

Equivalent-Ons-Replacement Self-Aware-Access LLC on Dual-Port SOT-MRAM by Sense-While-Replace

Keyang Zhang, Quanhai Zhu, Zhenghan Fang, Shuyu Wang, Hao Cai*
School of Integrated Circuits, Southeast University, Nanjing 210096, China.
Email: hao.cai@seu.edu.cn

Abstract—Last-Level Cache (LLC) is increasingly required to be energy-efficient and area-saving. Emerging Non-volatile memory (NVM), such as Magnetic-resistive Random Access Memory (MRAM), present potential solutions for LLC as its ultra-low leakage and area. However, the high replacement-latency caused by its high write-latency and power consumption hinders MRAM in LLC applications. Thus, this paper proposes a novel Sense-While-Replace (SWR) strategy for dual-port SOT-MRAM, which liberates the conflict between reading and writing to conceal the impact of high write-latency on system performance. Furthermore, Self-aware access circuits are proposed, which accelerate reading and obtain utmost writing-energy saving. Under 40-nm CMOS technology, the 4Kb Macro achieves $<3\text{ns}@32\text{bits}$ read and $<75\%$ energy-saving. Most crucially, SWR supports CPU continuously read whereas preserved from replacement latency, which improves performance by up to 8% even compared to SRAM.

Index Terms—LLC, SOT-MRAM, Dual-Port, Sense-While-Replace, Self-Aware access

I. INTRODUCTION

The performance difference between memory and CPU increases gradually, causing memory determining processor performance as “Memory Wall”. Considering capacity, access speed and cost of all kinds of memory comprehensively, the current memory hierarchy presents a pyramid structure. Cache plays an important role in memory systems, including three levels commonly. For multi-core processors, each core has its own cache, such as L1(Data/Instruction) cache and L2 cache whereas the LLC is shared by all cores. The number of processor cores has increased with the higher demand for computing power nowadays. Thus, the capacity requirement for LLC also rise accordingly, which leads into larger leakage power and area consumption for LLC based on SRAM.

To solve these challenges, there are some emerging non-volatile memory, such as eDRAM and MRAM, as alternatives to SRAM[1-9]. Although eDRAM has advantages in density, speed and access power, it is limited by its necessary refresh operation. Meanwhile, Spin-Transfer-Torque MRAM (STT-MRAM) also has higher density than SRAM and almost zero static power consumption, which is regarded as candidate to SRAM. However, the switching distribution of STT-MRAM has long-tail effect, so the pulse width of writing is often much longer than actually required. To sum up, the long writing pulse width and large write-energy hinders STT-MRAM becoming qualified candidate for write-intensive scenarios like LLC.

Spin-Orbit-Torque MRAM (SOT-MRAM) inherits the advantages of STT-MRAM, such as non-volatile, ultra-low-leakage and compatibility of CMOS, etc. Not only that, it also has

competence of faster access speed and immunity to read-disturbance. Thus, it is expected to become the memory medium of the next-generation LLC. The write pulse width and dynamic energy consumption of SOT-MRAM are still worse than those of SRAM, so a feasible solution to these two challenges is urgently needed.

Note that the essence of LLC replacement is the writing of memory. When the replacement of LLC occurs, CPU stops and will not access until the replacement is completed. To avoid such replacement-access conflicts, traditional solution adopt a multi-bank approach, because the conflict of replacement and accesses between different banks do not occur. But each bank needs to be configured with peripheral circuits. Therefore, the number of banks has a trade-off: lower probability of conflicts and more area and leakage overhead.

For SRAM-based LLC, this multi-bank solution can basically eliminate the impact of write latency. However, for memory with long writing pulse width such as MRAM, once the conflict occurs within the same bank, long waiting period of CPU will seriously affect system performance. However, if increasing the number of banks to reduce conflicts, the area and power consumption overhead of the introduced peripheral circuits will counteract the advantages of its low leakage and high density. Thus, we need a novel strategy to avoid the conflict rather than simple increase number of banks.

In this paper, we propose a holistic methodology for substituting conventional LLC made of SRAM with SOT-MRAM to meet the ultra-low leakage and competitive performance.

- To overcome the impact of write pulse width on replacement, we design a 4Kb LLC Testkey based on dual-port SOT-MRAM. Relying on its read-write separation feature, the designed LLC has the ability to read and write simultaneously. While a long pulse width, the desired data can be continuously read even in the same bank. Therefore, the dual-port design conceals the impact of high replacement latency on the performance of the LLC.
- To meet function of access in LLC, we propose a dual-referenced sense-amplifier (SA), increasing read margin to support 32-bits parallel reading and adopts self-aware strategy to latch the desired data to meet high-speed.
- To decrease the write energy consumption, we propose a self-aware parallel writing circuit, which monitors automatically resistance of written bit-cell in real time to prevent write redundancy and consumption waste after writing successfully. It reuses the SA circuit, so no additional area overhead is introduced except for logic control.

This work was supported by the Natural Science Foundation of Jiangsu Province (Grants No. BK20243042).

