STU
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

Journal of Electrical Engineering, Vol. 77, No. 1, 2026, pp. 10-20
https://reference-global.com/journal/JEE

Performance investigation of non-volatile memories for energy-efficient on-chip L2 caches

Inderjit Singh[1,2], Balwinder Raj[1*], Mamta Khosla[1] and Tajinder Kaur[3]

This paper explores the potential of emerging non-volatile memory (NVM) technologies – including Spin-Transfer Torque Magnetic RAM (STT-MRAM), Spin-Orbit Torque MRAM (SOT-MRAM), Voltage-Controlled Magnetic Anisotropy (VCMA)-based MRAM, and Resistive RAM (RRAM) – to reshape the CPU memory hierarchy landscape. These NVMs offer several advantages over conventional SRAM, such as ultra-low leakage, high density, non-volatility, and scalability. We present a comprehensive comparison between these technologies and SRAM across cache capacities ranging from 16KB to 8MB and associativity values from 1 to 16 at 22 nm node. The results reveal that SRAM outperforms NVMs at smaller cache sizes, while NVMs – particularly SOT-MRAM and aggressive STT-MRAM – become favourable beyond 128KB. For a 512KB L2 cache, SOT-MRAM and STT-A MRAM achieve energy-area-delay product (EADP) improvements of 73.7% and 67.8% over SRAM, which further improve to 97.2% and 98.4% at 2MB. However, RRAM and STT-MRAM show significantly higher write energy costs – up to 18.3× and 21.8× that of SRAM at 512KB. In terms of leakage, STT-A MRAM achieves the best results, reducing leakage by nearly 99%, followed by STT and SOT-MRAM. Interestingly, the impact of associativity variation becomes less pronounced at larger cache sizes.

Keywords: non-volatile caches, STT, SOT, RRAM, low power cache, NVSIM

1 Introduction

In modern computing systems, a carefully arranged stack of volatile and non-volatile data storage systems holds a delicate balance between cost and performance considerations. Within this hierarchy, the memory closest to the processor core experiences frequent access demands and must deliver the fastest operational speeds possible. However, this proximity also makes it the most resource-intensive segment due to substantial chip area requirements. Simultaneously, other levels within the memory hierarchy, as illustrated in Fig. 1, prioritize storage capacity and operational speed differently.

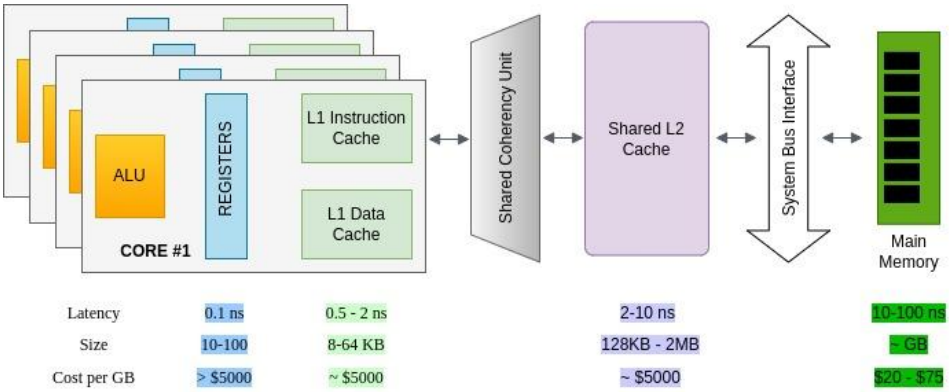


Fig. 1. Memory hierarchy: Different type and size of memories

[1]National Institute of Technology, Jalandhar, Punjab, 144011, India
[2]Professional Services, Cotiviti Pvt. Ltd., Mohali, Punjab, India
[3]SGTB Khalsa College, Sri Anandpur Sahib, Punjab, India
Corresponding authors: rajb@nitj.ac.in

https://doi.org/10.2478/jee-2026-0002, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

Within the realm of computer architecture, the consistent pursuit of improved performance and efficiency has prompted a quest for innovative solutions. As conventional Complementary Metal-Oxide-Semiconductor (CMOS) technology scaling faces formidable obstacles, novel avenues are being explored. This exploration has introduced concept of "More than Moore" [1], a response to the challenges of scaling.

In this evolving landscape, the transition from CMOS-based Static Random-Access Memory (SRAM) to cache solutions based on Non-Volatile Memory (NVM) is a natural progression. The benefits of CMOS scaling that once boosted cache performance are now overshadowed by increased leakage currents, diverse process variations, and heightened power inefficiencies [2]. Consequently, there is a keen search for alternative devices and architectural paradigms to ensure energy sustainability. In a world where multi-core systems and unconventional memory devices reign, the focus has shifted towards cache memory solutions that seamlessly align with contemporary computational demands [3]. Nano-magnetic storage devices, including Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM), Spin Orbit Torque MRAM (SOT-MRAM), Resistive RAM (RRAM), Phase-Change RAM (PC-RAM), and Voltage-Controlled Magnetic Anisotropy (VCMA), hold the promise to reshape cache hierarchies and memory architectures. These Non-Volatile Memory (NVM) technologies, traditionally associated with long-term data storage [4-6], are increasingly transcending their conventional boundaries to address modern computing needs.

In this context, our study embarks on a comprehensive exploration of emerging non-volatile memory (NVM) technologies for on-chip caches, highlighting their performance characteristics across diverse cache configurations. We highlight the overestimation of energy performance variations with associativity in the simulation tool, identifying a potential avenue for tool enhancement while offering valuable insights into the suitability of NVMs in cache hierarchy optimization.

2 Background

Over the last decade, MRAM's viability has been explored across all tiers of computing architectures. The relevant research works are listed in Tab. 1.

