Nanoscale Horizons

ROYAL SOCIETY OF CHEMISTRY

COMMUNICATION

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Nanoscale Horiz.*, 2026, 11, 909

Received 24th September 2025,
Accepted 6th January 2026

DOI: 10.1039/d5nh00653h

rsc.li/nanoscale-horizons

Surface-enhanced thermal dissipation in 3D vertical resistive memory arrays with top selector transistors

Arman Kadyrov,†[a] Seunghyun Lee,†[a] Batyrbek Alimkhanuly,[a] Shubham Patil,[a] Anupom Devnath,[a] Junseong Bae,[ac] Minwoo Lee,[ac] Jinsu Choi,[ac] Gisuk Hwang[b] and Seunghyun Lee *[ac]

Recent integration of 3D memory technologies such as high-bandwidth memory [HBM] into AI accelerators has enhanced neural network performance. However, the stacked structures of 3D memories result in notable heat accumulation because lateral interfaces obstruct vertical heat dissipation, thereby hindering effective cooling. An effective approach to mitigating energy consumption involves the utilization of nonvolatile memory technologies, such as resistive random-access memory (RRAM). Integration of selector transistors with RRAM devices mitigates sneak path leakage, increases nonlinearity, and improves the reliability of vertically stacked arrays. Nevertheless, executing core AI tasks—such as vector-matrix multiplication in neuromorphic computing—requires substantial current flow through these transistors, which in turn leads to heat generation, reduced power efficiency, and potential computational errors. Additionally, densely stacked layers create hotspots and restrict access to cooling interfaces. This study presents a comparative analysis of models with various selector transistor configurations, based on power parameters from microfabricated 3D RRAM structures. The results indicate that optimally positioning the selector transistor at the memory interface can reduce nanoscale heat accumulation by up to 11%, as verified through finite-element simulations and numerical calculations. Improved thermal management reduced peak local temperatures from over 160 °C to below 60 °C within 20 nanoseconds in configurations featuring 10 to 100 stacked layers.

New concepts

3D memory technologies, such as high-bandwidth memory, are essential components of AI accelerators and play a critical role in AI performance. However, the 3D architecture leads to significant heat accumulation, as lateral interfaces impede vertical heat dissipation and hinder efficient cooling. A portion of this heat arises from Joule heating due to electron transfer, particularly near selector transistors that collect current from individual memory cells. Another major contributor is sneak-path leakage—unintended current flow through unselected memory cells—caused by parasitic electrical paths between word and bit lines in the 3D structure. In this study, we demonstrate that in 3D-stacked resistive random-access memory, strategically positioning the selector transistor at the memory interface can reduce nanoscale heat buildup by up to 11%, as confirmed through finite-element simulations and numerical analysis. Ultra-low leakage transistors, such as IGZO selectors, offer lower latent power consumption compared to conventional CMOS technology. Improved thermal management reduced peak local temperatures from over 160 °C to below 60 °C within 20 nanoseconds in configurations with 10 to 100 stacked layers. Importantly, with these improvements, the costly silicon area underneath the RRAM—previously occupied by the selector transistor—is now available for other circuits, such as logic or peripheral components.

1. Introduction

The increasing demands for data processing tasks have highlighted the limitations of the conventional von Neumann computational architecture, which separates processing and memory components, causing a data transfer bottleneck. The processing-in-memory (PIM) methodology has been proposed as a potential solution to this problem and storage class memories (SCMs), encompassing resistive random access memory (RRAM), phase change memory (PCM), ferroelectric RAM (FeRAM), and spin-transfer torque magnetic RAM (STT-MRAM), have surfaced as suitable memory devices for PIM applications.[1–7] Among these alternatives, RRAM is gaining interest due to its benefits, including non-volatility, a simple two-terminal structure, compatibility with complementary metal-oxide-semiconductor (CMOS) technology, and fast operational speed compared to flash memories.[1–5,8–11]

The 3D vertical resistive random access memory (VRRAM) integration technique significantly enhances memory density by incorporating 3D vertical NAND (VNAND) technology and transforming the traditional 2D crossbar RRAM into a vertical

[a] Department of Electronic Engineering, Kyung Hee University, Yongin 17104, Republic of Korea. E-mail: seansl@khu.ac.kr
[b] Department of Mechanical Engineering, Wichita State University, KS, USA
[c] Department of Semiconductor Engineering, Kyung Hee University, Yongin 17104, Republic of Korea
† A. Kadyrov and S. Lee contributed equally.

configuration.[12,13] During the manufacturing process of 3D RRAM, layers of metal and oxide are deposited in sequence and then collectively etched to create a trench. A switching layer is then introduced to the trench through the atomic layer deposition (ALD) process and filled with a pillar metal, resulting in the formation of plane electrodes and pillar electrodes of the VRRAM. The 3D RRAM technology is particularly cost-efficient as it can effectively diminish the number of lithography process steps and masks.[12,14–16] Consequently, it is emerging as a promising technology that provides superior density and cost-effectiveness in comparison to the conventional cross-bar RRAMs.