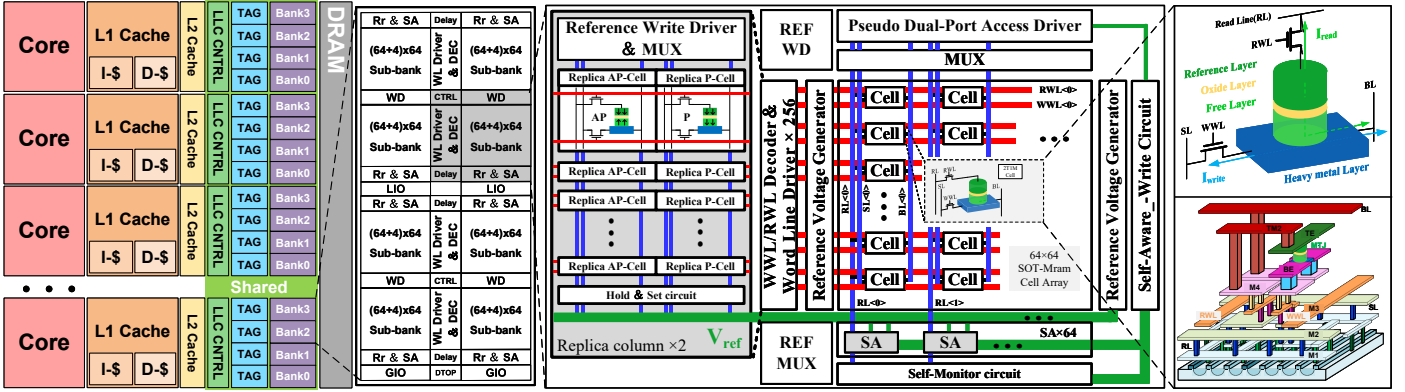


Fig. 1. Overview of the whole Framework, which demonstrates the framework and configuration from system architecture to memory floor-plan. The framework includes: LLC shared by multi-cores, sub-banks and peripheral circuits, bit-cell and corresponding layout.

The remainder of this paper is structured as follows: Section II introduces SOT-MRAM access principle and floorplan of memory array. Section III outlines the design of dual-port SOT-MRAM LLC for SWR operations and peripheral circuit for self-aware parallel access. Section IV presents the circuit simulation and performance evaluation of the proposed scheme. Section V provides this paper's conclusion.

II. DEVICE AND FLOORPLAN

A. Characteristics

Fig. 1 (a) illustrates the schematic of a typical three-terminal SOT-MTJ device. It comprises a perpendicular magnetic tunnel junction (MTJ) consisting of two ferromagnetic (FM) layers separated by an oxide barrier layer and stacked on a heavy metal channel. The fixed layer on the top has a fixed magnetization in one direction whereas the free layer below has reversible magnetization. The write current (ISOT), flowing through the heavy metal layer (HML), induces spin accumulation due to the spin Hall effect (SHE), resulting in torquing the FL's magnetization. The magnetization of the fixed layer and the free layer determine the resistance state of MTJ.

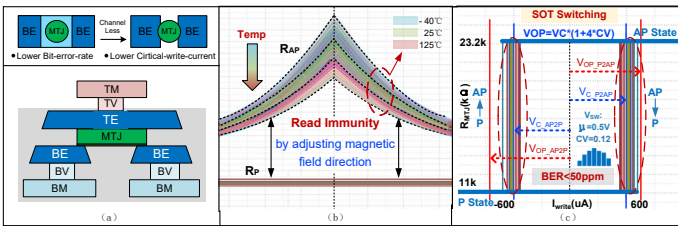


Fig. 2. (a) MTJ device Optimization and Schematic of a single SOT-MTJ device. (b) stable P-State Resistance, AP-State resistance variation in different temperature. (c) Normal distribution of MTJ switching with write voltage

Traditional BSL-type devices are updated to CHL-type devices, which achieve faster write speeds, write yield rates and lower critical switch current. The resistance with applied voltage shows as Fig.1(b). By adjusting the specific magnetic field direction, SOT can be immune to read disturbance, and MTJ can be applied with greater read voltage to meet the parallel high-margin reading requirements of LLC. Fig.1(c) shows the probability distribution of MTJ flipping with voltage.

B. Floorplan

The sub-bank is of $64 \times (64+4)$ specification with 4 columns serving as reference columns. The peripheral circuits mainly

include dual-port access driver, word line (WL) driver, SAs and self-aware circuits. The write voltage and read voltage applied in BL/SL and RL is generated by the access driver. WL driver is used to generate the WL voltage. In addition, we configured 32 SAs for each sub-bank with matching self-aware circuits since LLC reads data out in terms of word. Meanwhile, the 32 SAs share the 4 reference columns, which generate reference voltage. The detailed design shows as Section III-D.

As shown in Fig.1(b), differing from traditional 2T-1SOT structure, we split read-line (RL) from source-line (SL) to support SWR strategy. Separating the RL can also reduce the parasitic capacitance on the SL, accelerate the read operation and reducing the power consumption caused by power-on and power-off on SL. Before the RL is separated, the read-write transistor of the bit-cell can reuse the drain zone. However, if the RL is separated, two additional dummy poly is required to separate the RL from the SL, which approximately increase the area overhead by 37%. Fortunately, CMOS compatibility will gradually assist SOT-MRAM demonstrating an area advantage over SRAM as the device process shrinks. Therefore, area overhead is not the main focus of this design.

To achieve high-yield access, the size of access transistors are designed to be wider, which introduces larger parasitic capacitance on the SL and RL. Considering that the relatively high resistivity of the underlying metal M1, we adopt a parallel connection method between M1 and M2 to manufacture the SL and RL, further reducing their parasitic effects. Then, M4 and TM2 are designed as MTJ's BE and TE.