Table 1. Related works

Study	Contribution(s)	Method	Key Findings
F. Oboril et al. [7]	Proposed hybrid SOT-MRAM cache	L1-D as SRAM; L1-I, L2 as SOT-MRAM	Energy and performance improvement by 60% and 1%
M. P. Komalan et al. [8]	Cache design methodology to address STT-MRAM read penalty for single-core ARM system	Micro-architectural modifications and code transformations	Reduced performance penalty from ~54% to ~8%
Zhang et al. [9]	STT-MRAM scaling roadmap	Analysis of cell area, aspect ratios, cache capacities	Up to 5.4% latency improvement, 42.1% leakage power reduction
Saha et al. [10]	STT-MRAM vs SOT-MRAM performance in LLC	Benchmark simulation framework	STT-MRAM uses 50% less area, 74% less leakage power; SOT-MRAM 4× faster, better in large LLCs
Marinelli et al. [11]	Optimizing STT-MRAM as LLC	Simulations with technology models	No performance loss, over 60% LLC energy savings
S. Senni et al. [2]	Comparison of STT-MRAM and TAS-MRAM	Performance and energy consumption analysis	Demonstrated up to 90% reduction in LLC energy consumption for MRAM

Dong et al. [12]	Optimal memory hierarchy with ReRAM	Design space exploration framework	Identifies optimal ReRAM use for performance, energy, or cost efficiency
Mittal and Vetter [13]	Enhancing hybrid SRAM-NVM LLC lifetime	Data migration technique	Proposed AYUSH technique significantly increases cache lifetime up to 47.62×
Andrawis et al. [14]	Addressing MRAM high write current	Comparative analysis of magnetic switching techniques	Highlights trade-offs in MRAM design choices
W. Kang et. al. [15]	Evaluating VCMA-MTJ devices focusing on write energy and delay	Magnetization dynamics study, electrical model preparation, and strategy comparison	STT-assisted precessional VCMA most promising for VCMA-MRAM design

3 Magnetic random-access memory

Magnetic Random-Access Memory (MRAM) distinguishes itself from traditional volatile memories such as SRAM and DRAM by storing information through magnetic orientations rather than electric charge. The fundamental storage element in MRAM is the Magnetic Tunnel Junction (MTJ), an insulator sandwiched between two ferromagnetic layers whose relative magnetization orientation – parallel or anti-parallel – determines the binary state. This non-charge-based storage enables non-volatility, near-zero leakage power, high endurance, and robust data retention.

Spin-Transfer Torque MRAM (STT-MRAM) is one of the most mature MRAM variants and employs spin-polarized current to modulate the magnetization of the free layer in the MTJ. STT-MRAM writes data by bidirectional switching of the MTJ between R_{LOW} (parallel) and R_{HIGH} (anti-parallel) states, driven by the flow of spin-polarized electrons that impart spin-transfer torque to the free layer (Fig. 2a). However, STT-MRAM faces key design challenges. Since it uses a common read/write path, the optimization of write current and read reliability introduces a stringent trade-off. High write current requirements increase dynamic energy and pose reliability risks such as time-dependent dielectric breakdown (TDDB) of the tunnel barrier. Thus, balancing low junction resistance for writing and high magnetoresistance (TMR) for readability remains a critical constraint.

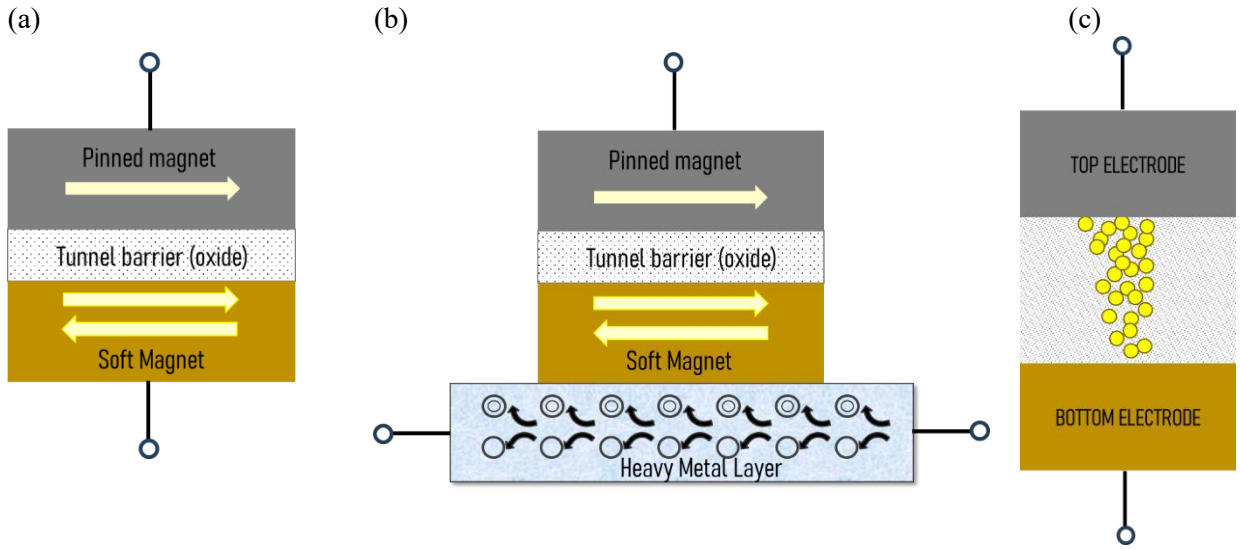


Fig. 2. Schematics of (a) STT-MRAM, (b) SOT-MRAM, and (c) RRAM

To overcome these limitations, the Spin-Orbit Torque MRAM (SOT-MRAM) architecture has recently emerged as a promising innovation (Fig. 2b). SOT-MRAM reverses magnetization by exploiting spin-orbit torque generated through the Spin Hall Effect (SHE). Unlike STT-MRAM, the write current does not pass through the MTJ, enabling independent read and write paths. As a result, SOT-MRAM provides significantly faster switching, lower write energy, greater endurance, and immunity to read-disturb issues – a major limitation in STT-based designs. The separate write path, however, requires an additional access transistor, slightly increasing cell area compared to STT-MRAM.

STT-Aggressive MRAM, developed further by NVSIM [16], and Voltage-Controlled Magnetic Anisotropy MRAM (VCMA-based MRAM) manipulate MTJ magnetic anisotropy through electric fields, markedly lowering write energy and delay [17, 18]. Resistive RAM (ReRAM), with a metal-insulator-metal (MIM) configuration, records data in resistive states, boasting lower power usage and greater density than MRAM and remains stable even at 10 nm process technologies (Fig. 2c), though it faces challenges with electrical property consistency.

4 Simulation framework

The framework illustrated in Fig. 3 facilitates the assessment of diverse memory array configurations suitable for on-chip caches. Through the utilization of SPICE simulations, we create a model for a single-bit cell and conduct a comprehensive circuit-level analysis spanning the entire memory array, including decoders, read/write circuitry, and I/O buffers. To deduce area, latency, and power consumption metrics for various NVMs, we employ NVSIM [16], a customized version of the open-source CACTI [19] tool.