Nonetheless, the principal obstacle linked with 3D RRAM arrays is the sneak path leakage that transpires between layers and adjacent cells.[17] Traditional strategies to mitigate sneak path leakage include either using selector transistors or self-rectifying RRAM.[17,18] The advantage of self-rectifying RRAM lies in the simplicity of the structure allowing for better scaling than the 1T-1R structure. Furthermore, analogue resistive switching of the self-rectifying RRAM is suited for in-memory and neuromorphic operations.[19,20] On the other hand, the selector transistor model is preferred when its enhanced control and multilevel behaviour can help with optimization of simulating synaptic processes.[21–23] The selector structure can effectively inhibit sneak path leakage by deactivating the transistor of the adjacent cell that is not the target for reading during the read operation. Several selector transistors based on different materials such as molybdenum disulfide (MoS_2) and tungsten disulfide (WS_2), or oxide-based transistors like indium gallium zinc oxide (IGZO) and indium zinc oxide (IZO) have been introduced.[24–26] Among these, the IGZO transistor is a potential candidate for 1T-1R operation due to its exceptionally low leakage current, large on/off ratio, and back-end-of-line (BEOL) compatibility.

In general, the switching speed of a transistor is determined by several factors including carrier mobility and subthreshold swing. While carrier mobility of IGZO FET is behind the state-of-the-art silicon transistor, the subthreshold swing values approach the sub 90 mV dec^{-1} values comparable to other transistors.[27]

Managing VRRAM with multiple stacked layers using a selector transistor presents a significant challenge due to thermal accumulation. When numerous layers of VRRAM are stacked and voltage is applied simultaneously to multiple plane electrodes, thermal energy accumulates at the base. This escalation increases the likelihood of operational errors or failures due to a substantial rise in the transistor's temperature. Moreover, higher temperatures result in changes in oxygen supply[28] and a reduction in endurance.[29,30] Therefore, it is crucial to analyse the structure for efficient heat dissipation in an ultra-high density VRRAM structure.

This study primarily concentrates on a fully BEOL-compatible monolithic 3D (M3D) stacked 1T-2VR device with a top FET configuration. We have devised a 1T-2VR structure with an oxide-based selector transistor interconnected in an M3D configuration, integrating an IGZO transistor atop a CMOS compatible VRRAM in a dual-layer arrangement. The characterization of the device attributes of the dual-layer VRRAM and IGZO selector transistor has been performed. Subsequently, we conducted a heat dissipation analysis using the COMSOL simulation toolkit, based on device measurement data and actual dimensions. We examined the feasibility of stacking VRRAM devices with hundreds or more layers by comparing the top FET and bottom FET structures and analysing heat accumulation relative to the number of stacked layers. The main goal of the study is to verify how the position of the thermal interfaces changes heat generation, flow, and accumulation in the vertically stacked RRAM device with a selector transistor.

2. Device structure

In Fig. 1, the overall structure as a 3D model, TEM cross-section images and segregated by layer RRAM measurements are displayed. Fig. 1(a) presents the cross-sectional view of the 3D VRRAM. The 3D VRRAM is composed of two layers of plane electrode, which are formed by alternating deposition of 7 nm of TiN and 25 nm of SiO_2 passivation layer. A trench is formed by simultaneously etching the two layers of the plane electrode using RIE, followed by a uniform deposition of a 5 nm HfO_2 switching layer by ALD. A significant 90 nm of Pd, with an adhesion layer of Ti (2 nm), fills the trench to create a drain and pillar electrode. The TEM image reveals the composition of each element through EDS analysis, as depicted in Fig. 1(b). The composition of an FET and double layer of memory can be observed in Fig. 1(c) with the inset showing precise dimensions.

Fig. 1(d) and (e) display the DC sweep measurement outcomes for the two-layer VRRAM, corresponding to L1 and L2, respectively. Each *I–V* curve represents a DC measurement outcome of 20 cycles. A comparison of these two graphs reveals that the variations in memory window, set/reset voltage, and current level are minimal, suggesting that the variation between layers of VRRAM is insignificantly small.