III. DESIGN

A. SWR strategy on Dual-Port SOT-MRAM

Studies [10-16] have shown that dual-port SOT can achieve simultaneous reading and writing on bit-cells of two different BLs. However, if using dual-port SOT to achieve reading data by CPU and implement replacement of LLC simultaneously, it is very likely that simultaneous reading and writing will occur on the same BL, which will lead into read errors. Y-MUX monitoring was adopted in [10] to control arbiter to prioritize write operations to avoid this situation, which limits the advantage of simultaneous read and write operations. This paper proposes a new solution to achieve SWR at any position within the same bank on the structure of the same bit-cell.

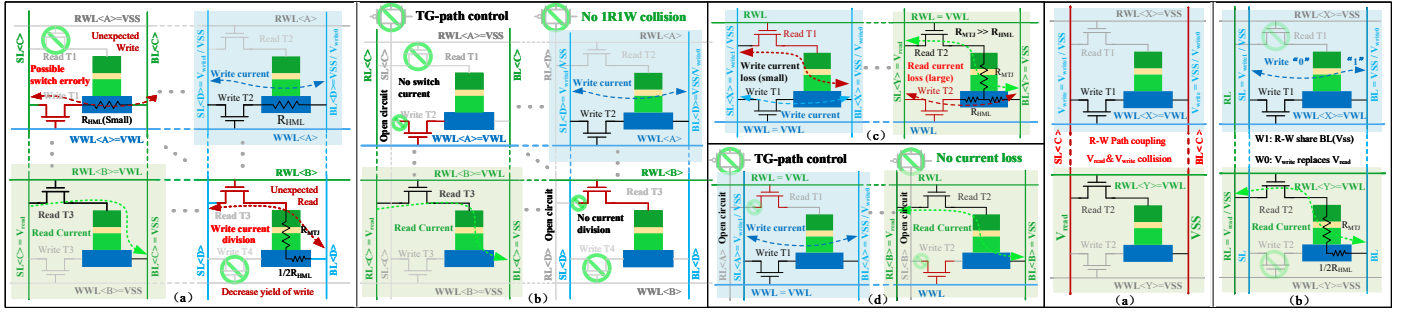


Fig. 3. Three cases within a sub-bank with and without the SWR strategy: (a) (b) Non-aligned Read/Write Cells: Unexpected write and read; (c) (d) Row-aligned Read/Write Cells: Write and read current loss; (e) (f) Column-aligned Read/Write Cell: path coupling and R&W voltage collision.

The proposed architecture comprises a main array and four reference columns, divided into two groups to handle write-1 and write-0 operations respectively. Each group contains a P-column and an AP-column, initialized by a power-on refresh circuit. During RWW operations, reference columns exclusively perform read operations without write current through the HML. The SOT-MTJ effectively prevents read disturbance through external magnetic field control, eliminating STT effects that might alter MTJ resistance states. This inherent read-disturb immunity ensures stable reference column states, requiring only one-time configuration by the refresh circuit.

To elaborate on the SWR Strategy in detail, four situations are enumerated as examples, which may occur when concurrent read-write in the same bank. As Fig.3 shows, the bit-cell with a blue background represent the written bit-cell while the one with a green background represent the read bit-cell.

1) Non-aligned Read/Write Cells:

Considering the most common scenario in the same sub-bank illustrated in the fig.3 (a) where read and write bit cells reside in different rows and columns, concurrent read-write operations will induce unexpected consequences. When $WWL < A >$ is activated, the read voltage applied to the HML of the top-left bit-cell may cause unexpected writes. Meanwhile, $RWL < B >$ activation introduces write voltage across the read path of the bottom-right bit-cell, inevitably degrades write-driving capability. In contrast, the SWR strategy effectively resolves these issues as Fig.3 (b) shows. The top-left bit-cell becomes open-circuited due to the turned-off transmission gate (TG) on SL, preventing current flow through its HML and eliminating erroneous writes. Furthermore, the independent RL ensures that write voltage remains isolated from the read path. This approach prevents write errors while maintaining full write-driving capability during concurrent read-write.

2) Row-aligned Read/Write Cells:

The second scenario involves read and write bit-cells located in the same row as Fig.3 (c). When both two access transistors are simultaneously activated, the MTJ resistance (over $10\times$ greater than that of the HML) ensures that only a minor portion of the write current is shunted. But the read bit-cell experiences significant read current shunting, leading to substantial degradation of read margin. By employing SWR, the read and write paths are isolated through TG disconnection, effectively eliminating mutual interference as Fig3 (d) shows.

3) Column-aligned Read/Write Cells:

The SWR addresses this limitation through two key mechanisms: isolation and reuse. For writing, the disconnection of read access transistors isolates the write bit-cell from read voltage interference, ensuring normal write operations. For reading, read bit-cell operation adapts to two distinct scenarios: When the co-column write bit-cell undergoes '1' programming (BL grounded), the read bit cell reuses BL while applying read voltage through RL. For '0' programming (BL biased high), the read bit-cell reuses the write voltage on BL as its read voltage, requiring only RL grounding. The MTJ resistance in the read path ensures negligible current diversion introduced by the reuse of write voltage on BL. This configuration imposes only minimal impact on write speed ($<5\%$ degradation in our measurements) while enabling concurrent read-write operations.

4) Same-cell Read/Write

The final scenario involves read/write bit-cells located in the same column and row. Due to cache data dependencies, the outdated unupdated data might be read out, potentially leading to subsequent computational errors in the LLC. This paper proposes a solution employing a small-capacity buffer to temporarily store the latest data pending write operations. This approach enables simultaneous writing to SOT-MRAM while reading the most recent data from the buffer, effectively resolving the data consistency issue.