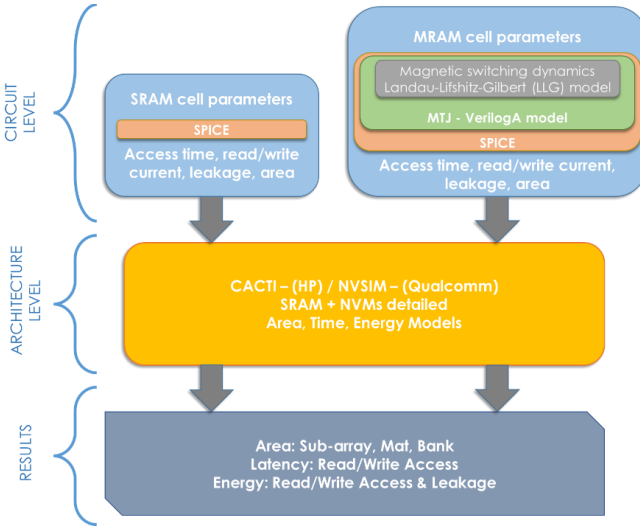


Fig. 3. Cross-tool performance evaluation framework

An indicative performance evaluation of various 512KB NVM technologies is briefed in Tab. 2.

Additionally, model parameters specific to SOT-MRAM are acquired through in-house Cadence Virtuoso circuit simulations using the TSMC 22 nm library [20]. These parameters are subsequently integrated into NVSIM to facilitate the evaluation of the memory array.

Table 2. Indicative performance comparison of different NVM technologies (Capacity 512KB)

Technology	Memory type	Cell size (F^2)	Access time (ns) R / W	Dynamic energy (pJ) R / W	Static power (mW)	Merits / Shortcomings [1]
SRAM [21]	CMOS Latch	146	4-5.5 /4-5.5	100 / 5	290	(++) Endurance (–) Scalability (–) Radiation vulnerable
SOT – MRAM [20, 22-29]	Magnetization	57	1.94 / 1.55	391 / 34	1.5	(+) Endurance (+) Scalability (+) Radiation immune
STT – MRAM [5, 30-35]	Magnetization	54	2.5 / >10	437 / 119	~1	(+) Scalability (+) Endurance (+) Radiation immune (–) Bit Failure rate
ReRAM [36, 37]	Resistance	4	2.5 / >20	418 / 133	1.7	(+) Scalability (–) Bit Failure rate (–) Endurance (–) Retention
PC RAM [38, 39]	Resistance	9	~1 / 40-150	30 / >8000	1.13	(±) Scalability (–) Endurance

NAND [40]	Floating gate	4	95 / 1.2×10^5	271 / 5.2×10^5	5.28	(–) Scalability (–) Endurance (–) Radiation vulnerable

NVSIM takes as input the device-level parameters like bit-cell switching energy, latency, leakage and the organization details like memory capacity, partitioning, optimization, and the technology node for CMOS peripheral circuitry. NVSIM then analytically extracts the per access (read/write) latencies, access energies, leakage, and area for the given memory architecture (Fig. 3). NVSIM is suitable for rapid prototyping of the memory chip with an acceptable degree of accuracy. For added precision SPICE simulations should be used.

In order to perform a comparative analysis among NVMs, the chosen cell model parameters for 22 nm technology node at 300 K are listed in Tab. 3.

The tool incorporates diverse cell models that include SRAM, RRAM, and two STT-MRAM variants (standard and aggressive). We included SOT-MRAM cell model from the prior research [20] and a VCMA model from [18] in the analysis. Essential cache parameters like size, associativity, line width, and technology node can be defined in the configuration files. Various optimization targets and device roadmap parameters – HP, LOP, and LSTP are explored. Excluding the HP roadmap due to unexpectedly high leakage, one limitation is NVSim's lack of multi-bank memory support.

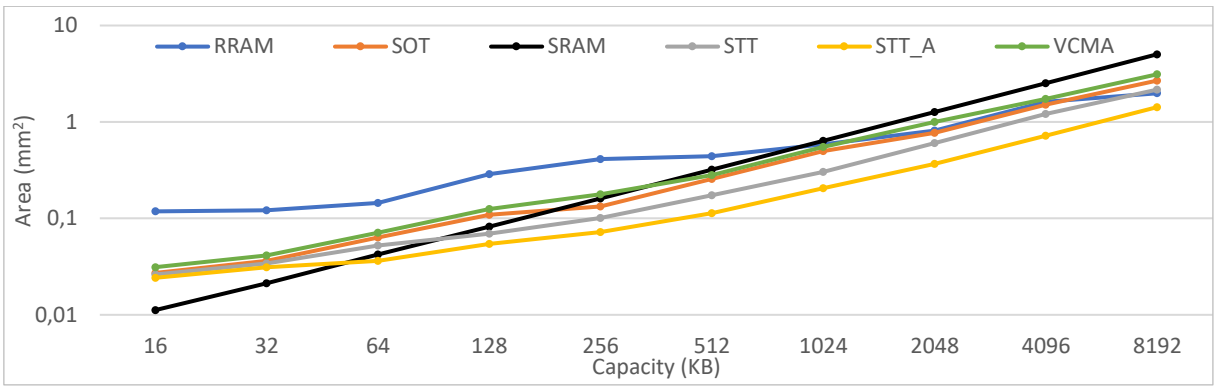
Table 3. Memory cell model parameters

Feature	SRAM	RRAM [6, 41]	STT MRAM [5, 30, 31]	STT-A [16]	VCMA [17-19]	SOT MRAM [20, 22, 24]
Cell area	146 F^2	4 F^2	54 F^2	34.5 F^2	69 F^2	57.5 F^2
R_P	Φ	100K	3000	6000	3500	16500
R_{AP}	Φ	10M	6000	12000	7500	32400
Read current/Voltage	Φ	0.4 V	12.5 µA	12.5 µA	18.5 µA	54.02 µA
Write current/Voltage	Φ	2.0 V	80 µA	40.82 µA	25 µA	54.02 µA
Write pulse	Φ	10 ns	10 ns	5 ns	3.3 ns	1 ns
Access CMOS width	1.31 F	Φ	8 F	4.5 F	8.5 F	8.5 F

