Adaptive MRAM Write and Read with MTJ Variation Monitor

Shaodi Wang, *Student Member, IEEE,* Hochul Lee, *Student Member, IEEE,*
Cecile Grezes, *Member, IEEE,* Pedram Khalili Amiri, *Member, IEEE,* Kang L. Wang, *Fellow, IEEE,*
and Puneet Gupta, *Senior Member, IEEE*

Abstract—Temperature and wafer-level process variations significantly degrade operation efficiency of Spin-transfer torque random access memory (STT-MRAM) and magnetoelectric random access memory (MeRAM), where the write and read reliability issues are exacerbated by the variations. We propose adaptive write and read schemes for highly efficient STT-MRAM and MeRAM programming and sensing that optimally selects write and read pulses to overcome process and temperature variation. With adaptive write, the write latency of STT-MRAM and MeRAM cache are reduced by up to 17% and 59% respectively, and application run time is improved by up to 41%. With adaptive read, the sensing margin is dramatically improved by 1.4X while maintaining read disturbance correctable by error-correcting-code (ECC) correction. To further mitigate read disturbance impact on memory system, additional adaptive read scheme can dynamically lower read voltage according to the proposed monitor result. It can extend memory service time by haft to one year, and reduce read disturbance induced memory failure by 59% to 84%. To better support these schemes, we also propose, design, and evaluate low-cost MTJ-based variation monitor, which precisely senses process and temperature variation. The monitor is over 10X faster, 5X more energy-efficient, and 20X smaller compared with conventional thermal monitors of similar accuracy.

Index Terms—MeRAM, STT-MRAM, adaptive write, adaptive read, thermal monitor, process variation, temperature variation, thermal activation, read disturbance, sensing margin

1 Introduction

Spin-transfer torque magnetoresistive random access memory (STT-MRAM) and magnetoelectric random access memory (MeRAM) are promising non-volatile memory technologies. STT-MRAM is designed with STT magnetic tunnel junctions (STT-MTJ) [1, 2], providing high endurance, fast programming and accessing time, and being identified as a possible replacement of current memory technologies, such as static RAM (SRAM) cache [3, 4] and Dynamic RAM (DRAM) memory [5]. MeRAM designed with voltage-control MTJ (VC-MTJ) [6–9] are switched by voltage-controlled magnetic anisotropy (VCMA) effect, providing more promising programming speed, lower programming energy and higher memory density [10, 11].

However, reliability issues are the main challenges for both STT-MRAM and MeRAM, including write error, read disturbance, and sensing error. In a MRAM write operation, thermal fluctuation can cause a write error. To reduce write error rate (WER) of STT-MRAM, traditionally, write pulse amplitude and duration should be increased, but as a trade-off, write energy increases, memory density decreases due to larger access transistors, and write latency increases. Nevertheless, for MeRAM, there is no previous way to avoid write errors [12]. For read disturbance, the STT-MTJ may falsely switch in a read operation due to the thermal activation, but MeRAM is free from this problem because its read current direction is opposite to write current, which strengthens VC-MTJ's thermal stability. The high-to-low resistance difference in MTJ is quantified by tunnel magnetoresistance (TMR, defined as $(R_H - R_L)/R_L$), and both STT-MRAM and MeRAM have low TMR, leading to a narrow sensing margin and possible read errors.

Process and temperature variation further exacerbates these problems [10, 13–15]. Local variation including etching-induced MTJ diameter and oxide tunnel barrier thickness variation leads to resistance change or MTJ functional failure [16]. Wafer-level variations, including thickness variation of free layer and oxide tunnel barrier layer, affect MTJ performance more severely than local variation [17, 18]. The wafer-level free layer thickness variation can dramatically change energy barrier and thermal stability, especially for out-of-plane MTJs. Temperature variation during operation also affects energy barrier, STT and VCMA effect, and MTJ resistance. Temperature and process variation together can change the energy barrier by 200%, indicating that extreme high write energy is required if STT-MRAM is designed for worst process and temperature corner. Unlike STT-MRAM, MeRAM requires precise voltage amplitude to achieve the least WER, but the voltage varies with energy barrier and hence is sensitive to process and temperature variation. Temperature and process variations also change MRAM's TMR dramatically [19]. For example, TMR drops from ∼205% to ∼140% with temperature rising from 200K to 300K [20]. The change mainly comes from the anti-parallel (AP) resistance change. This indicates that sensing margin in a read operation gets narrow at high temperature, that may result in read errors.

