2023 IEEE International Symposium on Workload Characterization (IISWC) | 979-8-3503-0317-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IISWC59245.2023.00022

NVMSurvey: Recent Advances and Comparative Analysis of Emerging Non-Volatile Memories (eNVMs)

Sandesh Ghimire* Shinsaku Kataoka* Lillian Pentecost

Amherst College

*These authors contributed equally to this work.

Abstract

Emerging non-volatile memories, specifically FeFET, PCM, RRAM, and STT-MRAM, promise significant advances in energy-efficient on-chip storage. Our research evaluates most recent (2020-2023) publications of such eNVMs against both optimistic and pessimistic projections sourced from 2016 to 2020 research using NVMExplorer, an open-source eNVMs simulator. Significant findings include FeFET's low area, PCM's improved energy efficiency and reduced read latency, and the evident advancements in STT-MRAM and RRAM. Moreover, we evaluate the potential benefits of recent eNVM solutions as memory resources for Deep Neural Network (DNN) accelerators and find that recent advances in RRAM and FeFET devices offer improved memory power and density for ResNet26 image processing. Overall, our study presents and analyzes continued research efforts on eNVM technologies as promising contenders to augment and replace conventional memory technologies.

1. Introduction

The exploration of emerging non-volatile memory technologies like FeFET, PCM, RRAM, and STT-MRAM, over the last three years offers a glimpse into the continuing evolution and innovation of memory technologies. This study dives deep into their performance and their potential to supplant conventional memories such as SRAM and DRAM by modeling and evaluating recent advances against standards from prior work based on 2016-2020 publications. Using DNN benchmarks, we also assess how these eNVMs can be optimized to improve the power efficiency of modern computing tasks.

2. Methodology

We reviewed eNVM-focused papers from ISSCC, IEDM, and VLSI (2020-2023), extracting cell configuration data. Based on the obtained data, we used NVMExplorer [10] to model array-level characteristics. We simulated eNVMs against six key optimization criteria mainly focused on energy, density, and power optimization. We then analyzed key metrics: Read Latency vs. Read Energy, Write Latency vs. Write Energy, and Area vs. Area Efficiency. Key findings on standout eNVMs are discussed in Section 3.1. We validated our simulations by comparing them against published values of fabricated test chips (see Section 3.3). Additionally, we evaluted applicability to DNN use cases using ResNet26, each storing either weights only or both weights and activations, at a 2 MB memory size; detailed results are presented in Section 3.2.

3. Results

3.1. Memory Array Characterization

We evaluated key metrics across technologies at iso-capacity (1MB) using published cell characteristics as inputs to NVMExplorer. We compared these results from 2020-2023 publications to NVMExplorer's provided optimistic and pessimistic projections per technology based on 2016-2020 publications. Our analysis highlights the top-performing technologies across several optimization criteria. Our data shows that recent examples of FeFET outperforms pre-existing pessimistic cell assumptions in area, generally achieving array characteristics close to prior optimistic assumptions, as depicted in Figure 1 (a). As per Figure 1 (b), recent advances in PCM surpass even optimistic prior characteristics, demonstrating reduced read latency and lower energy consumption. Meanwhile, Figure 1 (c) underscores the recent progress in STT-MRAM, showing examples of increased energy efficiency and quicker write times, consistently aligning with optimistic expectations.

3.2. DNN Inference Simulation

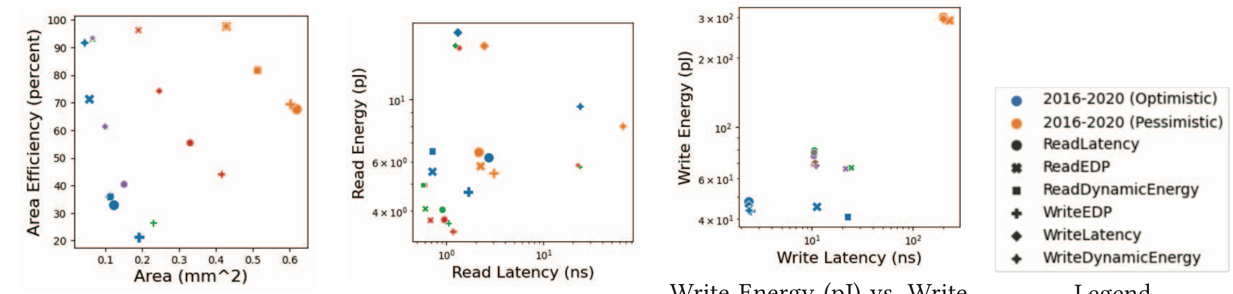
To evaluate recent eNVM configurations as potential replacements to SRAM in the deep learning accelerator architecture studied in [10], we evaluate 2MB capacity eNVM arrays under memory traffic patterns corresponding to a variety of ResNet26 image processing tasks.