3. VRRAM and FET *IV*-characteristics

Fig. 2 provides a detailed overview of the performance data for both the IGZO FET and the 3D VRRAM, as well as the combined performance of a single 1T-1R device. Fig. 2(a) depicts the resistance distribution for the low and high resistance states of RRAM, measured at a reference voltage of 0.2 V. The average on/off ratio is observed to be 460. Fig. 2(b) shows the distribution of programming voltages for SET and RESET cycles, with the average value for RESET being −2 V and for SET being 3V. Fig. 2(c) shows endurance pulse measurements for over 1000 cycles, where fluctuations around 5 MΩ for HRS and around 7 kΩ for LRS can be seen. The variability tends to increase with the number of cycles. Fig. 3(d) exhibits a retention test over 10 000 s for HRS and LRS values, with no significant change observed over the specified time interval, indicating robust retention results. Fig. 2(e) displays 20 cycles of 1T-1R measurements, with the average SET voltage around 4 V and the average RESET voltage around −4 V. The compliance current is 1 mA. Fig. 2(f) shows the cumulative probability for resistance levels of the 1T-1R structure, with the average LRS around 20 kΩ and the average HRS around

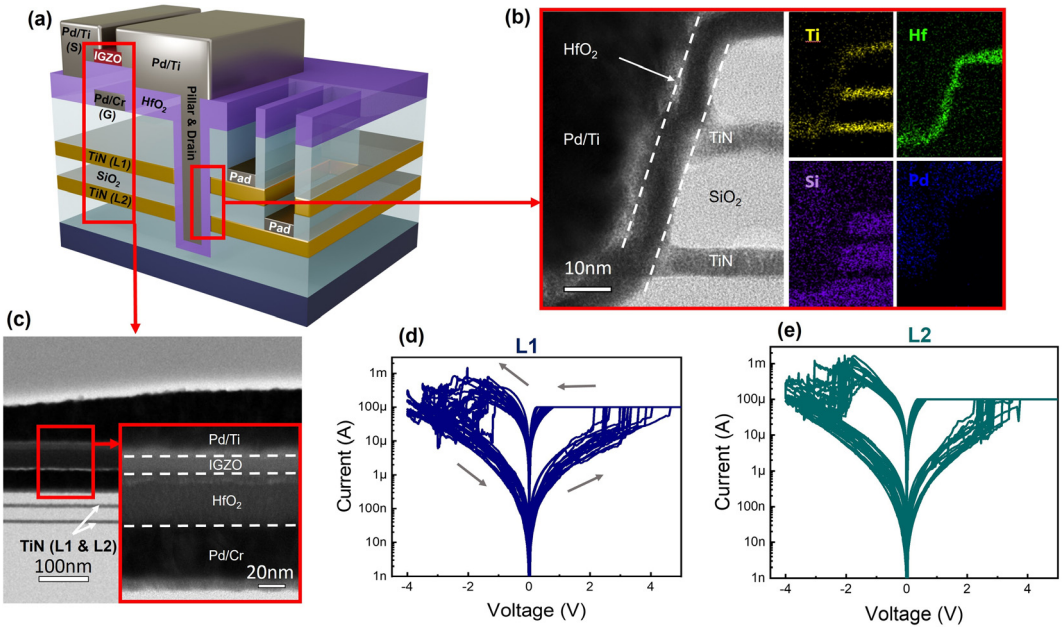

Fig. 1 3D model, TEM cross section view and layer characterization of a fabricated device. (a) A 3D model of the individual 1T-2VR device, which is characterized by a top FET design. The integration of the IGZO transistor onto the dual layers of the TiN plane electrode is illustrated in the (b) cross-sectional TEM image. (c) Cross-section view of the IGZO FET and both RRAM layers stacked on top of each other. (d) Memory performance of layer 1. (e) Memory performance of layer 2. The IGZO FET is designed with a localized back-gate structure, enabling the examination of each material and its corresponding thickness in the inset. The back-gate metal employs 30 nm of Pd, with an additional 5 nm of Cr serving as an adhesion layer. The gate oxide is made up of 30 nm of ALD HfO_2. A 13 nm thick IGZO channel is situated on top of it. Given the overlap of the source and gate metal, it is verified that there is 90 nm of Pd, which acts as the source metal, and 2 nm of Ti as an adhesion layer. Transistor dimensions are W = 140 μm and L = 6 μm. In the scenario of 1T-1VR operation, for example, voltage is applied to the L_1 plane, with the positive voltages being applied by probing the L_1 pad and gate metal, respectively, and the source is grounded. The proposed 1T-2VR structure combines the drain of the IGZO transistor with the pillar electrode of the 3D VRRAM into a single entity. This design effectively minimizes the photolithography steps and fabrication steps by eliminating the need for the deposition and pattern of the source and the pillar electrode, as well as the process of forming a *via* interconnect to link them.

5 MΩ, resulting in an on/off ratio of 447. Fig. 2(g) presents 20 cycles of I_D–V_{GS} characteristics of an IGZO FET with a local back gate and VDS at 1 V. The maximum current reaches 1 mA at 6 V, with positive voltage threshold switching and an off-current below 100 pA. Fig. 2(h) reveals that the leakage current remains below 1 nA throughout the measurement. Fig. 2(i) illustrates I_D–V_{DS} characteristics of the IGZO FET with a gate bias ranging from 0 to 6 V.