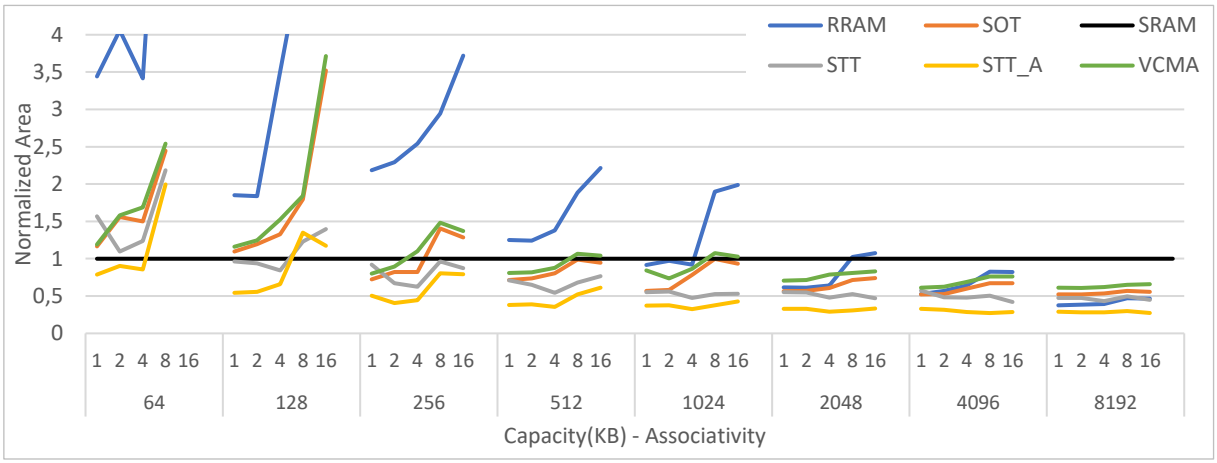
5 Performance evaluation of non-volatile memories

The analysis uncovers a clear trend in memory technology scalability. As depicted in Fig. 4a, it is apparent that with expanding memory capacity, the area required by all memory technologies also grows, showing a nearly proportional relationship. Notably, when compared to NVMs, SRAM exhibits a more rapid increase in area footprint. An intriguing observation is that MRAM technologies generally occupy less space beyond 128 KB capacity, except for RRAM. Among the NVMs, on average STT-A MRAM has the smallest footprint, followed by STT-MRAM and SOT-MRAM. Figure 4b shows that increasing associativity results in varying area consumption, with RRAM demonstrating the most significant increase. For A=1, 64KB STT-A cache saves area by 20%, further for a 256KB cache SOT-MRAM and STT-A saves 27.7% and 49.3% area respectively compared to SRAM. Similarly for larger caches NVMs tend to save area for a give capacity, with savings decrease with increase in associativity depicted in Fig. 4b.

It is worth noting that NVMs have smaller footprints overall, as detailed in Tab. 3, but their peripheral circuits tend to be larger in size pertaining to a modified sensing methods for smaller capacities. This suggests that memory scalability of MRAM technologies is constrained by the peripheral circuitry for small caches, while CMOS footprint size becomes a limiting factor for SRAM.



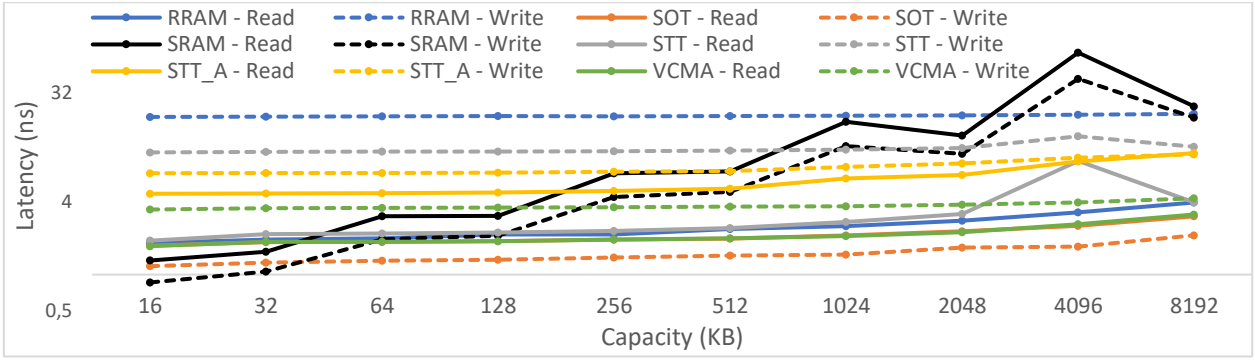
(a)



(b)

Fig. 4. (a) Area consumption – mm^2, (b) Normalized area – SRAM baseline

Figure 5 reveals an interesting relationship between access latencies in SRAM and NVMs across varying cache capacity and associativity. Notably, SRAM experiences a more rapid increase in access time as cache capacity grows compared to all NVMs. This phenomenon primarily arises due to the higher bit-line capacitance associated with larger 6-T SRAM cells compared to NVM memory cells, resulting in an accelerated access time increase in SRAM as memory array size expands. Above the 64 KB capacity threshold, it is worth noting that all NVMs consistently outperform SRAM in read times, except for STT-A MRAM. While most NVMs suffer with extended write times, SOT-MRAM emerges as a promising alternative to SRAM for larger caches, offering competitive read and write times. Furthermore, NVMs exhibit scalability in associativity with minor variations, as caches with higher associativity leverage parallel utilization of peripheral circuitry and cache layout adjustments to mitigate performance impacts.



