIM, addressing the limitations of the von Neumann bottleneck.[2,3] As deep neural networks (DNNs) continue to require extensive memory capacities, using novel CIM architectures implemented in the back-end-of-the-line (BEOL) fabrication has become essential. This is due to the high density achieved by BEOL-integrated monolithic three-dimensional (M3D) architectures, which enables efficient storage and processing of data.[4]

In M3D architecture for CIM, multiple memory tiers are integrated monolithically on top of the high-performance complementary metal-oxide-semiconductor (CMOS) logic, enabling a significant boost in memory density.[5] The bottom-tier circuits at the front-end-of-line (FEOL) can be fully or partially self-aligned through the employment of the M3D integration technique.[6] This results in denser integrated circuits and closer integration between different circuits, ultimately leading to improved efficiency.[7]

Despite its potential, M3D integration faces significant processing challenges. M3D integration requires the fabrication of BEOL upper-tier circuits at a relatively low temperature (below 400 °C) to prevent damage to the metal interconnects and existing bottom-tier circuits caused by elevated temperatures.[6,8] The thermal activation of dopants in silicon CMOS technologies typically occurs between 600 and 1000 °C for reliable device performance.[9] Recent literature has proposed successful integration for amorphous oxide semiconductors (AOS), such as indium tungsten oxide (IWO) and indium gallium zinc oxide (IGZO), along with ferroelectric (FE) hafnium zirconium oxide (HZO), using fabrication steps compatible with BEOL processes.[10–12]

Among emerging non-volatile memories, the ferroelectric thin-film transistor (Fe-TFT) used as a BEOL transistor stands out due to its excellent compatibility with CMOS technology, scalability, and high ON/OFF ratio.[10,12,13] Complex systems like DNN implemented on BEOL Fe-TFT M3D architectures (as illustrated in **Figure 1**a) face challenges due to the extensive operations involving multiple analog-to-digital converters (ADC), control circuits, and the inability to dissipate heat leads to increased operational temperature. This temperature rise, exceeding 100 °C in M3D CIM systems can affect device characteristics and result in incorrect read-out, reducing inference

S. Kumar, Y. S. Chauhan
Department of Electrical Engineering
Indian Institute of Technology
Kanpur, UP 208016, India

S. Kumar
Semiconductor Test and Reliability (STAR)
University of Stuttgart
70550 Stuttgart, Germany

H. Amrouch
Chair of AI Processor Design
TUM School of Computation
Information and Technology
Munich Institute of Robotics and Machine Intelligence
Technical University of Munich
80333 Munich, Germany
E-mail: amrouch@tum.de

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202400019.

DOI: 10.1002/aisy.202400019

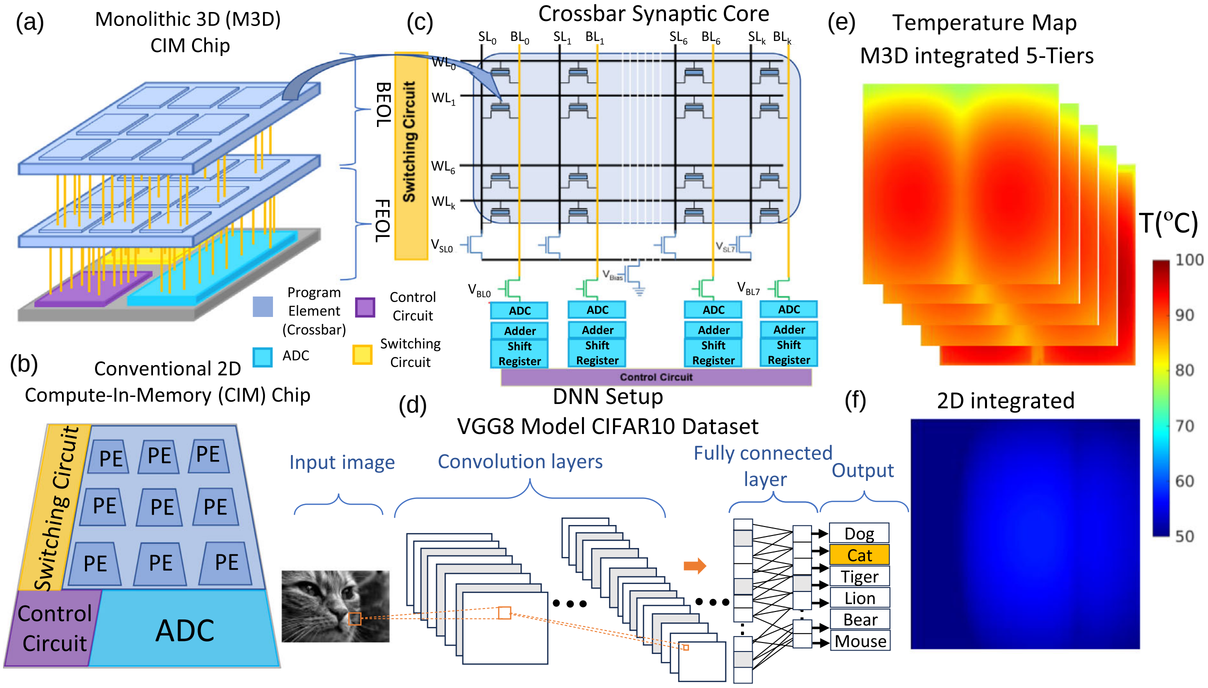


Figure 1. a) Schematic of monolithic 3D (M3D) integrated compute-in-memory (CIM) architecture having high density embedded non-volatile memory in BEOL and the peripheral circuits in FEOL, whereas b) conventional 2D CIM architecture has memory and peripheral circuits in FEOL. c) The crossbar of the synaptic core with other peripheral circuits for d) the deep neural network in the VGG8 network on the CIFAR10 dataset. e) Temperature maps of the 5-tier M3D integrated architecture compared with f) the 2D integrated architecture.

accuracy. These challenges are primarily driven by the high power density, resulting in elevated on-chip temperatures compared to traditional monolithic 2D (M2D) architectures[14] (Figure 1b).