Figure 2 presents the results of DNN inference simulations in which only weights are stored in eNVM array (2020-2023), leading to a read-only traffic during inference. We excluded those points that failed to meet software requirements (completing 60 frames-per-second image processing).

Table 1 compares the optimal eNVMs choices for three other DNN benchmarks based on DNN inference simulation results from 2016-2020 and 2020-2023 publications. It is noteworthy that FeFET has become the optimal choice for all benchmarks in terms of area. PCM, on the other hand, still performs the best for weights-only benchmarks, while RRAM has become the top performer for weights-and-activation benchmarks, surpassing STT.

2835-2238/23/$31.00 ©2023 IEEE
DOI 10.1109/IISWC59245.2023.00022

Authorized licensed use limited to: Amherst College. Downloaded on January 13,2025 at 19:03:05 UTC from IEEE Xplore. Restrictions apply.



Area Efficiency (%) vs. Area (mm^2), FeFET. Green: [8], Red: [6], Purple: [9]

Read Energy (pJ) vs. Read Latency (ns), PCM. Green: [2],Red: [5]

Write Energy (pJ) vs. Write Latency (ns), STT. Green: [1], Red: [4], Purple: [3], Brown: [7], Pink: [12], Grey: [11]

Legend

Figure 1: Example Array Characteristics for 1MB eNVMs using cell characteristics derived from 2020-2023 publications compared to existing optimistic / pessimistic example configurations from [10]

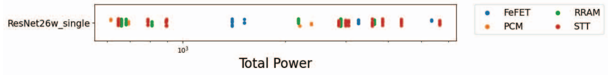


Figure 2: Total Power of eNVM-based weights memory for ResNet26 single-task image classification (2020-2023 data)

TABLE 1: Optimal eNVMs for each benchmark

Operating Mode	Priority	2016-2020	2020-2023
Single (Weights only)	Power	PCM	PCM
Single (Weights only)	Area	FeFET	FeFET
Single (Weights & Acts)	Power	PCM	**RRAM**
Single (Weights & Acts)	Area	FeFET	FeFET
Multiple (Weights only)	Power	PCM	PCM
Multiple (Weights only)	Area	STT	**FeFET**
Multiple (Weights & Acts)	Power	STT	**RRAM**
Multiple (Weights & Acts)	Area	STT	**FeFET**

4. Validation

We conducted a validation study on PCM [5] and STT-MRAM [1] memory chips, using available data from the literature at both device and array levels. We simulated array-level performance, based on provided cell parameters with NVMExplorer. The reported values consistently fell within our simulated range as depicted in Figure 3, validating that NVMExplorer's modeling capabilities effectively capture behaviors akin to those of fabricated memory arrays.

5. Conclusion

Our study advocates eNVMs as potential successors to conventional memory systems. Simulations reveal FeFET's superiority in area, power efficiency, and DNN application, while PCM, STT-MRAM, and RRAM show improved energy, latency, and reduced DNN inference power respectively.

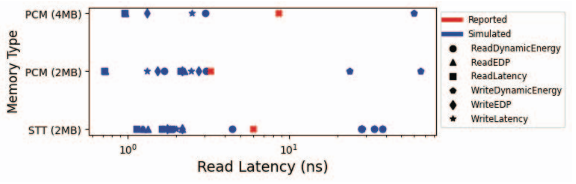


Figure 3: Simulated and Reported Results (Validation)

References

[1] H. Cai, Z. Bian, Y. Hou, Y. Zhou, J. Cui, Y. Guo, X. Tian, B. Liu, X. Si, Z. Wang, J. Yang, and W. Shan, "33.4 a 28nm 2mb stt-mram computing-in-memory macro with a refined bit-cell and 22.4 - 41.5tops/w for ai inference," in *ISSCC*, 2023.