Under the LLC architecture, the granularity of the replacement operation obeys the Cache-line unit, and the read operation accesses 8 bytes. In fact, due to the Temporal Locality and Spatial Locality of the cache, an entire cache-line is prefetched in one or more cycles generally. In concurrent replacement and read operations in LLC, case 1) occur most frequently and inevitably involve case 3) in the same column. There is also a very small probability of case 2) and case 4). A temporary buffer can solve these two situations effectively.

In this paper, since the sub-array uses a 64-bit cache-line, the prototype adopts 4 bytes (32 bits) as minimum of access granularity instead of traditional 8-byte (applicable similarly). Moreover, due to the equivalent zero-delay replacement implemented by SWR, replacement can be applied byte by byte, which greatly reduces the requirement for peripheral write circuits in each sub-array. The problem with doing this is that it is possible that the current replacement has not been completed when the next replacement starts, where there is a trade-off between the write granularity and the temporary buffer capacity.

The sub-array consists of data-array and 4 ref columns.

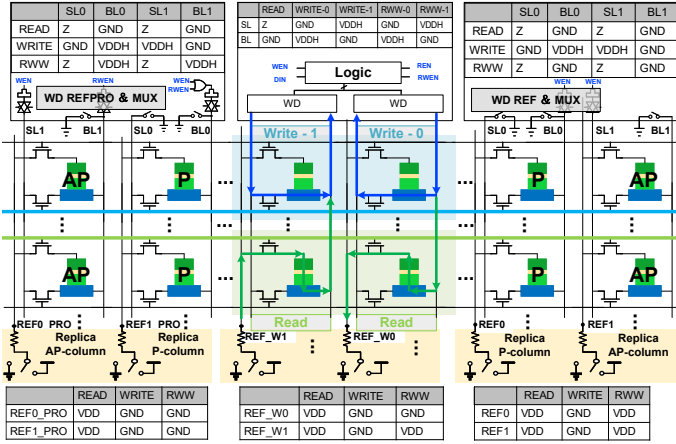


Fig. 4. Configuration of Reference and data-array in different working modes. The ref columns are divided into two groups, and each group has one P reference column and one AP reference column respectively. In addition, there are three working modes: READ (Read-only), WRITE (Write-only), RWW (read-while-write). Write operations include WRITE-0/1 and RWW-0/1. Three Write drivers are responsible for data-array, REF group and REF_PRO group, generating the required voltage as working modes. Each column is equipped with the same resistor, and the connected voltage is controlled by a 2-1 MUX. The overall configuration is as shown in the truth table in Fig.4.

B. Dual-Reference-Reused Reading Scheme

The reading scheme of this paper is 32-bit parallel sensing. To ensure the correct reading of LLC and increase the reading margin, each of 32 SAs configured is set with three inputs: two reference and one data input as Fig.5 shows. To reduce dynamic power consumption of reading, this paper reuses 4 reference branches for 32 SAs. In addition, a self-aware scheme is adopted to accelerate readout. A detailed analysis follows.

Reference resistance in conventional SA is set at $(R_P + R_{AP})/2$, where the effective margin is limited to $(R_{AP} - R_P)/2$. To address this limitation, we adopt dual-reference SAs incorporating both P-state and AP-state references. Leveraging the common-mode rejection of differential amplifiers, this design suppresses voltage differences from the same resistance state. Simultaneously, the cross-coupled positive feedback structure amplifies the voltage difference generated from opposite resistance states. Consequently, the read margin for complementary states is enhanced to approximately $(R_{AP} - R_P)$, theoretically doubling the sensing margin.

Operations feature two distinct modes: In the READ mode, read address is decoded to activate the selected read access transistor (T_r), then read voltage (V_{read}) is applied on both RL and RL_{ref} to generate stable read current through T_r , MTJ, and poly resistor R_r to ground. The voltage division between MTJ and R_r creates three node voltage ($V_{REFP}, V_{REFAP}, V_{DATA}$) as inputs of SAs. The voltage difference from different resistance states is amplified by positive feedback and self-aware circuit triggers latching when output differential exceeds certain threshold. As for the RWW mode, only the V_{read} configuration is different whereas reading process has no difference from the READ mode. Digital logic configures voltage based on column-aligned write data as Fig.4 shows.

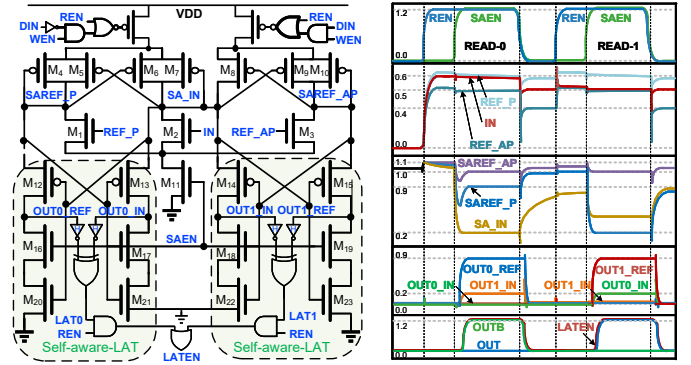


Fig. 5. Dual-reference SA circuit structure and simulation waveform

Conventional current SA [17-20] avoid reference reuse due to parasitic capacitances from the input transistors of SAs. When multiple SAs share one reference branch, the output variations during amplification, inducing slight fluctuations in reference path through parasitic capacitances. With multi SAs sharing one reference, this disturbance is amplified, severely degrading reference voltage stability and read yield. To mitigate the kickback noise, the proposed sensing scheme addresses it by generating reference with dc current. Traditional CSAs employ positive feedback amplification directly on the data and reference paths, causing output varying irreversibly. In contrast, our design decouples the input voltage, allowing disturbances caused by output variations to gradually recover by parasitic capacitances via the reference branch's dc path. Recovery time, determined by the number of shared SAs, remains below 1 ns for 32 SAs sharing 4 reference columns, as verified by array-level simulations, making this overhead acceptable.