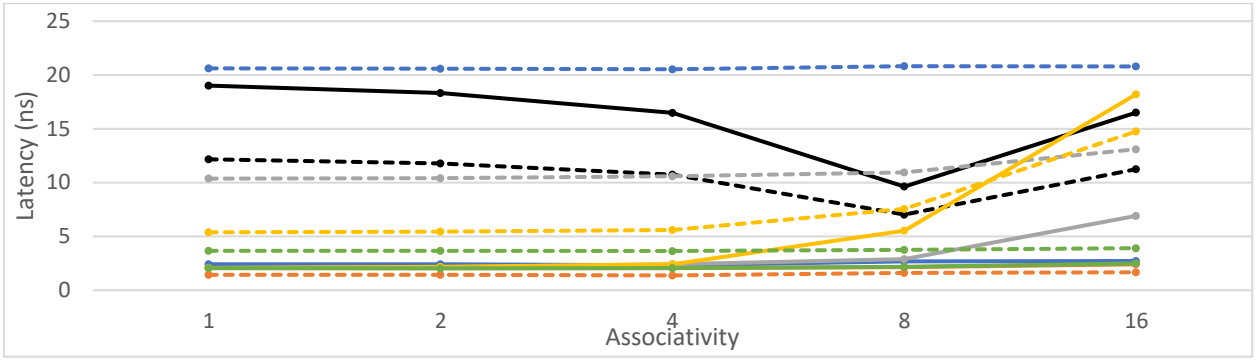


Fig. 5. Read/Write access latency

Dynamic energy, also called access energy: the energy consumed per access is depicted in Fig. 6. Like access latency, access energy increases more rapidly for SRAM as capacity scales up. When it comes to write energy, NVMs consistently require more energy than SRAM. However, for read operations, SRAM stands out as the preferable choice for smaller cache sizes. Notably, SOT and STT-A MRAM offer the best dynamic energy performance among NVMs Interestingly, when capacity remains constant but associativity is increased, NVMs experience a swifter rise in access energy. Notably, write energy increases for SOT and VCMA, while it decreases for STT and STT-A with associativity scaling (see Fig. 6b). This unexpected enhancement in write energy efficiency for STT and STT-A with higher associativity suggests the potential for deploying NVMs in lower levels of the memory hierarchy. Nevertheless, the observed behavior raises questions about the consistency and reliability of NVSIM for academic research purposes.

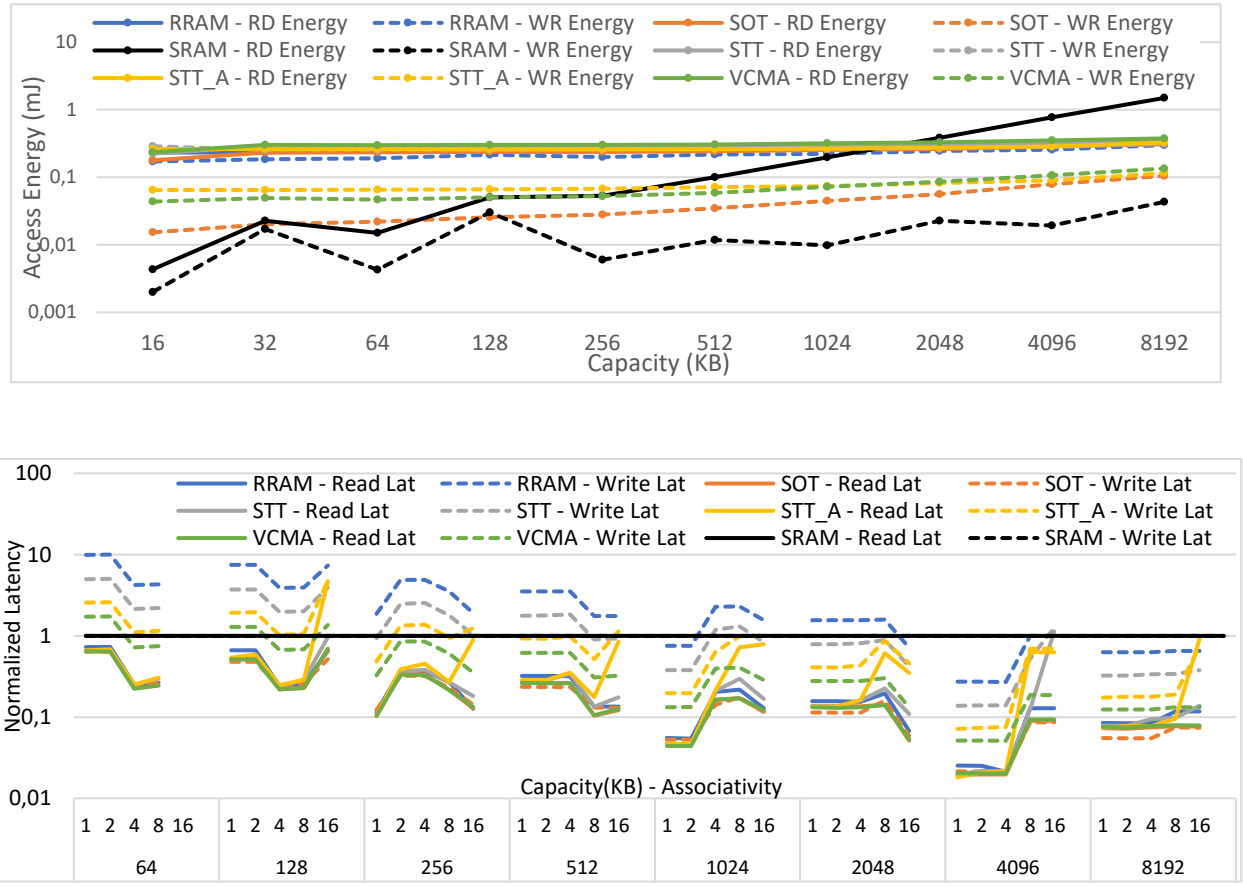
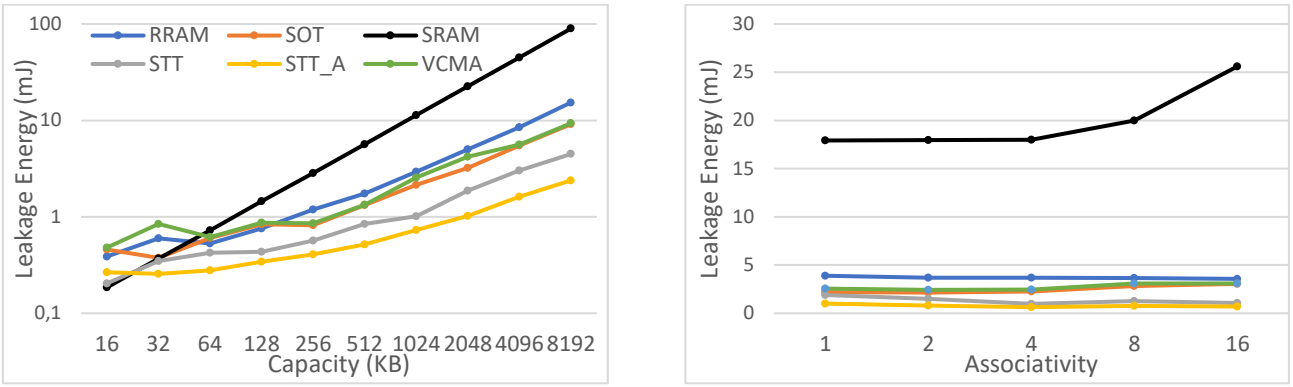