4. 3D model with simulation comparison

To establish a proper simulation several key steps are necessary: building a representative CAD model, defining the material boundaries and their parameters, defining the thermal boundary conditions, establishing the thermal equations and sources, defining the mesh of the model, and running the simulation.[31]

Material properties necessary for the simulation are collected in Table 1.[32–36] Fig. 3 presents a three-dimensional model that integrates a 3D RRAM with an FET device across 10 layers. This study uncovers the maximum potential temperature variance between the two models being compared.

Fig. 3(a) provides a cross-sectional image of single memory and FET units with respective labels put on each layer. Fig. 3(b) depicts the results for the lower FET structure, while Fig. 3(c) shows the results for the upper FET structure. A voltage pulse of 2 V was applied to the electrode L_{10} and L_1 over 1 μs for both the lower and upper FET structures, respectively, with grounding applied to the bottom electrode and top electrode corresponding to their structure. The data indicates that the maximum temperature variance between the two structures is approximately 500 degrees within a single pulse, taking into account the effect of active cooling. For both models, all sides were insulated except for the top, which had active cooling to keep the temperature at the top at 27 °C.

4.1. Model dimensions

Fig. 3 delineates the dimensions of a model utilized for COMSOL simulations. The top FET structure is portrayed in Fig. 3(d), whereas Fig. 3(e) exhibits the bottom FET structure for comparative analysis. The primary difference is the position of the FET, as indicated in the model names. The top FET necessitated a minor extension to ensure its source coincided with the Pd pillar of the 3D RRAM, reflecting the actual fabricated model that served as the primary reference tool. Our reference model had 2 layers, meaning that L_n = L_2. However, the model used for simulation had varying layer numbers from 10 to 100 layers to show the possible future implications of stacking on heat distribution. The dimensions