Considering the overhead, the area overhead from reference column is amortized to 1/16 per data-column. Furthermore, as these reference columns are integrated within the memory array, their parasitic parameters exhibit absolutely matching to those of the data columns, significantly enhancing tolerance to PVT (Process, Voltage, Temperature) variations, thereby ensuring robust read yield. Regarding read power consumption, the consumption from reference is amortized to 1/8 by reusing 4 reference columns for 32 SAs. Additionally, the read path current remains inherently low due to the high resistance of both the MTJ and R_r . Although the amplification circuits within the SA exhibits higher power consumption, this constitutes only a minor part of the total read consumption.

C. Self-aware access scheme

For write operations, the primary objective is to address the challenges of high write energy and latency, which is the main limit for MRAM as LLC. Numerous prior works have proposed solutions such as read-verify-write (RVW) [8] and self-termination schemes [9]. RVW involves a read operation before writing to avoid write redundancy. However, existing RVW necessitate an additional read cycle before writing, incurring latency overhead unacceptable for latency-critical LLC. Similarly, conventional self-termination typically detect the voltage change on BL upon write completion to disable the write path. A fundamental limitation of this approach is the inherent trade-off between detection margin and the voltage drop across MTJ. Achieving obvious voltage swing on BL

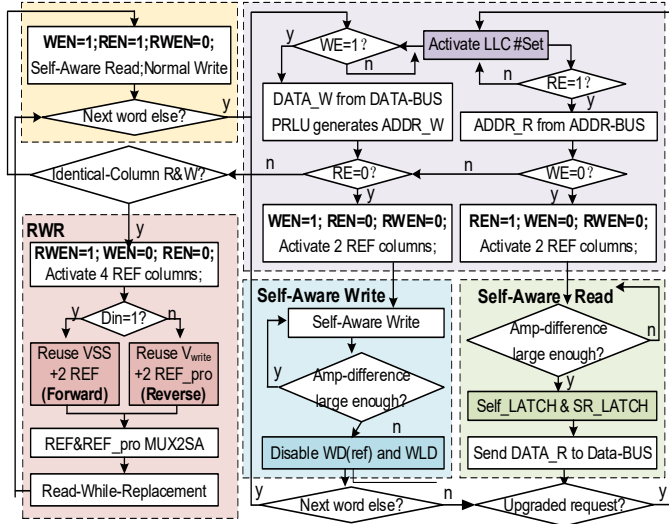


Fig. 6. The workflow of LLC under corresponding working modes

necessitates loss of voltage drop across MTJ, which directly degrades write driver capability. This degradation manifests as increased write latency and higher write power consumption.

To overcome the write power challenges and the limitations of conventional solutions, this work proposes a self-aware write circuit. Leveraging the previously introduced dual reference SA, this scheme concurrently reads the target bit-cell during writing. Specifically, the voltage drop across the HML serves as the read voltage. This read current flows through the MTJ, T_r , R_L and the R_r to VSS. An identical operation is performed concurrently on a dedicated reference column. By sensing the voltages developed across R_L and the R_{Lref} , data can be read out to avoid write redundancy and waste after switching.

Crucially, during the write operation, only the data-corresponding reference of SA is enabled (eg. write “1” and activate “AP” reference column) and its self-latching function is disabled. The enabled amplifier compares the voltage across R_L and the R_{Lref} . If the initial MTJ state matches the write data, the DC voltages on R_L and R_{Lref} become nearly equal once the write voltage stabilizes. Consequently, the amplifier outputs converge, triggering immediate logic-driven termination of write driver for that column via the write enable signal. Conversely, if the initial MTJ state differs from the target data, a significant DC voltage difference arises between R_L and R_{Lref} . This difference is amplified by the differential amplifier, reflected in the output voltage disparity, and the write operation maintains. Upon successful switching, the MTJ state aligns with the reference MTJ, equalizing the dc voltages on the R_L and R_{Lref} , causing the amplifier outputs to converge, again triggering immediate write termination.

This approach confers several distinct advantages over conventional schemes. First, the latency of the verified-read is entirely concealed within writing. Second, by utilizing the write voltage division across the HML to generate the read voltage, a higher write voltage directly results in a proportionally larger voltage difference between R_L and R_{Lref} . The detection margin exhibits a positive correlation with the write voltage. Thus, detection margin does not compromise write driver capability, whereas a stronger write drive inherently enhances detection

TABLE I
PHYSICAL PARAMETERS OF SOT-MTJ

Parameters	Description	SOT-MTJ
T_c	MTJ column Thickness	1nm
CD	MTJ Critical Dimension	80nm
RP	MTJ P-State Resistance	11.0kOhm
TMR	Tunneling Magnetoresistance Ratio	120%
σ	MTJ Resistance variation	7%
T_m	Spin-Hall-metal Thickness	4nm
R_m	Spin-Hall-metal Resistivity	250 $\mu\Omega$ cm
W	Spin Hall Angle	0.3
Hb	Bias magnetic Field	20mT

yield. Furthermore, since the MTJ resistance is significantly larger than the resistance of the HML, the read current constitutes only a negligible fraction of the total write current, resulting in minimal impact on the primary write operation.