Fig. 6. Access energy consumption

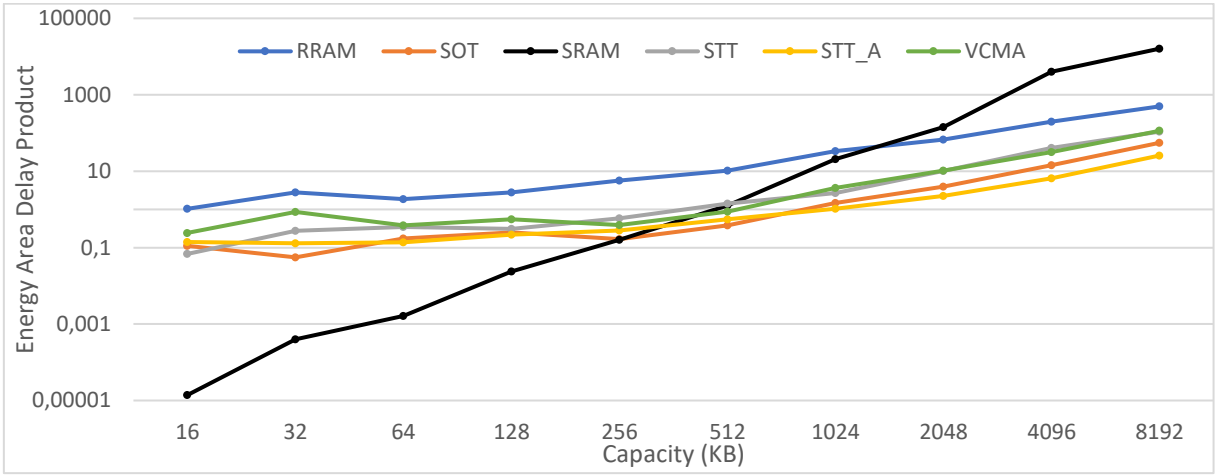
The unique performance feature among NVMs is their leakage energy, illustrated in Fig. 7. In SRAM, the accumulation of CMOS components results in a steep exponential rise in leakage energy. In contrast, MRAMs employing MTJs (Magnetic Tunnel Junctions) exhibit virtually no leakage. It is important to note that the leakage energy is primarily caused by peripheral circuitry and increases with capacity and associativity. The data clearly demonstrates that among NVMs, STT and STT-A MRAMs have the lowest leakage levels.



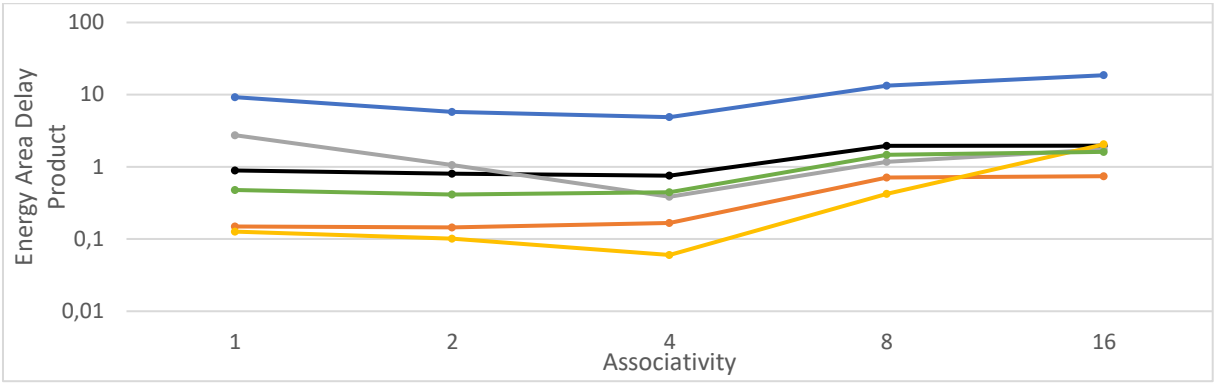
(a) (b)

Fig. 7. Average leakage energy consumption across (a) capacity, and (b) associativity

The Energy-Area-Delay-Product (EADP) serves as a robust metric for comparing various technologies and assessing NVSIM cache optimizations [11]. In Fig. 8, we present the EADP over different cache capacity and associativity. The results are quite evident: SRAM exhibits the most favourable balance of energy, area, and latency for smaller caches. However, as cache sizes exceed 256KB, NVMs take the lead in offering superior EADP, with STT-A and SOT MRAM consistently emerging as the optimal choices. Notably, there is a decrease in EADP as associativity (A) increases, maintaining this trend until A=4, after which it starts to rise. This behaviour aligns with the dynamic energy variations illustrated in Fig. 6 and is a significant factor contributing to these trends.



(a)



(b)

Fig. 8. Energy-Area-Delay-Product (EADP) for different (a) capacity and (b) associativity

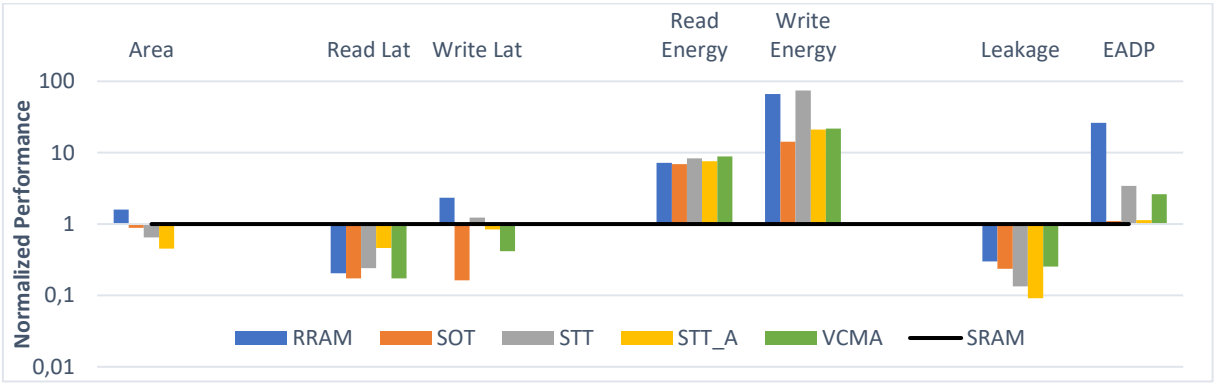


Fig. 9. Comparing non-volatile caches normalized average performance with SRAM as baseline