To leverage the benefits of 3D integration for CIM/accelerator hardware (Figure 1c), the rising trend in power densities brings significant challenges related to thermal effects. These challenges include inter-die thermal coupling and an increased occurrence of hotspots Figure 1e,f, which can have implications for performance and reliability. This arises from the variations in power densities within 3D integrated circuits compared to their monolithic 2D counterparts, with thermal performance not necessarily scaling linearly. Although advanced cooling methods like liquid cooling exist, their implementation introduces complexities and additional costs.[15] Liquid cooling devices are efficient heat exchangers but require more power than air-cooled heat sinks. Liquid cooling surpasses air cooling in thermal resistance only when adequately powered. Our work demonstrates the BEOL-integrated ferroelectric-based CIM architecture for DNN applications. Since the ferroelectric transistors are the most promising non-volatile memory, analyzing the implication of ferroelectric-based CIM for M3D integration is necessary.

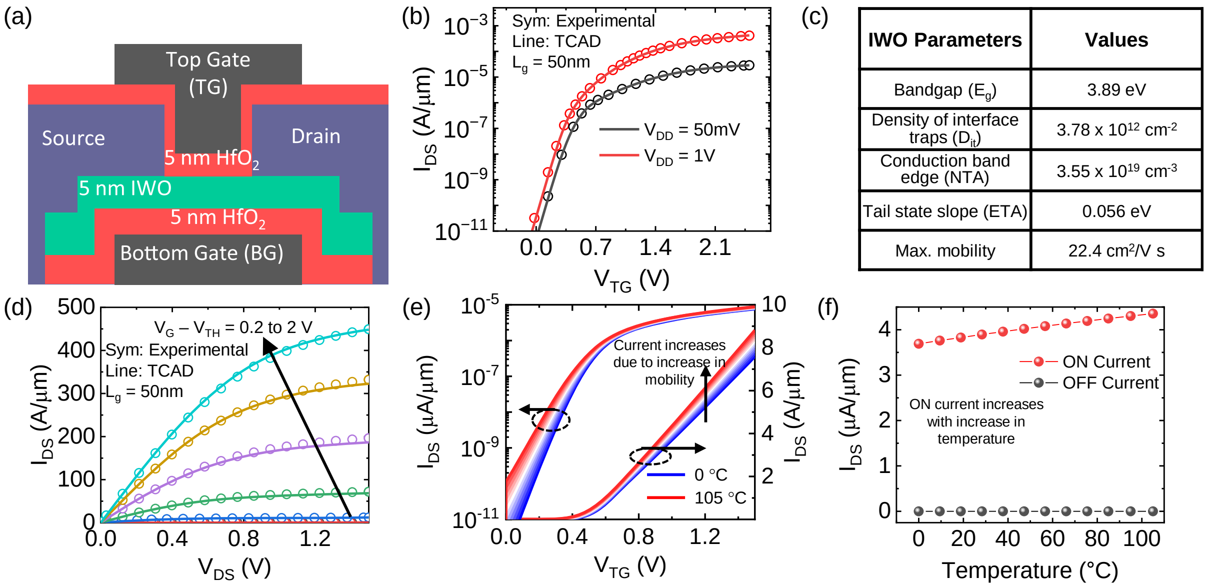
Our key contributions are as follows: 1) We demonstrate the BEOL-integrated Fe-TFT-based CIM architecture for DNN applications and present an extensive analysis aimed at quantifying the influence of the integration design in terms of run-time and design-time variability. 2) We examine how the properties of the device are affected at high temperatures and its influence on the accuracy of the DNN system. 3) We propose the most effective way for optimal performance even under high-temperature conditions. 4) We demonstrate the impact of various cooling methods on the temperature elevation within the CIM system, showcasing their effects and implications.

2. Electrical Characteristics of the Transistor

2.1. Device Calibration of Thin-Film Transistor

Figure 2a shows the simulated dual-gate (DG) TFT structure in Sentaurus technology computer-aided design (TCAD) with 50 nm channel length and 5 nm thick IWO channel. Figure 2b demonstrates the calibration of the experimentally measured[10] drain current (I_{DS}) as a function of top gate voltage (V_{TG}) characteristics for two different V_{DD}, 50 mV and 1 V. The calibration captures important features like subthreshold swing (SS), ON current, and OFF current of TFT. To accurately represent the density-of-states in the IWO channel, we employed an exponential band-tail density-of-states model and a Gaussian distribution of traps.[16] The bulk trap Coulomb scattering (BTCS) non-local mobility model and non-local tunneling for the drain and source were incorporated to account for trap-limited conduction. To ensure the reliability and accuracy of our TCAD simulation, we used density-of-states parameters of IWO from.[17] Figure 2c shows the parameter list of the IWO channel.

Additionally, we have calibrated I_{DS}–V_{DS} curve for different values of V_{TG} with experimental data[10] in TCAD as shown in Figure 2d. Subsequently, we examined the impact of temperature on the TFT device as shown in Figure 2e. The ON current



IWO Parameters	Values
Bandgap (E$_g$)	3.89 eV
Density of interface traps (D$_{it}$)	3.78 x 10^{12} cm^{-2}
Conduction band edge (NTA)	3.55 x 10^{19} cm^{-3}
Tail state slope (ETA)	0.056 eV
Max. mobility	22.4 cm^2/V s

Figure 2. a) Schematic of dual-gate thin-film transistor (TFT) with 5 nm thick IWO channel and channel length of 50 nm. b) I_{DS}–V_{TG} TCAD calibration with the measured data[10] for V_{DD} of 50 mV and 1 V. c) List of the density of states parameters within the bandgap of IWO channel. d) I_{DS}–V_{DS} at different V_{TG} TCAD calibration with measured data.[10] e) The impact of temperature on the I_{DS}–V_{TG} characteristic. The ON current increases with temperature due to an increase in mobility. f) The plot of ON and OFF current of TFT with different temperatures.

increases (Figure 2f) with rising temperature, contrary to silicon transistors, due to enhanced mobility resulting from trap-limited percolation-dominated carrier conduction.[17]

2.2. Ferroelectric Transistor

Figure 3a shows the simulated DG Fe-TFT structure in Sentaurus TCAD with 50 nm channel length and 5 nm thick IWO channel.[18,19] The Preisach model captures the FE characteristics. The change in FE polarization (P) as a function of the applied voltage (V) is shown in Figure 3b. The hysteresis loop of the P–V curve is calibrated against a fabricated metal-ferroelectric-metal (MFM) capacitor.[20] FE parameters such as remanent polarization (P_r) = 22.8 μC cm^{-2}, saturation polarization (P_s) = 32.5 μC cm^{-2}, and coercive field (E_c) = 1.76 MV cm^{-1} are calculated from the P–V hysteresis loop and listed in Figure 3c.