Note that performing an identical write operation on the dedicated reference column to generate the reference voltage incurs an additional write power overhead. However, for high-bandwidth parallel write operations characteristic of LLC, sharing 2 reference columns among 32 write bit-cells amortizes this overhead to a mere 1/16 per bit. This modest overhead represents a highly favorable trade-off considering the substantial gains in performance and power efficiency achieved. Consequently, the proposed self-aware write scheme is exceptionally well-suited for LLC applications featuring parallel writes.

IV. EVALUATION

The proposed SOT-MTJ based LLC prototype circuits are simulated using SMIC 40-nm CMOS process, MTJ VerilogA compact model and Hspice platform[21-24]. The parameters of the SOT devices used are shown in Table 1. Among them, the CD of the MTJ cylindrical device is 80nm, the resistance value of the P state is little affected by PVT. However, the resistance value of the AP state has a fluctuation of approximately $\pm 7\%$. The resistance value of the HML is much smaller compared to MTJ. As for the write aspect, the write time shows a negative correlation with the write voltage. Specifically, based on the actual tape-out data, it can achieve 4ns writing at a 770uA write current, whose high-speed is more suitable for LLC.

A. Evaluation of EZLR-LLC

Fig. 7 shows the mixed-signal co-simulation waveforms of the read, write and SWR operations. All waveforms were co-simulated under the condition of different temperature (-40/25/100°C) with specific magnetic-field assistance. In addition, both the read and write operations can be performed at

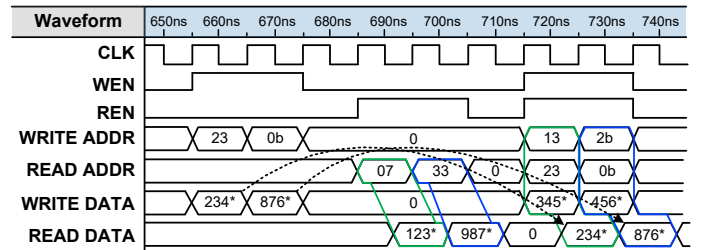


Fig. 7. Digital-analog hybrid waveform diagram: Achieve normal reading and writing functions and verify the realization of simultaneous read-write functions

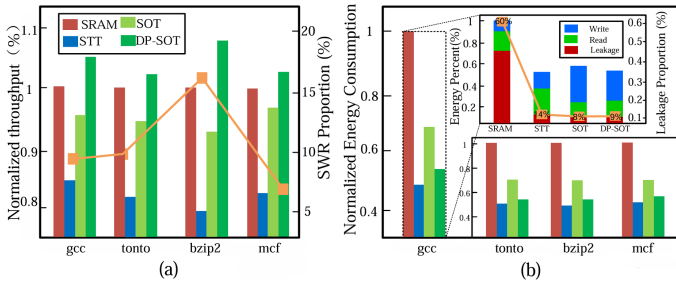


Fig. 8. Four kinds of memory under different instruction sets: (a) throughput (b) power consumption (c) power consumption proportion power supply ($VDD/VDDH=1.2/2.5V$), From the co-simulation waveforms, we have confirmed that the SOT-MRAM can perform 330-MHz 51 μA -read operation, 250-MHz 770 μA -write operation and 250MHz Sense-while-Replace Operation.

Fig.8 shows the throughput and energy consumption under four workloads with NVsim[25] and Gem5 after normalization with SRAM. Due to the long write delay of MRAM, especially STT, its throughput is very low compared to SRAM. However, after the introduction of SWR, the throughput of SOT surpassed SRAM, with a maximum increase of 8%. Meanwhile, the line chart shows the percentage increase in throughput of SOT with and without SWR. Fig.8 (b) shows the power consumption comparison, among which the one of SRAM is significantly higher than those of MRAM. After introducing Self-aware, the access power consumption of SOT is reduced, especially the write power consumption. The upper right part shows the proportion of each part. The leakage of SRAM accounts for the majority, and the actual dynamic power consumption is lower than that of MRAM. The line graph shows the leakage only accounts for a very small part for MRAM and is mainly generated by the peripheral circuits.

Fig. 9 shows a test chip layout of the SOT-MRAM and its cache configuration specifications. This chip is fabricated using a hybrid process technology of 40-nm CMOS and 80nm-CD SOT-MTJ. Fig.9 (a) shows the part of overall chip layout, including EZR-LLC and pins with connection of global IO. Fig.9 (b) shows the layout of the prototype with data-array and peripheral circuits. Fig.9 (c) shows the layout of bit-cell. The memory capacity is 4 kb and the scale of cache-line is 64-bits. The bandwidth of this cache is 32-bits.