In summary, the analysis reveals distinct characteristics across various memory technologies and cache configurations. SRAM excels in smaller caches, offering lower latency, and EADP. However, NVMs, particularly STT-A and SOT MRAMs, outperform SRAM in larger caches, showing superior EADP. Leakage energy is a notable strength of NVMs, with MRAMs showing minimal leakage. Dynamic energy considerations favor SRAM for read operations in smaller caches, while NVMs perform well in write operations for large and associative caches. Notably, associativity primarily impacts the performance of caches but it is more pronounced in smaller capacities. Figure 9 summarizes the relative performance of NVMs normalized to SRAM for cache sizes > 64KB.

6 Conclusion

This work explores the modern non-volatile memory technologies – STT, SOT, STT-A, VCMA and RRAM for on-chip cache memories. The experiments evaluate the performance comparison of SRAM with NVMs across different cache capacity and associativity to identify the usage space of NVMs in memory hierarchy. For L1 cache implementations (size < 128KB), SRAM remains a sweet spot offering unparalleled energy efficiency, compact area utilization, and minimal latency. However, as cache sizes expand, NVMs exhibit notable performance enhancements. The influence of associativity on cache performance becomes less pronounced in larger caches. For larger L2 caches, STT-A and SOT MRAM outperform SRAM demonstrating superior EADP and leakage characteristics. RRAM and VCMA suffer from high write latency and write energy. Notably, STT and STT-A MRAMs excel in leakage mitigation among NVMs. Our analysis highlights the significance of NVMs, particularly SOT, STT-A, and VCMA, which demonstrate substantial EADP improvements of approximately 73.7%, 67.8%, and 35.7%, respectively, for a 512KB cache. These findings provide valuable guidance for informed memory hierarchy decisions customized to specific application prerequisites and energy constraint.

References

[1] "More Moore," *Int. Technol. Roadmap Semicond.*, pp. 1–52, 2015.

[2] S. Senni, L. Torres, G. Sassatelli, A. Gamatie, and B. Mussard, "Exploring MRAM Technologies for Energy Efficient Systems-On-Chip," *IEEE J. Emerg. Sel. Top. Circuits Syst.*, vol. 6, no. 3, pp. 279–292, 2016.

[3] H. S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nat. Nanotechnol.*, vol. 10, no. 3, pp. 191–194, 2015.

[4] S. Mittal and J. S. Vetter, "A Survey of Software Techniques for Using Non-Volatile Memories for Storage and Main Memory Systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 5, pp. 1537–1550, 2016.

[5] Z. Guo, K. Cao, K. Shi, and W. Zhao, "Ultra-low power consumption spintronics devices," in *2019 IEEE 13th International Conference on ASIC (ASICON)*, 2019, pp. 1–4.

[6] T. C. Chang, K. C. Chang, T. M. Tsai, T. J. Chu, and S. M. Sze, "Resistance random access memory," *Mater. Today*, vol. 19, no. 5, pp. 254–264, Jun. 2016.

[7] F. Oboril, R. Bishnoi, M. Ebrahimi, and M. B. Tahoori, "Evaluation of hybrid memory technologies using SOT-MRAM for on-chip cache hierarchy," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 34, no. 3, pp. 367–380, 2015.

[8] M. P. Komalan, C. Tenllado, J. I. G. Perez, F. T. Fernandez, and F. Catthoor, "System level exploration of a STT-MRAM based level 1 data-cache," in *Proceedings -Design, Automation and Test in Europe, DATE*, 2015, vol. 2015-April, pp. 1311–1316.

[9] Z. Zhang, W. Wang, P. Yu, and Y. Jiang, "Cache performance of NV-STT-MRAM with scale effect and comparison with SRAM," *Int. J. Electron.*, vol. 109, no. 3, pp. 391–409, 2022.

[10] R. Saha, Y. P. Pundir, and P. Kumar Pal, "Comparative analysis of STT and SOT based MRAMs for last level caches," *J. Magn. Magn. Mater.*, vol. 551, p. 169161, Jun. 2022.

[11] T. Marinelli, J. I. G. Pérez, C. Tenllado, M. Komalan, M. Gupta, and F. Catthoor, "Microarchitectural Exploration of STT-MRAM Last-level Cache Parameters for Energy-efficient Devices," *ACM Trans. Embed. Comput. Syst.*, vol. 21, no. 1, pp. 1–20, 2022.

[12] X. Dong, N. P. Jouppi, and Y. Xie, "A circuit-architecture co-optimization framework for evaluating emerging memory hierarchies," *ISPASS 2013 - IEEE Int. Symp. Perform. Anal. Syst. Softw.*, pp. 140–141, 2013.

[13] S. Mittal and J. S. Vetter, "AYUSH: A Technique for Extending Lifetime of SRAM-NVM Hybrid Caches," *IEEE Comput. Archit. Lett.*, vol. 14, no. 2, pp. 115–118, Jul. 2015.

[14] R. Andrawis, A. Jaiswal, and K. Roy, "Design and Comparative Analysis of Spintronic Memories Based on Current and Voltage Driven Switching," *IEEE Trans. Electron Devices*, vol. 65, no. 7, pp. 2682–2693, Jul. 2018.

[15] W. Kang, Y. Ran, Y. Zhang, W. Lv, and W. Zhao, "Modeling and Exploration of the Voltage-Controlled Magnetic Anisotropy Effect for the Next-Generation Low-Power and High-Speed MRAM Applications," *IEEE Trans. Nanotechnol.*, vol. 16, no. 3, pp. 387–395, 2017.

[16] X. Dong, C. Xu, N. Jouppi, and Y. Xie, "NVSim: A circuit-level performance, energy, and area model for emerging non-volatile memory," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 31, no. 7, pp. 994–1007, 2012.