We designed an MTJ-based variation monitor [21] utilizing thermal activation and VCMA effect [21]. The monitor enables in-situ process and temperature variation sensing.

The authors are with the Department of Electrical Engineering, UCLA, Los Angeles, CA, 90095, (e-mails: shaodiwang@g.ucla.edu, chul0524@ucla.edu@ucla.edu, grezes.cecile@gmail.com, pedramk@gmail.com, wang@ee.ucla.edu and puneet@ee.ucla.edu).
Manuscript received August 13, 2017;

The monitor achieves remarkable area, power, and latency improvement compared with conventional on-chip thermal monitors. We proposed an adaptive write scheme which selects optimized write pulse for STT-MRAM and MeRAM to achieve faster write speed based on run-time variation sensing [21]. We also proposed an adaptive read scheme, which smartly selects sensing voltage and sensing resistance to optimize the trade-off between read disturbance rate (RDR) and sensing error rate.

Our contributions are summarized as follows.

- We have designed an MTJ-based variation monitor to sense process and temperature variations. Compared with conventional thermal monitors, the monitor is 10X faster, 5X energy-efficient, and 20X smaller. The monitor directly utilizes MTJs from regular MRAM array without adding fabrication cost overhead.
- We propose an adaptive write scheme that selects write pulse according to ambient process and temperature variation to achieve fast write. We evaluate the proposed method in both circuit-level and system-level. The write latency of MRAM based caches are improved by up to 59%. Applications can be sped up by up to 41%.
- We propose an adaptive read scheme to dynamically select read voltages and reference resistors to maintain read disturbance rate under control while improving sensing margin.

2 Background

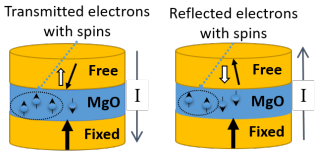

Fig. 1: Spin-transfer torque induced switching.

STT-MTJ and VC-MTJ are resistive memory devices and share a similar device structure, their resistance is determined by the two ferromagnetic layers. One layer has a fixed magnetic direction (referred as reference layer) while the other one has a switchable magnetic direction (referred as free layer). A low ("1") and high ("0") resistance are present when magnetic directions are parallel (P state) or anti-parallel (AP state) respectively. The difference in resistance is quantified by tunnel magnetoresistance (TMR, defined as $(R_H - R_L)/R_L$), where TMR of 180% [22] has been demonstrated in a 8Mb STT-MRAM chip. Based on the magnetization direction, MTJs are classified into in-plane and out-of-plane (perpendicular magnetized) devices. In this paper, we consider out-of-plane MTJs, which have more efficient write, less fabrication challenge, and higher thermal stability (retention time) [23–25].

By contrast, STT-MTJ is switched by bidirectional current, while VC-MTJ is switched by one-directional current pulse. Fig. 1 shows the STT effect. Polarized electrons flowing from the fixed layer to the free layer switch the magnetization of the free layer to P state; when electrons flow in the opposite direction, the reflected electrons from the fixed layer switch the free layer to AP state. Fig. 2 shows the VCMA effect and the fast precessional switching in VC-MTJs. The energy barrier (E_B) separates two stable states of the free layer magnetization (pointing up and down). When a positive voltage is applied across the VC-MTJ, E_B decreases due to VCMA effect, and the thermal activation probability increases. When the voltage reaches V_C (the voltage that fully activates precessional switching), the magnetization spins to the other direction for about 0.5 ns (precessional switching), and the switching can be completed by removing the applied voltage.

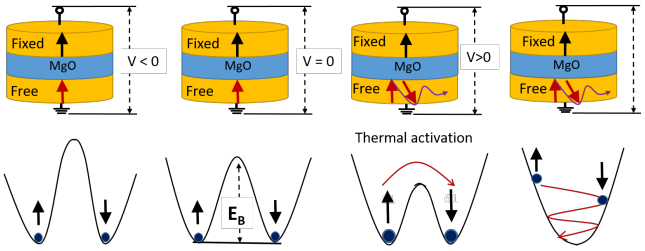

Fig. 2: VCMA-induced precessional switching. A positive (negative) voltage on an MTJ reduces (increases) the energy barrier separating the two magnetization states. A positive voltage over V_C gives rise to a full energy barrier reduction and precessional switching.