[2] F. Disegni, A. Ventre, A. Molgora, P. Cappelletti, R. Badalamenti, P. Ferreira, G. Castagna, A. Cathelin, A. Gandolfo, A. Redaelli, D. Manfrè, A. Maurelli, C. Torti, F. Piazza, M. Carfì, F. Arnaud, M. Perroni, M. Caruso, S. Pezzini, R. Annunziata, G. Piazza, O. Weber, and M. Peri, "16mb high density embedded pcm macrocell for automotive-grade microcontroller in 28nm fd-soi, featuring extension to 24mb for over-the-air software update," in *VLSI*, 2021.

[3] G. Hu, G. Lauer, J. Sun, P. Hashemi, C. Safranski, S. Brown, L. Buzi, E. Edwards, C. D'Emic, E. Galligan, M. Gottwald, O. Gunawan, H. Jung, J. Kim, K. Latzko, J. Nowak, P. Trouilloud, S. Zare, and D. Worledge, "2x reduction of stt-mram switching current using double spin-torque magnetic tunnel junction," in *IEDM*, 2021.

[4] T. Ito, T. Saito, Y. Taito, K. Sonoda, G. Watanabe, K. Matsubara, A. Kanda, T. Shimoi, K. Takeda, and T. Kono, "A 20mb embedded stt-mram array achieving 72logic process," in *IEDM*, 2021.

[5] W. Khwa, Y. Chiu, C. Jhang, S. Huang, C. Lee, T. Wen, F. Chang, S. Yu, T. Lee, and M. Chang, "A 40-nm, 2m-cell, 8b-precision, hybrid slc-mlc pcm computing-in-memory macro with 20.5 - 65.0tops/w for tiny-al edge devices," in *ISSCC*, 2022.

[6] S. Kuk, S. Han, B. Kim, S. Baek, J. Han, and S. Kim, "Comprehensive understanding of the hzo-based n/pfefet operation and device performance enhancement strategy," in *IEDM*, 2021.

[7] P. Lee, C. Lee, Y. Shih, H. Lin, Y. Chang, C. Lu, Y. Chen, C. Lo, C. Chen, C. Kuo, T. Chou, C. Wang, J. Wu, R. Wang, H. Chuang, Y. Wang, Y. Chih, and T. Chang, "33.1 a 16nm 32mb embedded stt-mram with a 6ns read-access time, a 1m-cycle write endurance, 20-year retention at 150°c and mtj-otp solutions for magnetic immunity," in *ISSCC*, 2023.

[8] Z. Liang, K. Tang, J. Dong, Q. Li, Y. Zhou, R. Zhu, Y. Wu, D. Han, and R. Huang, "A novel high-endurance fefet memory device based on zro2 anti-ferroelectric and igzo channel," in *IEDM*, 2021.

[9] Z. Lin, M. Si, Y.-C. Luo, X. Lyu, A. Charnas, Z. Chen, Z. Yu, W. Tsai, P. C. McIntyre, R. Kanjolia, M. Moinpour, S. Yu, and P. D. Ye, "High-peformance beol-compatible atomic-layer-deposited in2o3 fe-fets enabled by channel length scaling down to 7 nm: Achieving performance enhancement with large memory window of 2.2 v, long retention > 10 years and high endurance > 108 cycles," in *IEDM*, 2021.

Authorized licensed use limited to: Amherst College. Downloaded on January 13,2025 at 19:03:05 UTC from IEEE Xplore. Restrictions apply.

[10] L. Pentecost, A. Hankin, M. Donato, M. Hempstead, G. Wei, and D. Brooks, “Nvmexplorer: A framework for cross-stack comparisons of embedded non-volatile memories,” in *HPCA*, 2022.

[11] T. Shimoi, K. Matsubara, T. Saito, T. Ogawa, Y. Taito, Y. Kaneda, M. Izuna, K. Takeda, H. Mitani, T. Ito, and T. Kono, “A 22nm 32mb embedded stt-mram macro achieving 5.9ns random read access and 5.8mb/s write throughput at up to tj of 150 °c,” in *VLSI*, 2022.

[12] Z. Wei, W. Kim, Z. Wang, L. Hu, D. Jung, J. Zhang, and Y. Huai, “Accurate and fast stt-mram endurance evaluation using a novel metric for asymmetric bipolar stress and deep learning,” in *VLSI*, 2022.

Authorized licensed use limited to: Amherst College. Downloaded on January 13,2025 at 19:03:05 UTC from IEEE Xplore. Restrictions apply.