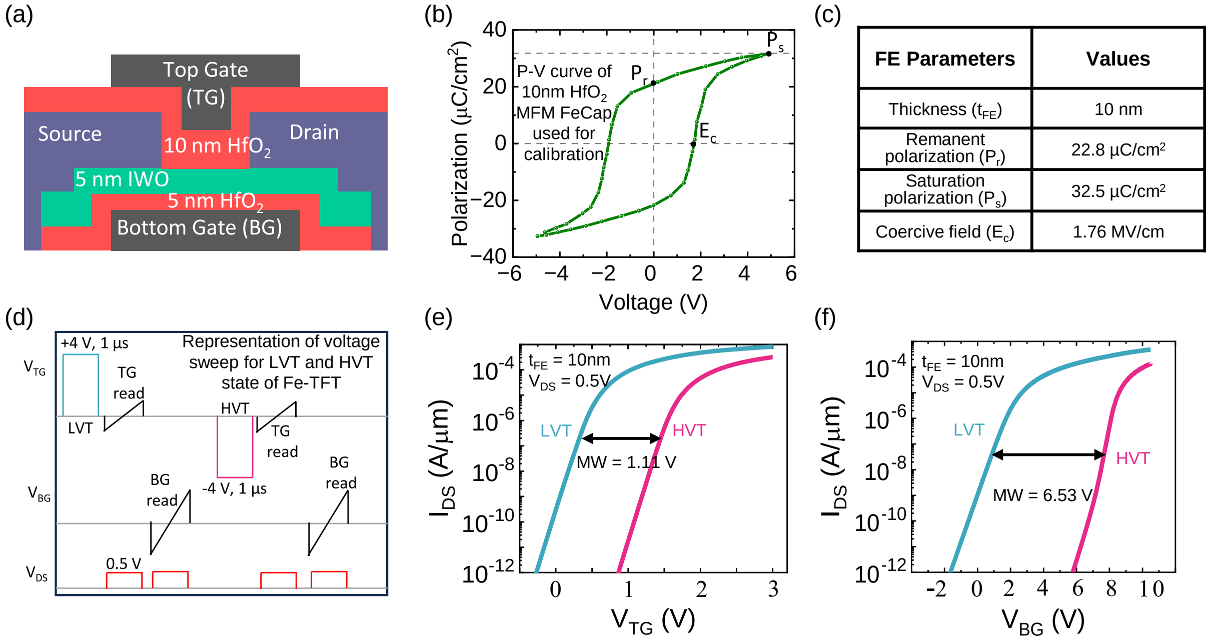
For the Fe-TFT, the distinct memory state is obtained by applying a voltage pulse at the top gate (TG) and then sweeping the voltage to read the state as shown in Figure 3d. The memory state in Fe-TFT is characterized as different V_{th}. The low V_{th} (LVT) is obtained by applying a voltage pulse of 4 V, 1 μs at TG, whereas high V_{th} (HVT) is obtained by applying a voltage pulse of −4 V, 1 μs at TG. To read the memory state TG is swept and I_{DS}–V_{TG} curve is obtained as shown in Figure 3e. The MW of the Fe-TFT is calculated as the difference between HVT and LVT. For the TG read method, an MW of 1.11 V is achieved. The conventional TG read of Fe-TFT cannot have a high MW at FE thickness t_{FE} = 10 nm[21] and can be approximated as:

$$\mathrm{MW} = 2 \cdot \gamma \cdot E_c \cdot t_{FE} \tag{1}$$

where γ is the ideality factor for the FE that accounts for second-order effects and is less than 1. Considering the typical values for FE films are E_c = 1 MV cm^{-1} and t_{FE} = 10 nm, the maximum MW of the Fe-TFT is 2 V. A thick FE layer can increase the MW since the MW is directly proportional to t_{FE}. But, FE properties degrade for thick t_{FE} and also hinder device scaling. Further, applying both read and write voltages at the same TG terminal can flip the polarization of the FE layer during the read operation and cause read disturbance. To overcome these issues, we have used DG Fe-TFT in which the write voltage is applied at the TG and read voltage at bottom gate (BG) has been proposed.[3,21] This provides a disturb-free read due to separate read and write terminals and also amplifies the MW. Figure 3f shows the I_{DS} as a function of V_{BG} of Fe-TFT. The MW of 6.53 V is obtained for the BG read. Large MW for the case of BG read is due to the large coupling between the BG and TG.[22] The BG method to read Fe-TFT states presents certain difficulties. In a recent study,[22] it was shown that BG read amplifies transistor variability. This increased variability can disrupt the functionality of integrated circuits and systems.