3 Related Works

A MTJ-based sensor has been proposed in [26] to sense magnetic field attack to STT-MRAM. However, this monitor used smaller sized MTJs than data MTJs to sense magnetic attack for the reason that small MTJs have low retention time and are switched earlier than bigger MTJs. However, the smaller sized MTJs would have unexpected physic phenomenons from data MTJs, (e.g., single magnetic domain vs multi magnetic domain), and fabrication would be more challenging to print smaller sized monitor MTJs. In [27], an early write termination methodology has been proposed to complete STT-MRAM write upon MTJ switching through sensing voltage change on bit-lines. However, modern STT-MTJs are designed with low resistance leading to little voltage change on bit-lines during MTJ switching. Moreover, the scheme cannot assist MeRAM due to its long sensing latency of over 0.5ns. In [28, 29], a negative differential resistance (NDR)-assisted sensing scheme has been proposed to amplify sensing margin. The NDR's lowest resistance should designed between MTJ's high and low resistance states. However, the MTJ resistance varies with temperature, hence the reliability of the NDR sensing scheme can be improved by the proposed method in Section 7 through designing adaptive NDR resistance. In [30], an adaptive write scheme has been proposed for STT-RAM. Slow switching MTJ columns are marked, and are written with a boosted current. However, temperature variation was not considered. In [31–33], several self-monitored programming schemes have been proposed, where write current is terminated once an MTJ switching is detected. Two main drawbacks of such schemes exist: 1) with a write current through an MTJ, its resistance gets easier to osculate due to the stochastic switching behavior (i.e., fluctuation of magnetization), where a false switching (i.e. resistance

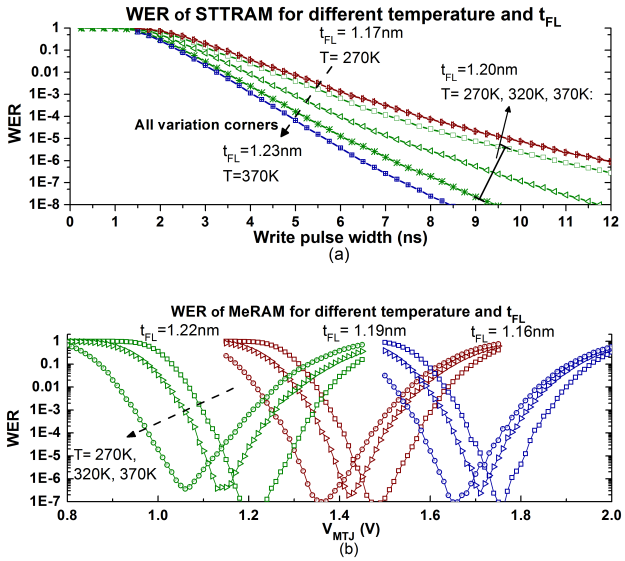

Fig. 3: (a) The STT-MRAM P-to-AP WER as a function of write pulse width under different t_{FL} and temperature corners. In STT-MRAM, P-to-AP switching is more difficult and dominates write latency. (b) The average AP-to-P and P-to-AP WER of MeRAM as a function of write voltage.

changes and recovers back) leads to a false resistance change detection, and then a false write termination and a write error. 2) The monitoring operation is performed every write operation adding to energy overhead. In [34], a current boosting scheme has been proposed. In this scheme, a write current is boosted up if the MTJ state has not toggled after certain write time.

In [35], a variation-tolerant sensing scheme has been proposed to use a same sensing path to sense data MTJ and reference resistor, which eliminated variation impact from CMOS transistors. But large systematic-variation induced read disturbance was not handled, e.g. 100 ^oC temperature change, which can be handled by the proposed monitor in this work. In [36], to avoid read disturbance rate, one more terminal is added to the two-terminal MTJ. during a read operation, the net torque acting on the storage cell always acts in a direction to refresh the data stored in the cell. However, three terminals make it difficult to access the MTJ as well as hurt cell density. Other recent works have approached the read disturbance mitigation from different angles [37–41].