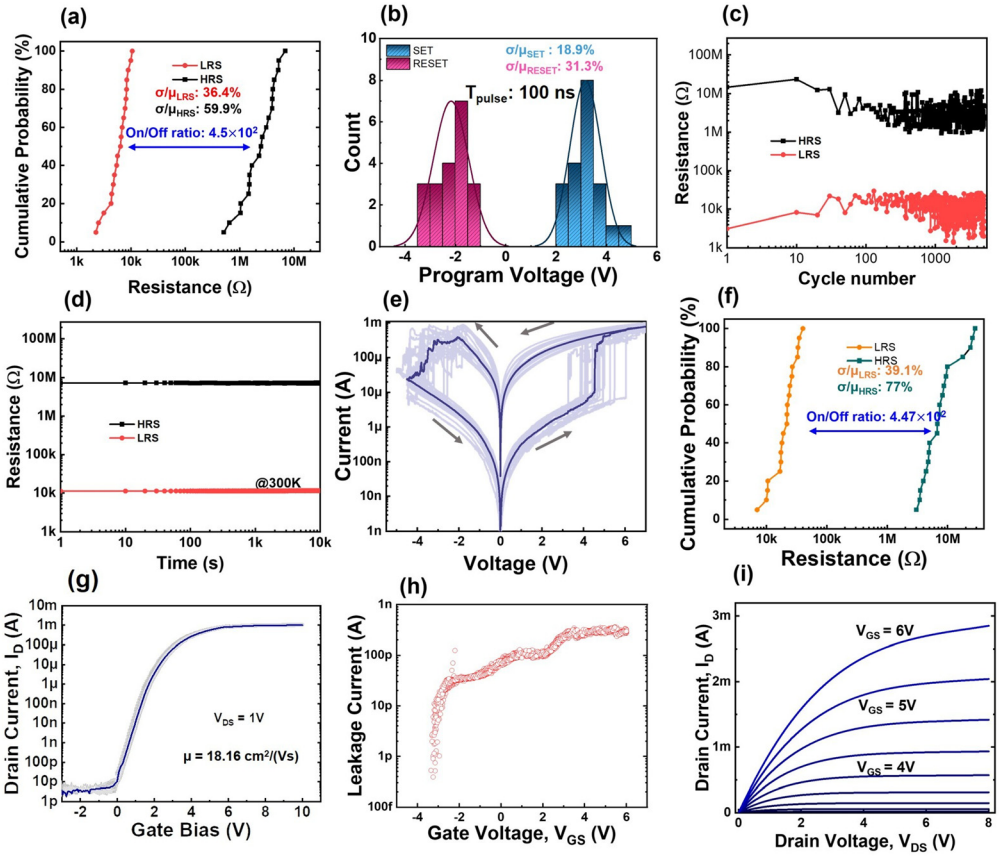

Fig. 2 Measurement outcomes of the 3D VRRAM, IGZO FET, and 1T-1VR single device are as follows: (a) The distribution of resistance for the LRS and HRS states at a read voltage of 0.2V. (b) The distribution of program voltage for Set and Reset. (c) Pulse endurance exceeding 10^4 cycles. (d) Retention test lasting more than 10^4 s. (e) 20 cycles of 1T-1VR SET and RESET measurements. (f) The distribution of LRS and HRS resistances of the 1T-1VR measurement. (g) 20 cycles of I_D–V_{GS} characteristics of the locally back-gated IGZO FET with 1 V for V_{DS}. (h) Gate leakage. (i) I_D–V_{DS} characteristics of the IGZO FET from a gate bias ranging from 0 V to 6 V.

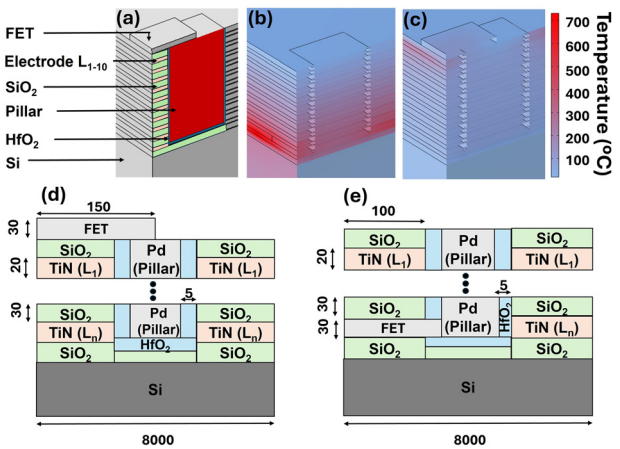

Fig. 3 3D visualization of a 10-layer 3D RRAM + FET, highlighting the highest temperature differences among the models. The illustration includes (a) descriptive annotations for the model, (b) the temperature of the lower FET structure when a voltage bias is applied at L_{10}, and (c) the temperature of the upper FET structure when a voltage bias is applied at L_1. Cross section view of a model showcasing the dimensions used in COMSOL simulation (nanometer scale). The FET location is (d) at the top, and (e) at the bottom of the structure.

of the model closely correspond to those of the actual fabricated device. Slight variations in dimensions are attributable to the complexity of the model, which can interfere with mesh operations when there are several objects with substantial dimensional disparity. In our endeavour to render the Si heat sink as realistic as possible, its relative dimensions became disproportionately large compared to individual layers, thereby necessitating a minor augmentation in layer thickness.

4.2. 2D model for heat distribution analysis

The next step in assessing the thermal energy distribution properties in a specific model entails simulating the dispersion of heat within the set boundary conditions. For this, the models

Table 1 Properties of materials

Material	Electrical conductivity [S m^{-1}]	Thermal conductivity [W m^{-1} K^{-1}]	Heat capacity [J kg^{-1} K^{-1}]
TiN	1×10^5	19	49.69
SiO_2	1×10^{-14}	1.4	730
Pd	1×10^7	71.8	244
Si	1	34	678
HfO_x	1×10^4	23	144

depicted in Fig. 4 are thermally insulated on both the left and right sides to replicate situations where a single memory and transistor cell remain unaffected by the cumulative heat of neighbouring memories. Simultaneously, both structures are subjected to cooling at the top and bottom.

The top FET structure in Fig. 4(a) shows a propensity for heat to gather near the top side of the entire structure, moving away from the direction of the top FET; furthermore, illustrating the effect of heat dissipation on the temperature of the Si heat sink after 20 ns. It can be seen that the effect on the silicon heat sink situated at the bottom of the structure is negligible. In contrast, Fig. 4(b) portrays the heat distribution in a bottom FET configuration, where it shows that the initial heat disperses throughout the entire body starting from the bottom FET, with more heat tending to gather towards the bottom Si heat sink in comparison to the Top FET model. It is significant that major heat is transferred to the silicon heat sink at the bottom of the structure, since reduced heat accumulation at the bottom of the top FET structure can result in more efficient use of the space in place of a silicone heat sink, for instance, other electronic components. Moreover, the lack of immediate access to the cooling element on the top of the structure leads to an overall higher maximum temperature inside the structure for the second model, increasing from 54 to 163 °C.

From Fig. 4(c), continuation of the trend can be observed. More memory stacks lead to an increase in the maximum temperature. At the same time, the heat accumulated in the second structure continues to be significantly higher than the heat accumulated in the top FET structure.

Owing to the widespread distribution of heat in the bottom FET structure, it becomes difficult to effectively dissipate heat using cooling systems, even if they are located at both the top and bottom. This is due to limited access to the middle layers of the structure, resulting in a lower cooling efficiency compared to a top FET structure where heat is mainly accumulated at the top.