3. Fe-TFT-Based M3D CIM Configuration

We have used Fe-TFT to build the CIM architecture and compare the two types of read schemes. A comprehensive system-level framework is employed to evaluate the hardware performance and inference accuracy of the CIM accelerator. This accelerator incorporates BEOL Fe-TFT integrated within a M3D design, comprising multiple tiers of memory and logic circuits. This study demonstrates the temperature rise occurring in each tier following the inference operation. By analyzing the temperature variations, we assess how the increased temperature affects the performance and reliability of the BEOL Fe-TFT. Subsequently,



FE Parameters	Values
Thickness (t_{FE})	10 nm
Remanent polarization (P_r)	22.8 μC/cm^2
Saturation polarization (P_s)	32.5 μC/cm^2
Coercive field (E_c)	1.76 MV/cm

Figure 3. a) Schematic of Fe-TFT made by stacking of 10 nm ferroelectric layer in the top gate oxide of TFT structure. b) Polarization in the ferroelectric layer as a function of Voltage hysteresis loop for ferroelectric model parameter calibration with the measurement data[20] using 10 nm ferroelectric layer in the metal ferroelectric metal capacitor. c) The list of ferroelectric model parameters. d) The waveform of the pulse scheme at each terminal of Fe-TFT used to read and write the states. To write the memory state, we have applied +4 V/−4 V, 1 μs to the TG of Fe-TFT to set them in LVT/HVT. We have used two types of read methods, TG and BG read. In the case of TG read, a read voltage is applied at the TG terminal keeping the BG terminal at 0 V while for the BG read, we apply the read voltage at the BG terminal keeping TG at 0 V. e) For the I_{DS}–V_{TG} characteristic, a memory window of 1.11 V is obtained, f) while for the case of BG read, a memory window of 6.53 V is obtained from I_{DS}–V_{BG} characteristic.

we examine the impact of device degradation on the inference accuracy of CIM accelerator-based DNN.

We have used the simulation flow depicted in **Figure 4** to quantify 1) the inference accuracy of the M3D CIM system, 2) the elevated temperature of each tier stack, and 3) the thermal effects of various M3D integrated BEOL Fe-TFT design parameters on CIM inference accuracy. This simulation flow combines the CIM inference accuracy estimation framework (DNN + NeuroSim[23]) with a finite volume method (FVM)-based thermal modeling framework.[24]

To evaluate the hardware performance and inference accuracy on a large-scale model, we have applied the M3D design of CIM

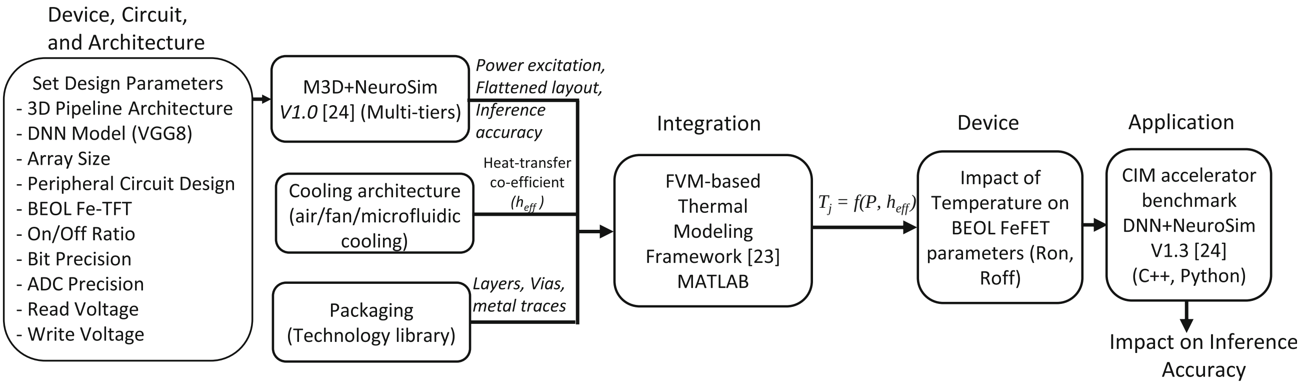

Device, Circuit, and Architecture
Set Design Parameters
- 3D Pipeline Architecture
- DNN Model (VGG8)
- Array Size
- Peripheral Circuit Design
- BEOL Fe-TFT
- On/Off Ratio
- Bit Precision
- ADC Precision
- Read Voltage
- Write Voltage
M3D+NeuroSim V1.0 [24] (Multi-tiers)
Power excitation, Flattened layout, Inference accuracy
Cooling architecture (air/fan/microfluidic cooling)
Heat-transfer co-efficient (h_{eff})
Packaging (Technology library)
Layers, Vias, metal traces
Integration
FVM-based Thermal Modeling Framework [23] MATLAB
$T_j = f(P, h_{eff})$
Device
Impact of Temperature on BEOL FeFET parameters (Ron, Roff)
Application
CIM accelerator benchmark DNN+NeuroSim V1.3 [24] (C++, Python)
Impact on Inference Accuracy

Figure 4. The assessment of inference accuracy within the thermal-driven M3D CIM-based framework follows a systematic simulation procedure. Initially, we input essential data, including the architectural configuration of the network (pipeline[25]), the flattened layout of each tier, device-specific parameters, the cooling architecture, and the technological node specifications. These inputs enable the computation of temperature distributions at a steady state for each tier using the thermal framework.[24] Subsequently, we proceed to evaluate the influence of temperature on device performance while considering variations in device parameters arising from temperature fluctuations. Following this analysis, we proceed to quantify the loss in accuracy within the neural network due to elevated temperatures.

Table 1. Parameter list of Beol Ferroelectric thin film Transistor and simulation Setup.

Parameters	TG read Fe-TFT	BG read Fe-TFT
Technology node (Logic) [nm]	7	7
Technology node (Memory) [nm]	22	22
Before Thermal Analysis (Simulation at room temperature)		
R_{ON} [Ω]	1.96E3	8.19E3
ON/OFF ratio	4.71E1	2.68E4
Inference accuracy	91.44%	91.92%
After Thermal Analysis		
Max. Temperature [°C]	84.74	85.24
R_{ON} [Ω]	4.76E2	1.55e3
ON/OFF ratio	1.63E1	1.33E3
Resistance drift [%]	25.2	27.1
Inference accuracy [%]	81.11	67.85

accelerators to the VGG8 network using the CIFAR10 dataset. The subarray size is set to 128 × 128 and is constructed using 1-bit/cell Fe-TFT, with the parameters listed in **Table 1**. Partial sums within the 128 × 128 synaptic arrays are linearly quantized using a 5-bit ADC. From an architectural perspective, we consider a pipelined (PP) system with 3D interleaved logic and memory tiers.[25] This architecture offers high speed but consumes high power, providing a suitable framework for our analysis.