4 Write Error and Read Disturbance Rate Under Variation

The switching behavior of STT-MRAM and MeRAM are affected by temperature and free layer thickness (t_{FL}) [14, 42]. We simulate the switching behaviour of STT-MRAM and MeRAM under different t_{FL} and temperature corners to obtain WER using an LLG-based numerical simulator[1] including temperature dependence, VCMA effect, STT effect, and thermal fluctuation, which has been verified against experimental data in [10]. In the simulations, the t_{FL} variation is assumed to be within 5% across wafer [18],

1. Available at http://nanocad.ee.ucla.edu/Main/DownloadForm

the temperature varies from 270K to 370K, and the local variations including resistance variation are simply treated as random Gaussian variation in the simulations together with variation of access transistors [43] due to line edge roughness, and random doping fluctuation.

The WER of STT-MRAM and MeRAM under different temperature and t_{FL} corners are shown in Fig. 3. According to simulation results, the variation can change WER by over 1,000X. The WER of STT-MRAM is mainly affected by temperature only, while MeRAM is affected by both t_{FL} and temperature. To reduce WER, adaptive write pulses should be chosen according to the temperature and process variation.

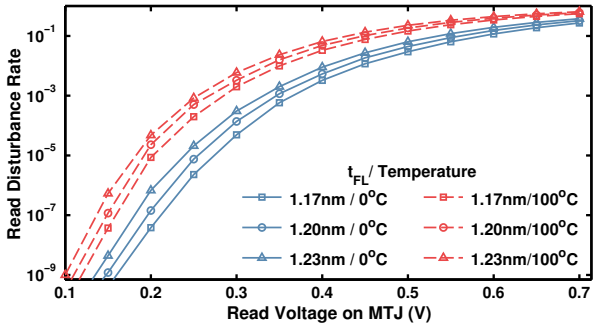

Fig. 4: The STT-MRAM P-to-AP read disturbance rate as a function of voltage drop on P MTJ for a set of temperature and free layer thickness variation corners. The read disturbance rate is obtained from Monte-Carlo simulation with sensing time of 3ns.

The read disturbance rate of STT-MRAM under variation is shown in Fig. 4. In STT-MRAM, P-to-AP is selected as the read current direction due to its high resistance to spin polarized switching and hence results in lower read disturbance rate than AP-to-P switching. As expected, read disturbance increases with read voltage and temperature. Thicker free layer thickness also increases read disturbance because of the reduced perpendicular magnetic anisotropy and hence thermal stability [44]. Fortunately, MeRAM is free from read disturbance because its read current direction is opposite to the direction that can switch the MTJ. Actually, the read current strengthens data retention rather than destroying it. Overall, the variations can shift read disturbance rate by over 10X.

5 MTJ based Variation Monitor

In this section, we propose an MTJ-based variation monitor offering a cheaper solution for in-situ variation monitoring application than exhausting chip testing and expensive conventional thermal monitors. The monitor senses combined temperature and wafer-level t_{FL} variation.

5.1 Sensing Principle

Monitoring variation through directly WER measurement is expensive, which requires large number of writes and reads. The proposed monitor utilizes thermal activation and VCMA effect to indirectly monitor variation by sensing the thermal activation rate in MTJs under different stress voltage and current.

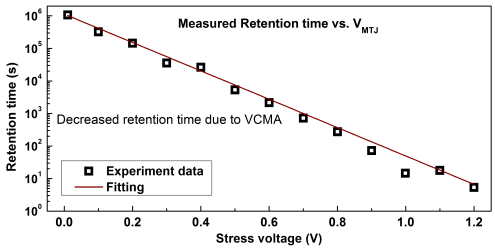

Fig. 5: The experimentally measured retention time as a function of stress voltage on MTJs.

$$t_{R,STT} = \exp\left(\Delta\left(1 - I_{MTJ}/I_C(\Delta)\right)\right)$$
$$t_{R,VC} = \exp\left(\Delta\left(1 - V_{MTJ}/V_C(\Delta)\right)\right) \quad (1)$$

As described by (1) [45, 46], the retention time (i.e., the mean of switching time under non-write state) of STT-MTJ ($t_{R,STT}$) and VC-MTJ ($t_{R,VC}$) exponentially depends on thermal stability (Δ, proportional to energy barrier), critical current of STT-MTJs ($I_C(\Delta)$), and critical voltage of VC-MTJs ($V_C(\Delta)$). The current and voltage across STT-MTJ and VC-MTJ respectively can shorten retention time. The write pulse width and voltage that create instantaneous switching (<10ns) for STT-MRAM and MeRAM depend on $I_C(\Delta)$ and $V_C(\Delta)$, which also depend on Δ. This indicates that knowing the $t_{R,STT}$ and $t_{R,VC}$ changes due to temperature and process variation can predict the MRAM write behavior change.