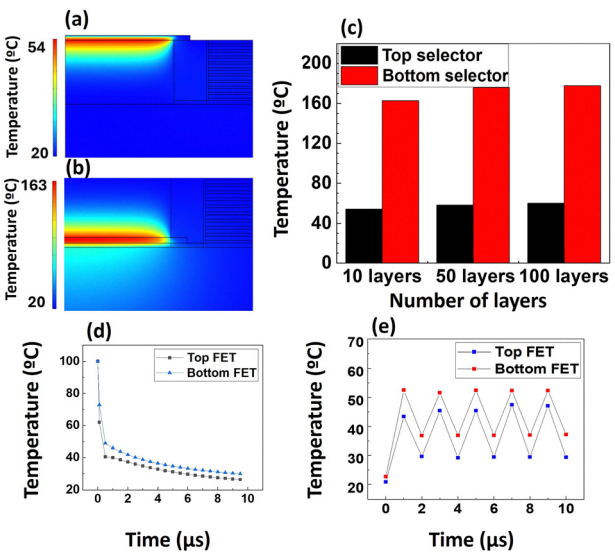

Fig. 4 Examination of heat dispersion for a structure comprising 10 layers, with emphasis on (a) the initial heat source after 20 ns for the top FET structure, (b) the initial heat source after 20 ns for the bottom FET structure and (c) the comparison of max temperatures of both models for 10-, 50- and 100-layer structures, (d) heat dissipation comparison and (e) the complete cycle of thermal elevation and reduction for both FET structures.

4.3. Complete cycle thermal variance

Upon the completion of the heat accumulation examination, the subsequent pivotal phase is the dispersion of heat. In this context, two models were scrutinized based on their post-heating cooling rate. The top FET model exhibited heat dissemination in the upper half, as unveiled by the prior simulation, as illustrated in Fig. 4(a). In contrast, the heat in the bottom FET model was concentrated at the bottom, as illustrated in Fig. 4(b). The results of the cooling rate simulations are illustrated in Fig. 4(d). It is apparent that within a duration of 1 µs, the top FET and bottom FET attain temperatures of 40 °C and 45 °C, respectively, making the top FET structure have a temperature which is 11% lower than the bottom FET structure.

The final critical evaluation involves gauging the performance discrepancy between the two models under recurrent operational conditions. A 2V, 1 µs pulse was administered to L_{50} of both the top FET and bottom FET structures of a 100-layer model, succeeded by a 1 µs cooling interval to simulate working conditions. This on and off cycle was reiterated across multiple cycles. The findings are documented in Fig. 4(e). The general trend suggests that the top FET consistently sustains a lower temperature than the bottom FET structure. The overall thermal efficiency of the top FET appears to be 11% higher in comparison to the bottom FET model.

4.4. Model verification

One of the primary equations used for the local temperature for the COMSOL calculations is the Fourier law of heat conduction:[37]

$$1 \times \sqrt{2}/2 \times \sqrt{7} \quad (1)$$

And its derivation for a single dimension:[37]

$$1 \times \sqrt{5}/2 \times 4 \quad (2)$$

where k is the thermal conductivity, A – active area, and ΔT – temperature difference between the initial and final stages of an object.

Eqn (2) is used for calculating the temperature increase in the equivalent model to verify the COMSOL calculations.

In the case of the current work, the heat equation for the simplified model has been solved and compared to the results presented by COMSOL on the model with similar dimensions. This process ensures that the error margin of calculations remains low and verifies the outcome of an experiment.

To successfully calculate the simplified equivalent model, it is necessary to first calculate the Biot number (Bi):[38]

$$1 \times 1/2 \times \sqrt{3} \quad (3)$$

where k is the thermal conductivity, h – convective heat transfer coefficient, and L – characteristic length of an object.

In our case, the calculation results after using Eqn (3) showed that our Bi $\ll$ 0.1, from which it was concluded that the convection to conduction ratio of our model allows us to use the simplified equivalent model for further mathematical derivations. Using this model and applying the empirical results from Fig. 1, the temperature calculated by COMSOL reached 27.1 °C, whereas the temperature calculated using the conventional method reached 25.5 °C with the estimated error being 5.9%, which can be considered a reasonable error margin.