Using the open-source tool 3D CIM thermal v1.0,[24] we conduct steady-state thermal analysis for each tier of the architecture. For steady-state thermal analysis, we employ the flattened layout of the modeled CIM configuration, including memory arrays ADCs, Global buffers, accumulators, shift and add, activation and pooling circuits. A flattened layout implies that the hierarchical structure of the CIM, which consists of multiple layers or tiers, is collapsed into a simplified 2D representation for modeling and simulation purposes. This representation facilitates the thermal analysis of the CIM within the framework. Additionally, we incorporate power excitation maps of each active tier based on the flattened layout and a description of the die stack-up, encompassing bulk material, interconnects, and dielectrics, along with their respective thermal properties, such as thermal conductivity and specific heat capacity. To account for different cooling architectures, each with varying effective heat-transfer coefficients (h_{eff}) and assumptions regarding tier-to-tier interconnections (vias, I/Os) for the assumed PP architecture are summarized in **Table 2**. Our analysis employs interlayer vias (ILVs) for tier-to-tier interconnections, assuming a diameter of 0.1 μm.[26] The tier-to-tier I/O bonding pitch is assumed to be 0.1 μm, with a bond height of 0.5× the bonding pitch.

Table 2. Interconnect assumptions.

Attribute	M3D
ILV diameter [μm]	0.1
Number of vertical vias between two tiers	5.9×10^6
ILV total area [mm^2]	0.24
bonding pitch [μm]	0.1
Bonding layer thickness [μm]	0.05

All these parameters serve as inputs to the FVM-based thermal framework. The thermal model incorporates three primary input parameters: 1) Power consumption of individual functional blocks within the chip. 2) Geometric details of the M3D stack, including the dimensions of its constituent elements. 3) Material properties relevant to the stack components.[24] This algorithm discretizes the entire layout of the stack and deduces each non-zero element within the h_{eff} matrix, which exhibits both sparsity and symmetry characteristics. The relationship between matrix h_{eff}, power consumption vector P, and the junction temperature vector T_j is as in Equation (2):

$$T_j = P/h_{\mathrm{eff}} \tag{2}$$

The output of the FVM-based thermal framework is the maximum junction temperature ($T_{j,\max}$) for each tier. The elevated temperature impacts various Fe-TFT parameters, such as the R_{ON} value and the ON/OFF ratio (Table 1). Consequently, we calculate the drift in the R_{ON} value caused by the increased temperature and assess its impact on the inference accuracy of the network.

4. Results and Discussion

Our study considers three M3D partitioning configurations: 1) two-tiers logic-on-memory (L-M), 2) three-tiers (L-M-L), and 3) five-tiers (L-M-L-M-L). These configurations are then compared with the M2D configuration. To assess the hardware performance, we employ two methods to read the stored states in the Fe-TFT: TG and BG read. We analyze various hardware estimation metrics such as throughput (TOPS), energy efficiency (TOPS/W), compute efficiency (TOPS/mm^2), and chip area (**Figure 5**). As we move from M2D to higher-tier M3D configurations, we observe a reduction in chip area, resulting in improved TOPS/mm^2. The results demonstrate that Fe-TFT devices using BG read exhibit higher performance due to their larger R_{ON} values and ON/OFF ratio.

The power consumption per block, the number of blocks, as well as the number of memory components (memory array and switch matrix), and peripheral logic elements (shift-add, ADC, accumulation, activation, pooling, and global buffer) for the VGG8 network are calculated directly using NeuroSim. To further analyze the thermal impact, we have conducted a comprehensive thermal analysis that takes into account the actual power distribution across different locations on the chip floorplan. The analysis enables us to generate temperature contours, as depicted in **Figure 6**. For the M3D two-tier configuration, we present the block-based power density distribution and the floorplans for each logic and memory tier as shown in Figure 6a,b, respectively. In our assumptions, we place the ILVs in the middle and evenly distribute the synaptic memory arrays across the memory tiers. In the logic tiers, we group the ADCs and other peripheral circuits used for one synaptic memory array into a single logic block. These logic blocks are evenly placed in the same manner