5.2 Circuit Implementation and Simulation

Retention time of MTJs is too long to be measured directly. Fortunately, we observe that, as illustrated by the Equation (1), applying current/voltage on MTJs reduces retention time exponentially. This observation is demonstrated in experiment measurement, where retention time decreases exponentially with increasing stress voltage due to VCMA effect in Fig. 5. Inspired by this observation, we introduce a stress operation in the proposed variation monitor. We apply low voltage or current across MTJs to reduce retention time and hence to increase thermal activation rate, and we call them stress voltage or stress current for simplicity.

$$P_{SW,STT} = 1 - \exp\left(-t_S/t_{R,STT}\right)$$
$$P_{SW,VC} = 1 - 1/2 * \exp\left(-t_S/t_{R,VC}\right) \quad (2)$$

When the retention time reduces to sub-μs, the MTJ switching rate (P_{SW}) due to thermal activation during stress time (t_S in tens of ns) can be measured as explained in Eqn. (2). Then P_{SW} (correlated to $t_{R,STT}$ and $t_{R,VC}$) inherently reflects the ambient variation.

We use an example in Fig. 7 to simply illustrate the proposed sensing principle. The top MTJ is assumed to have retention time of 10 years, while the bottom one suffers from variation and has retention time of only 10 hours at normal conditions. To sense the variation difference, we apply the same stress voltage across the two MTJs. As stated in this section, their retention time are exponentially reduced. They reduced to 100 μs and 10 ns respectively. During the voltage stressing time of 20 ns, the bottom MTJ is more possible to be thermally activated, while the top MTJ state most likely remains unchanged. Therefore, thermal activation switching rate can be obtained by performing more such tests on single MTJ or an MTJ array. We choose to do tests simultaneously on an array to speed up sensing operation. We set a threshold for the thermal activation rate. If the switching rate reaches selected threshold after a stress operation, the stress level is output to reflect ambient variation. Otherwise, the monitor continues to try a higher stress level of voltage/current.

The monitor design is shown in Fig. 6. To minimize fluctuation of sensing results caused by MTJ stochastic switching, a number of MTJs are sensed simultaneously, and hence the individual stochastic switching fluctuation is averaged out. In a stress operation, all MTJs in the monitor are in AP state initially. The write control circuit applies a stress current (for STT-MRAM) or voltage (for MeRAM) simultaneously on all MTJs in the monitor array for 20ns. The stress current (for 256-MTJ bit-line) ranges from 2.5mA to 10mA, which is precisely controlled by the effective width of transistors in the stress current selection array, where the stress current variation is close to 0 due to the large transistor width guaranteeing monitor accuracy. The stress voltage on VC-MTJs is adjusted by dividing voltage on bit-lines and resistors (vary from 200Ω to 700Ω). The stress voltage variation is also close to 0 because the equivalent parallel resistance of all VC-MTJs on a bit-line averages out individual MTJ resistance variation.

After a stress operation, the read control circuit selects each MTJ one by one and reads its state. In the read, the bit-line (BL) and reference bit-line (BL_ref) are pre-charged and pulled down by the read MTJ and reference resistor separately. The difference between $Vsense$ and $Vref$ creates an output to S Latch, and a switched MTJ raises S's output from 1 to 0, then the XOR of S Latch and D Latch (output is constantly 1) creates a rise edge, which is counted by Counter2. At last a switched MTJ is reset by a write pulse for future stress operations.

We simulate the monitor design using a 65nm technology node commercial cell library. Please note that the simulation results in this section are compared with other works [47–50] designed in 65nm. In the following sections, designs are simulated with advanced 32nm technology. The stress pulses are shown in Fig. 8 (a). Stress current has $< 0.3\%$ and $< 4.7\%$ variation due to temperature (27^oC to 100^oC) and oxide thickness variation (9% resistance change) respectively, while stress voltage has $< 1\%$ and $< 2\%$ variation accordingly. In addition, switched MTJs (e.g., 30%) during stress time can cause up to 10% and 2% stress current and voltage change respectively. The low variation demonstrates the proposed monitor accuracy.