Table II summarizes the performance comparison of the proposed SOT-MRAM with other memories presented in recent years. Even though the proposed SOT-MRAM is designed

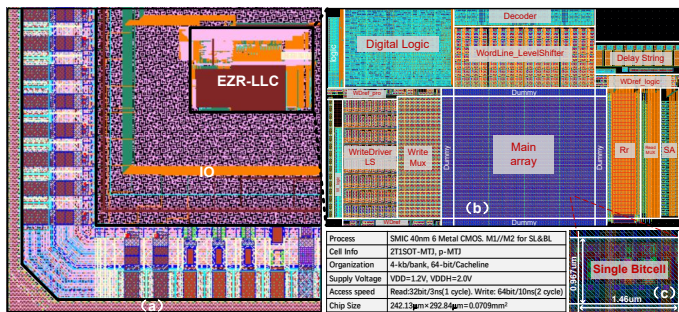


Fig. 9. (a) IO on the main chip and the layout of the LLC (b) The layout composition of each module inside the LLC (c) The layout of a single bitcell (d) The spec of the LLC

TABLE II
PERFORMANCE COMPARISON WITH EXISTING MRAM

	JSSC'23 [13]	ISSCC'16 [8]	JSSC'23 [10]	IEDM'24 [21]	This work
Process	40nm	65nm	55nm	180nm	40nm
Memory device	SRAM	STT-MTJ	SOT-MTJ	SOT-MTJ	SOT-MTJ
Cell Type	8T	2T-2R	2T-1R(+RL)	2T-1R	2T-1R(+RL)
Cell Size(μm^2)	0.9434	0.9975*	5.2185	N/A	1.4118
Memory Capacity	38Kb	4Mb	32Kb	128Kb	4Kb
Write speed	1.6ns	10.82ns	16.7ns	5ns	4ns
@write bit-width	@N/A	@32bit	@8bit	@8bit	@32bit
Read speed	1.6ns	3.3ns	11.1ns	15ns	3ns
@Vread	@0.71-1.1V	@1.25V	@1.2V	@1.4V	@1.2V
MTJ Tape-out support	N/A	Yes	Yes	Yes	Yes
Dual-Port	Yes	No	Yes	No	Yes
Read-while-While	Yes @even in the same row	No	Yes @different row/column	No	Yes @any position

without the use of advanced process nodes, its performance is comparable to that of SRAM which has no nonvolatile memory capability. In addition, through dual-port configuration, this paper realizes the equivalent 0ns replacement of SOT-MRAM in LLC applications, solving the main challenge that hinders MRAM from being used as LLC. Since this paper only targets verifying SOT-MRAM replacing SRAM for equivalent 0ns-replacement LLC, its capacity is smaller compared to other works. However, we have conducted an array-level simulation under scale of 512*512, and the simulation verification for 3ns@1.2V read and 4ns@32bit write has been passed. Note that the MRAM listed in this article all support tape-out, rather than some simulation-level physical models that have not yet been tape-out verified.

For SRAM, adopting dual-port does not significantly enhance its performance, as the replacement delay overhead is not the main factor affecting system performance. Moreover, changing from 6T to 8T would introduce additional area and leakage overhead. Therefore, the dual-port configuration is more suitable for MRAM in LLC applications. As for the cell size, the size is comparable to the SRAM fabricated using the same process rule in [5], which is also under a dual-port configuration.

V. CONCLUSION

This article presents a nonvolatile LLC, using MTJ devices with SOT switching. The SOT-MRAM was fabricated using a 40-nm CMOS process with 80nm-CD SOT-MTJ device. Thanks to the read-disturbance-free characteristic and self-aware access strategy, the 4-kb SOT-MRAM achieves 250-MHz write and 330-MHz read operations with 1.2/2.5V supply voltage under a specific magnetic-field. The SOT-MRAM configured with dual-port, realizes an equivalent-0ns-replacement LLC with SWR strategy applicable to high-speed applications.

We believe that the fabrication of the DP-SOT chip, particularly with innovations that equivalently eliminate write latency and self-aware access, paves the way for scalable, high-performance, and ultra-low-leakage LLC. This achievement is a crucial step toward realizing a new computing paradigm for energy-efficient computing, which is fundamental to the sustainable development of next-generation IoT and AI computing.