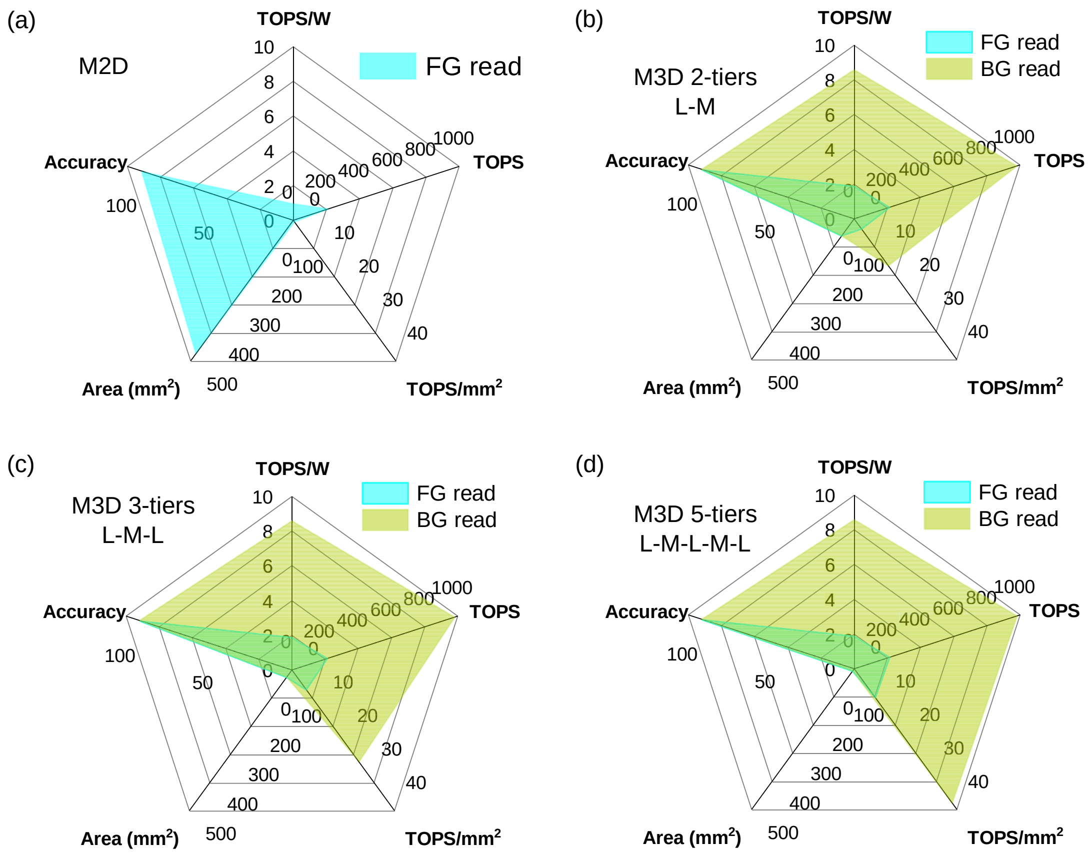


Figure 5. The evaluation of the performance and hardware estimation metrics for various architectural configurations a) M2D architecture, including metrics such as chip area, inference accuracy, TOPS, TOP/W, and TOPS/mm^2. b) M3D architecture with 2 tiers (memory on logic), considering TG and BG read BEOL Fe-TFT devices. c) M3D architecture with 3 tiers, specifically configured as L-M-L (Logic-Memory-Logic). d) M3D architecture with 5 tiers, configured as L-M-L-M-L (Logic-Memory-Logic-Memory-Logic).

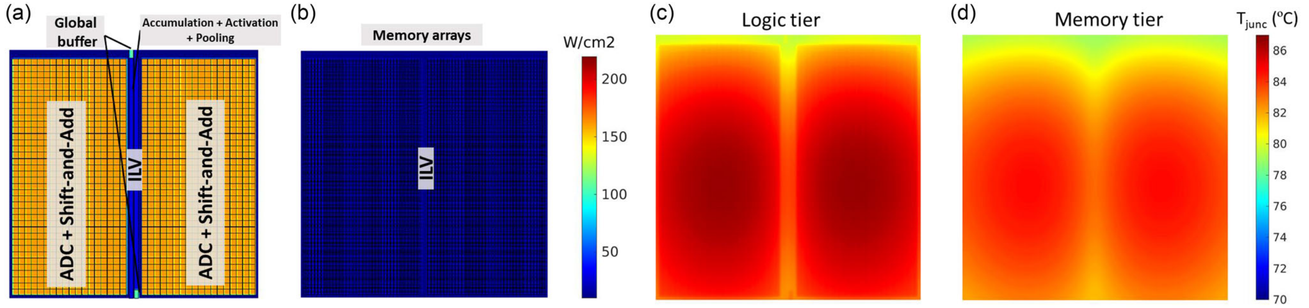


Figure 6. The floorplans and block-based power densities for a two-tier architecture, a) the logic tier, and b) the M3D memory tier. c) The steady-state tier junction temperature contours for the two-tier M3D logic tier and d) the memory tier. These steady-state tier junction temperature contours for each tier are calculated using the thermal framework.[24]

as the memory arrays. The global buffers and other blocks, such as accumulation, activation, and pooling circuits, are located in the middle along with the ILVs to facilitate intermediate data processing. Regarding data transfer, we assume that data is exchanged among logic blocks and synaptic memory arrays through H-tree interconnects within the logic and memory tiers, respectively.

The temperature contours obtained from the FVM-based thermal modeling framework, considering air-cooling with h_{eff} of 4.4×10^3 W K^{-1} m^2,[27,28] are presented for both the logic and memory tiers in Figure 6c,d, respectively. It is worth noting that the relative temperatures among all tiers within the M3D configurations are similar. This similarity arises from the low thickness of each tier, resulting in a low interior thermal resistance.

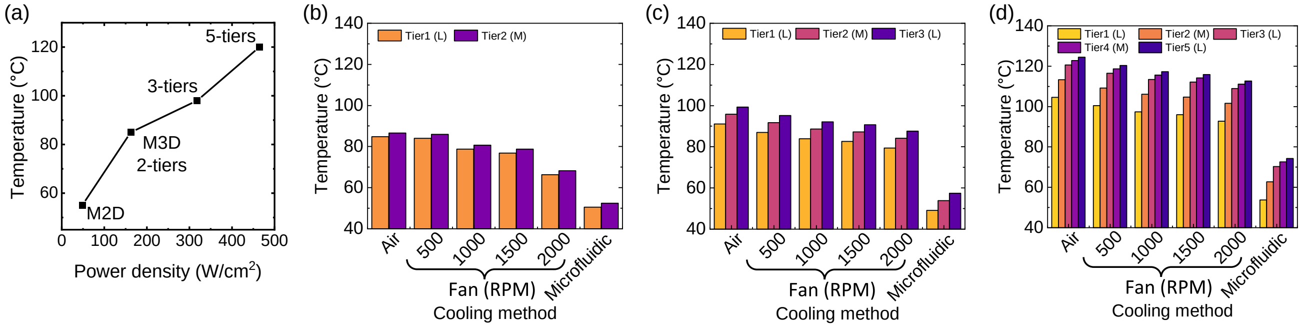


Figure 7. a) The Investigation of chip temperature as a function of power density variations among different CIM architecture types, including M2D, and M3D with varying numbers of tiers. As the power density increases from M2D to higher-tier M3D architectures, there is a corresponding increase in temperature within the tiers. b) The assessment of the maximum temperature rise for different cooling methodologies applied to the two-tier M3D CIM setup, as well as extending to c) three-tier M3D and d) five-tier M3D architecture. It is noteworthy that microfluidic cooling effectively mitigates temperature rise within the chip compared to air-based cooling; however, it incurs a linear increase in cooling costs.