Conclusions

This research involves fabricating and integrating three-dimensional VRRAM with the top FET. The integration with the selector FET aids in controlling the RRAM and programming specific memory cells, enhancing performance. The study focused on the feasibility of using the top FET structure instead of the conventional bottom FET structure. RRAM and FET tests showed strong performance across multiple layers of a manufactured device. Additionally, COMSOL simulations for thermal dissipation differences between the top and bottom FET structures were performed. Analyses concluded that the temperature variance between both structures is notable, with heat accumulating at the top for the top FET and dispersing in the bottom layers for the bottom FET. This led to heat confinement and a decreased cooling rate for the bottom FET. The top FET structure's temperature was approximately 11% lower than that of the bottom FET in the middle layers, making it a more efficient model. The COMSOL model was verified by conventional calculations, with an error margin of 5.9%, supporting its validity. Moreover, the hotspot temperature decreased from over 160 °C to below 60 °C within 20 nanoseconds pulse during the stacks' temperature comparison between two models.

Author contributions

Conceptualization: A. K. and S. L. investigation: S. L. and J. C. data curation: J. B. and M. L. formal analysis: A. K. and A. D. writing – original draft: A. K. and S. L. writing – review and editing: all the authors. Supervision: G. H. and S. L. project administration: S. L. visualization: A. K., S. L., B. A. and S. P. For more information about the CRediT taxonomy, please refer to https://credit.niso.org

Conflicts of interest

There are no conflicts to declare.

Data availability

The data supporting this article have been included as part of the supplementary information (SI). Supplementary information is available. See DOI: https://doi.org/10.1039/d5nh00653h.

Acknowledgements

This work was supported in part by the National R&D Program through the National Research Foundation of Korea (NRF) funded by Korean Ministry of Science and ICT, under Grant 2018R1A6A1A03025708, Grant 2009-0082580, and Grant RS-2020-NR047143; in part by the National Program supervised by Institute of Information and Communications Technology Planning and Evaluation (IITP) under Grant 2022-0-00248-001; and in part by the Industrial Fundamental Technology Development Program ''Public-private joint investment semiconductor R&D program(K-CHIPS)'' Grant RS-2024-00406182 funded by the Korean Ministry of Trade, Industry & Energy.

Notes and references

1 F. Zahoor, T. Z. Azni Zulkifli and F. A. Khanday, *Nanoscale Res. Lett.*, 2020, **15**, 90–106.
2 H. S. P. Wong, H. Y. Lee, S. Yu, Y. S. Chen, Y. Wu, P. S. Chen, B. Lee, F. T. Chen and M. J. Tsai, *Proc. IEEE*, 2012, **100**, 1951–1970.
3 J. S. Meena, S. M. Sze, U. Chand and T.-Y. Tseng, *Nanoscale Res. Lett.*, 2014, **9**, 526.
4 F. Pan, S. Gao, C. Chen, C. Song and F. Zeng, *Mater. Sci. Eng., R*, 2014, **83**, 1–59.
5 H. S. P. Wong and S. Salahuddin, *Nat. Nanotechnol.*, 2015, **10**, 191–194.
6 S.-T. Lee and J.-H. Lee, *Nanoscale Horiz.*, 2024, **9**, 1475–1492.
7 Z. Cao, B. Sun, G. Zhou, S. Mao, S. Zhu, J. Zhang, C. Ke, Y. Zhao and J. Shao, *Nanoscale Horiz.*, 2023, **8**, 716–745.
8 G. W. Burr, M. J. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, L. A. Lastras, A. Padilla, B. Rajendran, S. Raoux and R. S. Shenoy, *J. Vacuum Sci. Technol. B*, 2010, **28**, 223–262.
9 V. Gupta, S. Kapur, S. Saurabh and A. Grover, *IETE Tech. Rev.*, 2020, **37**, 377–390.
10 D. Kumar, R. Maddu, H. J. Chung, H. Rahaman, T. Jin, S. Bhatti, S. Ter Lim, R. Sbiaa and S. N. Piramanayagam, *Nanoscale Horiz.*, 2024, **9**, 1962–1977.
11 E. K. Koh, P. A. Dananjaya, H. Y. Poh, L. Liu, C. X. X. Lee, J. R. Thong, Y. S. You and W. S. Lew, *Nanoscale Horiz.*, 2024, **9**, 828–842.
12 Y. Deng, H. Y. Chen, B. Gao, S. Yu, S. C. Wu, L. Zhao, B. Chen, Z. Jiang, X. Liu, T. H. Hou, Y. Nishi, J. Kang and H. S. P. Wong, 2013 IEEE International Electron Devices Meeting, 2013, 1, 25.27.21-25.27.24.
13 Y.-R. Jeon, D. Akinwande and C. Choi, *Nanoscale Horiz.*, 2024, **9**, 853–862.
14 M. Yu, Y. Fang, Z. Wang, G. Chen, Y. Pan, X. Yang, M. Yin, Y. Yang, M. Li, Y. Cai and R. Huang, *Nanotechnology*, 2016, **27**, 205202.
15 S. G. Park, M. K. Yang, H. Ju, D. J. Seong, J. M. Lee, E. Kim, S. Jung, L. Zhang, Y. C. Shin, I. G. Baek, J. Choi, H. K. Kang and C. Chung, 2012 International Electron Devices Meeting, 2012, 1, 20.28.21-20.28.24.
16 P. Y. Chen and S. Yu, 2014 IEEE 6th International Memory Workshop (IMW), 2014, 1, 1-4.