[17] W. Kang, Y. Ran, Y. Zhang, W. Lv, and W. Zhao, "Modeling and Exploration of the Voltage-Controlled Magnetic Anisotropy Effect for the Next-Generation Low-Power and High-Speed MRAM Applications," *IEEE Trans. Nanotechnol.*, vol. 16, no. 3, pp. 387–395, May 2017.

[18] S. Shreya and B. K. Kaushik, "Modeling of Voltage-Controlled Spin-Orbit Torque MRAM for Multilevel Switching Application," *IEEE Trans. Electron Devices*, vol. 67, no. 1, pp. 90–98, Jan. 2020.

[19] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi, "CACTI 6.0: A Tool to Model Large Caches," *Symp. A Q. J. Mod. Foreign Lit.*, vol. HPL-2009-8, no. HPL-2009-85, pp. 0–24, 2009.

[20] I. Singh, B. Raj, M. Khosla, and B. K. Kaushik, "Comparative Analysis of Spintronic Memories for Low Power on-chip Caches," *SPIN*, vol. 10, no. 04, p. 2050027, Dec. 2020.

[21] N. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, 4th ed. USA: Addison-Wesley Publishing Company, 2010.

[22] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. B. Tahoori, "Architectural aspects in design and analysis of SOT-based memories," *Proc. Asia South Pacific Des. Autom. Conf. ASP-DAC*, pp. 700–707, 2014.

[23] T. Wang, J. Q. Xiao, and X. Fan, "Spin-Orbit Torques in Metallic Magnetic Multilayers: Challenges and New Opportunities," *Spin*, vol. 7, no. 3. p. 1740013, 2017.

[24] R. Ramaswamy, J. M. Lee, K. Cai, and H. Yang, "Recent advances in spin-orbit torques: Moving towards device applications," *Appl. Phys. Rev.*, vol. 5, no. 3, p. 031107, Aug. 2018.

[25] H. D. Kallinatha, S. Rai, and B. Talawar, "A Detailed Study of SOT-MRAM as an Alternative to DRAM Primary Memory in Multi-Core Environment," *IEEE Access*, vol. 12, pp. 7224–7243, 2024.

[26] X. Xu, H. Zhang, C. Jiang, J. Li, S. Lu, Y. Li, H. Du, X. Zhang, Z. Wang, K. Cao, *et al.*, "Full reliability characterization of three-terminal SOT-MTJ devices and corresponding arrays," *IEEE Int. Reliab. Phys. Symp. Proc.*, vol. 2023-March, 2023.

[27] Y. Seo and K. W. Kwon, "Ultra High-Density SOT-MRAM Design for Last-Level On-Chip Cache Application," *Electronics*, vol. 12, no. 20, Oct. 2023.

[28] K. Tian, B. Wu, K. Chen, and W. Liu, "High-Performance Co-Processing Architecture Using SOT-MRAM-Based In-memory Computing Scheme," *Int. Symp. Circuits Syst.*, 2025.

[29] P. Kumar, D. E. Shim, and A. Naeemi, "Comprehensive Device to System Co-Design for SOT-MRAM at the 7 nm Node," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, 2025.

[30] A. Sura and V. Nehra, "Performance Comparison of Single Level STT and SOT MRAM Cells for Cache Applications," *2021 25th Int. Symp. VLSI Des. Test, VDAT 2021*, pp. 1–4, Sep. 2021.

[31] G. Yu, "Two-terminal MRAM with a spin," *Nat. Electron.*, vol. 1, no. 9, pp. 496–497, Sep. 2018.

[32] D. Mondal, A. Singh, S. Bhatt, and R. Mishra, "Hybrid Spin-Orbit Torque/Spin-Transfer Torque-Based Multibit Cell for Area-Efficient Magnetic Random Access Memory," *IEEE Trans. Electron Devices*, vol. 70, no. 12, pp. 6318–6323, Dec. 2023.

[33] S. Han, Q. Wang, and Y. Jiang, "MRAM-based Cache System Design and Policy Optimization for RISC-V Multi-core CPUs," *IEEE Trans. Magn.*, pp. 1–14, 2023.

[34] S. Han, Q. Wang, and Y. Jiang, "Hierarchical cache configuration based on hybrid SOT- and STT-MRAM," *AIP Adv.*, vol. 13, no. 2, Feb. 2023.

[35] S. Zou, X. Zhao, Y. Xue, J. Gao, Y. Cui, and J. Luo, "Extremely Low Switching Current STT-MRAM Device With Double Spin Transfer Torque," *IEEE Electron Device Lett.*, vol. 46, no. 4, pp. 584–587, 2025.

[36] L. Chang, Z. Wang, Y. Gao, W. Kang, Y. Zhang, and W. Zhao, "Evaluation of spin-Hall-assisted STT-MRAM for cache replacement," *Nanoscale Archit. (NANOARCH), 2016 IEEE/ACM Int. Symp.*, pp. 73–78, 2016.

[37] D. Ielmini and G. Pedretti, "Resistive Switching Random-Access Memory (RRAM): Applications and Requirements for Memory and Computing," *Chem. Rev.*, vol. 125, no. 12, pp. 5584–5625, Jun. 2025.

[38] S. Kang, W. Y. Cho, B. H. Cho, K. J. Lee, C. S. Lee, H. R. Oh, B. G. Choi, Q. Wang, H. J. Kim, M. H. Park, *et al.*, "A 0.1-μm 1.8-V 256-Mb Phase-change Random Access Memory (PRAM) with 66-MHz synchronous burst-read operation," *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 210–216, 2007.

[39] S. Raoux, Y.-C. Chen, G. W. Burr, D. Krebs, R. M. Shelby, S.-H. Chen, H.-L. Lung, C. H. Lam, C. T. Rettner, M. J. Breitwisch, and M. Salinga, "Phase-change random access memory: A scalable technology," *IBM J. Res. Dev.*, vol. 52, no. 4.5, pp. 465–479, 2010.

[40] Y. Li and K. N. Quader, "NAND Flash memory: Challenges and opportunities," *Computer (Long. Beach. Calif).*, vol. 46, no. 8, pp. 23–29, 2013.

[41] D. S. Jeong, R. Thomas, R. S. Katiyar, J. F. Scott, H. Kohlstedt, A. Petraru, and C. S. Hwang, "Emerging memories: resistive switching mechanisms and current status," *Reports Prog. Phys.*, vol. 75, no. 7, p. 76502, 2012.

Received 20 October 2025