The average power density in the PP system is relatively high, consumes significant power, and operates at high temperatures. Additionally, the power density further increases with a smaller ILV diameter. **Figure 7**a illustrates the temperature rise of the chip as the power density increases. In the case of the M2D configuration, the larger chip area leads to a lower power density. However, as we move to higher-tier M3D configurations, the power density increases due to the smaller chip area, resulting in more concentrated heat dissipation.

4.1. Impact of Temperature on M3D CIM Configuration

The comprehensive thermal framework calculates a maximum temperature rise of 85 °C for the two-tier M3D CIM configuration. To assess the impact of temperature on the BEOL Fe-TFT, we examine the drift in the R_{ON} value and the ON/OFF ratio, as detailed in Table 1. During our evaluation, the memory states are programmed at room temperature, maintaining the FE parameters constant, and subsequently read out at higher temperatures. Since the memory states are programmed at room temperature, i.e., FE parameters remain unchanged.[29] However, there will be a shift in the threshold voltage of the transistor due to the substrate effect and its dependence on temperature influences the basic FET. This shift in the threshold voltage induces drift in the R_{ON} and ON/OFF ratio of the transistor. As the temperature increases, the current for both LVT and HVT states shows an increasing trend, primarily due to the characteristics of the amorphous oxide material channel.[17]

The drift in the R_{ON} due to the increased temperature has more impact on the BG read scheme due to high variation for the case of BG read Fe-TFT.[22] The variation parameter of the Fe-TFT for both TG and BG read is introduced in the NeuroSim as a percentage of variation of desired resistance.[23] We re-run the framework using the updated Fe-TFT parameters and observe a loss in the inference accuracy for the same network.[23] Specifically, the inference accuracy reduces to 81.11% for the TG read and 67.85% for the BG read case, in contrast to the accuracy achieved at room temperature, which was approximately 91%. This decline in accuracy can be attributed to the variations caused by temperature changes, leading to incorrect multiply-accumulate (MAC) outputs in relation to fixed ADC references used during the inference operation. It is important to note that although the BG read scheme demonstrates superior performance compared to the TG read scheme, it also experiences a more significant loss in inference accuracy due to the elevated temperature when incorporated into the M3D integrated configuration.

4.2. Thermal Resiliency for M3D Chips

The temperature of the metal lines results from a combination of heating within the BEOL stack and heating within the FEOL. The degree of heat transfer between FEOL and BEOL is closely tied to the specific packaging and cooling solution employed for the chip, making it highly dependent on the intended application. In our framework, we have explored various cooling methods ranging from air cooling to fan cooling to liquid cooling. The h_{eff} for each cooling type is provided in **Table 3**.[30] Figure 7b–d illustrate the maximum temperature rise experienced in each tier of the CIM system while performing computations under different cooling methods. We observe that higher values of h_{eff} result in lower temperature rise within each tier. In other words, improved cooling capabilities lead to better thermal management. However, it is essential to consider that the cooling cost also increases significantly as we move from air cooling to liquid cooling. A liquid cooling system functions as a heat exchanger, capable of effectively dissipating a significant amount of heat. However, it typically demands a higher power input compared to air-cooled heat sinks.

Table 3. Parameters for different cooling methods.[30]

Cooling type	h_{eff}	R_{th}
Air cooled	4.4×10^3	2.25
Fan cooled (500 RPM)	8.33×10^3	1.2
Fan cooled (1000 RPM)	16.67×10^3	0.6
Fan cooled (1500 RPM)	25×10^3	0.4
Fan cooled (2000 RPM)	100×10^3	0.1
Microfluidic	333.33×10^3	0.03

Achieving a lower thermal resistance results in superior cooling efficiency, but it necessitates increased power consumption for the cooling process. Liquid cooling systems can attain a lower thermal resistance when adequately powered, but when supplied with an equal amount of power, they exhibit higher thermal resistance compared to air-cooled heat sinks. Therefore, there will be tradeoffs between the cost of cooling and the performance of the CIM-based DNN system. There are alternative approaches to mitigate high junction temperatures such as thermal-aware design-time partitioning[31] and introducing a temperature-sensing and bias-adaptive solution to minimize the device degradation and maintain the accuracy of DNN.[32]

5. Conclusion

In conclusion, our extensive thermal analysis has provided valuable insights into the integration design of a CIM-based BEOL-Fe-TFT for DNN applications. By examining the influence of device properties on system accuracy under varying temperature conditions, we aimed to identify optimal performance strategies, particularly in the context of M3D integrated systems. Our analysis considered both TG and BG read schemes for Fe-TFT devices, with the BG read demonstrating higher performance due to larger R_{ON} values and ON/OFF ratio. However, we noted that the BG read scheme experienced a more significant inference accuracy loss due to temperature variations when integrated into the M3D configuration. Furthermore, we highlighted the impact of various cooling methods on the temperature elevation within the CIM system, showcasing their effects and implications on thermal management. The thermal analysis of the VGG8 network revealed the effectiveness of different cooling methods in managing temperature rise, but tradeoffs between cooling cost and performance must be considered.

Acknowledgements

The authors would like to thank Ankit Kaul, Madison Manley, and Muhannad S. Bakir from the Georgia Institute of Technology for their valuable help in thermal modeling and for providing their framework for thermal simulations. The authors also thank Om Prakash and Swetaki Chatterjee for their valuable support and discussions regarding the modeling of the thin-film transistors.

Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Keywords

BEOL, deep neural network, ferroelectric, monolithic 3D, thin-film transistor

Received: January 6, 2024
Revised: May 5, 2024
Published online: July 3, 2024

[1] D. Marković, A. Mizrahi, D. Querlioz, J. Grollier, *Nat. Rev. Phys.* **2020**, *2*, 499.

[2] S. Kumar, S. Chatterjee, S. Thomann, P. R. Genssler, Y. S. Chauhan, H. Amrouch, in *2022 IFIP/IEEE 30th Int. Conf. Very Large Scale Integration (VLSI-SoC)*, IEEE, Piscataway, NJ **2022**, pp. 1–6.

[3] S. Kumar, S. Chatterjee, S. Thomann, Y. S. Chauhan, H. Amrouch, *IEEE Trans. Circuits Syst. I: Regul. Pap.* **2023**, *70*, 2891.