17 U. Dilna and S. N. Prasad, 2020 Third International Conference on Multimedia Processing, Communication & Information Technology (MPCIT), 2020, 1, 138-145.
18 J. M. Cho, S. S. Kim, T. W. Park, D. H. Shin, Y. R. Kim, H. J. Park, D. Y. Kim, S. H. Lee, T. Park and C. S. Hwang, *Nanoscale Horiz.*, 2025, **10**, 113–123.
19 L. Hu, J. Yang, J. Wang, P. Cheng, L. O. Chua and F. Zhuge, *Adv. Funct. Mater.*, 2021, **31**, 2005582.
20 T. D. Dongale, G. U. Kamble, D. Y. Kang, S. S. Kundale, H.-M. An and T. G. Kim, *Phys. Status Solidi RRL*, 2021, **15**, 2100199.
21 Q. Wang, Y. Wang, R. Luo, J. Wang, L. Ji, Z. Jiang, C. Wenger, Z. Song, S. Song, W. Ren, J. Bi and G. Niu, *Neuromorphic Comput. Eng.*, 2022, **2**, 044012.
22 Z. Xia, X. Sun, Z. Wang, J. Meng, B. Jin and T. Wang, *Nano-Micro Lett.*, 2025, **17**, 217.
23 B. Chen, X. Wang, B. Gao, Z. Fang, J. Kang, L. Liu, X. Liu, G.-Q. Lo and D.-L. Kwong, *Sci. Rep.*, 2014, **4**, 6863.
24 C.-J. Liu, Y. Wan, L.-J. Li, C.-P. Lin, T.-H. Hou, Z.-Y. Huang and V. P.-H. Hu, *Adv. Mater.*, 2022, **34**, 2107894.
25 F. A. McGuire, Y.-C. Lin, K. Price, G. B. Rayner, S. Khandelwal, S. Salahuddin and A. D. Franklin, *Nano Lett.*, 2017, **17**, 4801–4806.
26 X. Wang, X. Chen, J. Ma, S. Gou, X. Guo, L. Tong, J. Zhu, Y. Xia, D. Wang, C. Sheng, H. Chen, Z. Sun, S. Ma, A. Riaud, Z. Xu, C. Cong, Z. Qiu, P. Zhou, Y. Xie, L. Bian and W. Bao, *Adv. Mater.*, 2022, **34**, 2202472.
27 Y. Li, L. Zhu, C. Chen, Y. Zhu, C. Wan and Q. Wan, *Chin. Phys. Lett.*, 2022, **39**, 118501.
28 Y. Y. Chen, M. Komura, R. Degraeve, B. Govoreanu, L. Goux, A. Fantini, N. Raghavan, S. Clima, L. Zhang, A. Belmonte, A. Redolfi, G. S. Kar, G. Groeseneken, D. J. Wouters and M. Jurczak, 2013 IEEE International Electron Devices Meeting, 2013, 1, 10.11.11-10.11.14.
29 B. Chen, J. F. Kang, B. Gao, Y. X. Deng, L. F. Liu, X. Y. Liu, Z. Fang, H. Y. Yu, X. P. Wang, G. Q. Lo and D. L. Kwong, *IEEE Electron Device Lett.*, 2013, **34**, 1292–1294.
30 S. Mao, B. Sun, G. Zhou, T. Guo, J. Wang and Y. Zhao, *Nanoscale Horiz.*, 2022, **7**, 822–848.
31 Q. Chen, Z. Wang, M. Yu, Y. Fang, Z. Yu, Y. Yang, Y. Cai and R. Huang, *Phys. Scr.*, 2019, **94**, 045001.
32 A. Munster, K. Sagel and G. Schlamp, *Nature*, 1954, **174**, 1154–1155.
33 L. Olimov and I. Khojimatov, *E3S Web Conf.*, 2023, **458**, 01022.
34 M. J. Laubitz and T. Matsumura, *Can. J. Phys.*, 1972, **50**, 196–205.
35 M. H. Brodsky and R. J. Gambino, *J. Non-Cryst. Solids*, 1972, **8–10**, 739–744.
36 W. B. J. Zimmerman, *Multiphysics Modeling with Finite Element Methods*, World scientific, 2006, vol. 18, pp. 1–26.
37 I. S. Liu, *Continuum Mech. Thermodyn.*, 1990, **2**, 301–305.
38 M. G. Christensen and J. Adler-Nissen, *Appl. Therm. Eng.*, 2015, **86**, 214–221.