REFERENCES

- [1] G. Hu et al., "Spin-transfer torque MRAM with reliable 2 ns writing for last level cache applications," IEDM, San Francisco, CA, USA, pp. 2.6.1-2.6.4, 2019.
- [2] G. K. Chen, P. C. Knag, C. Tokunaga and R. K. Krishnamurthy, "An Eight-Core RISC-V Processor With Compute Near Last Level Cache in Intel 4 CMOS," in *IEEE Journal of Solid-State Circuits*, vol. 58, no. 4, pp. 1117-1128, April 2023.
- [3] E. Cheshmikhani, H. Farbeh and H. Asadi, "ROBIN: Incremental Oblique Interleaved ECC for Reliability Improvement in STT-MRAM Caches," 2019 24th Asia and South Pacific Design Automation Conference (ASP-DAC), Tokyo, Japan, 2019.
- [4] G. Prenat et al., "Ultra-Fast and High-Reliability SOT-MRAM: From Cache Replacement to Normally-Off Computing," in *IEEE Transactions on Multi-Scale Computing Systems*, vol. 2, no. 1, pp. 49-60, 1 Jan.-March 2016.
- [5] K. C. Chun, H. Zhao, J. D. Harms, T. -H. Kim, J. -P. Wang and C. H. Kim, "A Scaling Roadmap and Performance Evaluation of In-Plane and Perpendicular MTJ Based STT-MRAMs for High-Density Cache Memory," in *IEEE Journal of Solid-State Circuits*, vol. 48, no. 2, pp. 598-610, Feb. 2013.
- [6] F. Oboril, R. Bishnoi, M. Ebrahimi and M. B. Tahoori, "Evaluation of Hybrid Memory Technologies Using SOT-MRAM for On-Chip Cache Hierarchy," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 3, pp. 367-380, March 2015.
- [7] J. G. Alzate et al., "2 MB Array-Level Demonstration of STT-MRAM Process and Performance Towards L4 Cache Applications," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019.
- [8] H. Noguchi et al., "7.2 4Mb STT-MRAM-based cache with memory-access-aware power optimization and write-verify-write / read-modify-write scheme," 2016 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2016.
- [9] H. Farkhani, M. Tohidi, A. Peiravi, J. K. Madsen and F. Moradi, "STT-RAM Energy Reduction Using Self-Referenced Differential Write Termination Technique," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 2, pp. 476-487, Feb. 2017, doi: 10.1109/TVLSI.2016.
- [10] M. Natsui et al., "Dual-Port SOT-MRAM Achieving 90-MHz Read and 60-MHz Write Operations Under Field-Assistance-Free Condition," in *IEEE Journal of Solid-State Circuits*, vol. 56, no. 4, pp. 1116-1128, April 2021.
- [11] Y. Seo, K. -W. Kwon, X. Fong and K. Roy, "High Performance and Energy-Efficient On-Chip Cache Using Dual Port (1R/1W) Spin-Orbit Torque MRAM," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 3, pp. 293-304, Sept. 2016.
- [12] J. Keane et al., "17.2 5.6Mb/mm² 1R1W 8T SRAM arrays operating down to 560mV utilizing small-signal sensing with charge-shared bitline and asymmetric sense amplifier in 14nm FinFET CMOS technology," 2016 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2016, pp. 308-309.
- [13] Y. Yokoyama, K. Nii, Y. Ishii, S. Tanaka and K. Kobayashi, "Disturbance Aware Dynamic Power Reduction in Synchronous 2RW Dual-Port 8T SRAM by Self-Adjusting Wordline Pulse Timing," in *IEEE Journal of Solid-State Circuits*, vol. 58, no. 7, pp. 2098-2108, July 2023.
- [14] H. Fujiwara et al., "A 5-nm 254-TOPS/W 221-TOPS/mm² Fully-Digital Computing-in-Memory Macro Supporting Wide-Range Dynamic-Voltage-Frequency Scaling and Simultaneous MAC and Write Operations," 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2022.
- [15] Y. Seo, X. Fong, K. -W. Kwon and K. Roy, "Spin-Hall Magnetic Random-Access Memory With Dual Read/Write Ports for On-Chip Caches," in *IEEE Magnetics Letters*, vol. 6, pp. 1-4, 2015.
- [16] K. Zhang, B. Liu and H. Cai, "Cache-Like Dual-Port SOT-MRAM with Read-while-Write Access Strategy," 2024 IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA), Hangzhou, China, 2024.
- [17] M. Y. Song et al., "High speed (1ns) and low voltage (1.5V) demonstration of 8Kb SOT-MRAM array," 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 2022.
- [18] T. -C. Chang et al., "13.4 A 22nm 1Mb 1024b-Read and Near-Memory-Computing Dual-Mode STT-MRAM Macro with 42.6GB/s Read Bandwidth for Security-Aware Mobile Devices," 2020 IEEE International Solid-State Circuits Conference - (ISSCC), San Francisco, CA, USA, 2020.
- [19] H. Noguchi et al., "7.5 A 3.3ns-access-time 71.2μW/MHz 1Mb embedded STT-MRAM using physically eliminated read-disturb scheme and normally-off memory architecture," 2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers, San Francisco, CA, USA, 2015.
- [20] K. Ali, F. Li, S. Y. H. Lua and C. -H. Heng, "Area Efficient High Through-put Dual Heavy Metal Multi-Level Cell SOT-MRAM," in *IEEE Transactions on Nanotechnology*, vol. 19, pp. 613-619, 2020.
- [21] C. Jiang et al., "Demonstration of 128 Kb SOT-MRAM Chip with 5 ns Write and 15 ns Read Speed, High Endurance Over 1010 and Low ECC-on Bit Error Rate," 2024 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2024.
- [22] E. Liu et al., "A Novel Channel-Less SOT-MRAM with 115% TMR, 2 ns Switching, and High Bit Yield (99.9%)," 2024 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2024.
- [23] W. Yang et al., "Achieving High Yield of Perpendicular SOT-MTJ Manufactured on 300 mm Wafers," in *IEEE Electron Device Letters*, vol. 45, no. 11, pp. 2094-2097, Nov. 2024.
- [24] Q. Zhu and H. Cai, "Surrogate MTJ Model for Early-Stage MRAM Macro Reliability Analysis," 2025 9th IEEE Electron Devices Technology & Manufacturing Conference (EDTM), Hong Kong, Hong Kong, 2025.
- [25] X. Dong, C. Xu, Y. Xie and N. P. Jouppi, "NVSIm: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 7, pp. 994-1007, July 2012.
- [26] J. Su, Q. Zhu, Y. Hou, Q. Shao, B. Liu and H. Cai, "Modeling of Endurance Degradation and Hard Breakdown for MRAM-OTP Demonstration," in *IEEE Electron Device Letters*, vol. 46, no. 8, pp. 1333-1336, Aug. 2025, doi: 10.1109/LED.2025.3581530.