[4] S. Dutta, H. Ye, W. Chakraborty, Y.-C. Luo, M. San Jose, B. Grisafe, A. Khanna, I. Lightcap, S. Shinde, S. Yu, S. Datta, in *2020 IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2020**, pp. 36.4.1–36.4.4.

[5] S. Salahuddin, K. Ni, S. Datta, *Nat. Electron.* **2018**, *1*, 442.

[6] S. Datta, S. Dutta, B. Grisafe, J. Smith, S. Srinivasa, H. Ye, *IEEE Micro* **2019**, *39*, 8.

[7] M. M. Shulaker, T. F. Wu, M. M. Sabry, H. Wei, H.-S. P. Wong, S. Mitra, in *2015 Design, Automation & Test in Europe Conf. & Exhibition (DATE)*, IEEE, Piscataway, NJ **2015**, pp. 1197–1202.

[8] A. Vandooren, L. Witters, J. Franco, A. Mallik, B. Parvais, Z. Wu, A. Walke, V. Deshpande, E. Rosseel, A. Hikavyy, W. Li, in *2018 Int. Conf. IC Design & Technology (ICICDT)*, IEEE, Piscataway, NJ **2018**, pp. 145–148.

[9] P. Batude, B. Sklenard, C. Fenouillet-Beranger, B. Previtali, C. Tabone, O. Rozeau, O. Billoint, O. Turkyilmaz, H. Sarhan, S. Thuries, G. Cibrario, in *IEEE Int. Interconnect Technology Conf.*, IEEE, Piscataway, NJ **2014**, pp. 373–376.

[10] H. Ye, J. Gomez, W. Chakraborty, S. Spetalnick, S. Dutta, K. Ni, A. Raychowdhury, S. Datta, in *2020 IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2020**, pp. 28.3.1–28.3.4.

[11] S. Dutta, H. Ye, A. A. Khandker, S. G. Kirtania, A. Khanna, K. Ni, S. Datta, *IEEE Electron Device Lett.* **2022**, *43*, 382.

[12] F. Mo, Y. Tagawa, C. Jin, M. Ahn, T. Saraya, T. Hiramoto, M. Kobayashi, in *2019 Symp. VLSI Technology*, IEEE, Piscataway, NJ **2019**, pp. T42–T43.

[13] T. Böscke, J. Müller, D. Bräuhaus, U. Schröder, U. Böttger, *Appl. Phys. Lett.* **2011**, *99*, 102903.

[14] P. Shukla, A. K. Coskun, V. F. Pavlidis, E. Salman, in *Proc. 2019 on Great Lakes Symp. VLSI*, ACM, Tysons Corner, VA **2019**, pp. 439–444.

[15] Y.-H. Gong, J. Kong, S. W. Chung, *IEEE Trans. Emerging Top. Comput.* **2019**, *9*, 854.

[16] P. Zhang, S. Samanta, X. Fong, *IEEE Trans. Electron Devices* **2020**, *67*, 2352.

[17] W. Chakraborty, H. Ye, B. Grisafe, I. Lightcap, S. Datta, *IEEE Trans. Electron Devices* **2020**, *67*, 5336.

[18] S. Kumar, S. Thomann, O. Prakash, K. Ni, Y. S. Chauhan, H. Amrouch, *IEEE Trans. Electron Devices* **2023**, *71*, 368.

[19] S. Kumar, O. Prakash, Y. S. Chauhan, H. Amrouch, *IEEE Trans. Electron Devices* **2023**, *70*, 6286.

[20] K. Ni, J. Smith, H. Ye, B. Grisafe, G. B. Rayner, A. Kummel, S. Datta, in *2019 IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2019**, pp. 28.8.1–28.8.4.

[21] H. Mulaosmanovic, D. Kleimaier, S. Dunkel, S. Beyer, T. Mikolajick, S. Slesazeck, *Nanoscale* **2021**, *13*, 16258.

[22] S. Chatterjee, S. Thomann, K. Ni, Y. S. Chauhan, H. Amrouch, *IEEE Trans. Electron Devices* **2022**, *69*, 5316.

[23] X. Peng, S. Huang, Y. Luo, X. Sun, S. Yu, in *2019 IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2019**, pp. 32–35.

[24] Y. Zhang, Y. Zhang, M. S. Bakir, *IEEE Trans. Compon., Packag., Manuf. Technol.* **2014**, *4*, 1914.

[25] X. Peng, A. Kaul, M. S. Bakir, S. Yu, *IEEE Trans. Electron Devices* **2021**, *68*, 5598.

[26] M. M. Shulaker, G. Hills, R. S. Park, R. T. Howe, K. Saraswat, H.-S. P. Wong, S. Mitra, *Nature* **2017**, *547*, 74.

[27] A. Kaul, S. K. Rajan, M. O. Hossen, G. S. May, M. S. Bakir, in *2020 IEEE 70th Electronic Components and Technology Conf. (ECTC)*, IEEE, Piscataway, NJ **2020**, pp. 1459–1467.

[28] A. Kaul, Y. Luo, X. Peng, M. Manley, Y.-C. Luo, S. Yu, M. S. Bakir, *IEEE Trans. Electron Devices* **2022**, *70*, 485.

[29] T. Ali, K. Kühnel, M. Czernohorsky, C. Mart, M. Rudolph, B. Pätzold, M. Lederer, R. Olivo, D. Lehninger, F. Müller, R. Hoffmann, *IEEE Trans. Electron Devices* **2020**, *67*, 2793.

[30] D. Shin, S. W. Chung, E.-Y. Chung, N. Chang, *IEEE Trans. Ind. Inf.* **2010**, *6*, 340.

[31] S. S. Kumar, A. Zjajo, R. van Leuken, *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **2017**, *25*, 1549.

[32] S. Chatterjee, S. Kumar, A. Sunil, S. De, D. Lehninger, M. Jank, T. Kämpfe, Y. S. Chauhan, H. Amrouch, in *2023 Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2023**, pp. 1–4.