Fig. 8 (b) shows the simulated waveforms of read, counting, and reset operations. The first and third reads are performed on switched MTJs, where write pulses follow reads to reset MTJs, and the counter increases because of the detected MTJ switching. The second read is on a non-switched MTJ, and hence no action is taken after the read. If the counted number reaches the selected threshold (e.g., 64 out of 256 MTJs), it sends out a completion signal and outputs the current stress level, which presents the ambient variation level. If the selected threshold is not reached after reading all MTJs, the counter is reset, and a higher stress level is selected in the next variation sensing cycle.

We simulate the circuitry to obtain the switching rate

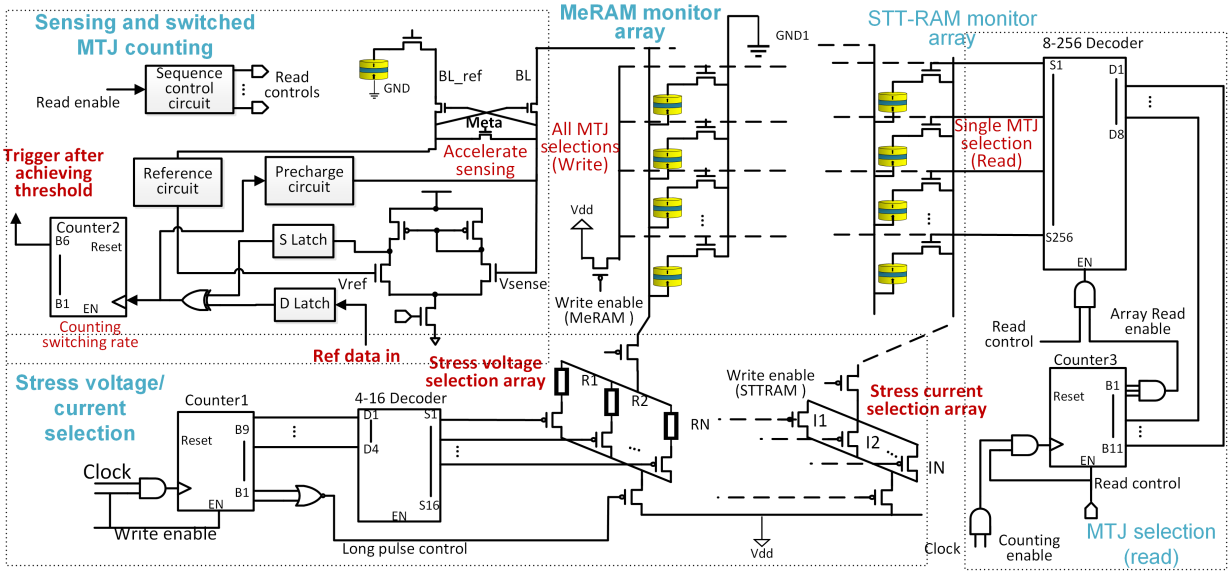

Fig. 6: The schematic of STT-MRAM and MeRAM based variation monitor. Variation monitoring operations: 1) apply stress voltage/current on MRAM array controlled by stress voltage/current selection circuitry; 2) select every MTJ (controlled by MTJ selection circuit) one by one to read and count MTJ switching rate (controlled by sensing and switched MTJ counting circuit).

and standard deviation (σ) of a 256-MTJ variation monitor with different stress levels and variation corners as shown in Fig. 9. If we select a switching rate threshold to any value between 10% to 30%, the voltage levels to reach the threshold for different variation levels (10^oC temperature difference between two consequent curves) can be well differentiated, e.g., the dotted curves show the standard deviation (accuracy of the monitor) is much smaller than curve gaps, and the variation levels can be determined. Hence, temperature variation of 100^oC can be distinguished with ten stress levels, achieving the accuracy of 10^oC.

Previously, we show that the proposed monitor can sense thermal stability by appropriately selecting stress voltage levels. With sensed thermal stability, it can assist to optimize MRAM variability and reliability. However, the stress voltage level selection is not straightforward. We use one example application to show how these levels are selected. In this example, the monitor can warn retention hazard when MTJ's retention error rate reaches a threshold E_r. Though E_r is too low to be easily sensed, we are able to find a stress voltage V_s such that stressing such MTJ for 20ns can increase the switching rate to 20%. When stress time and stressed switching rate threshold are given, V_s is only determined by E_r. The mapping of V_s and E_r can be extracted from chip test. Therefore, in this application, the proposed monitor reaching switching rate threshold

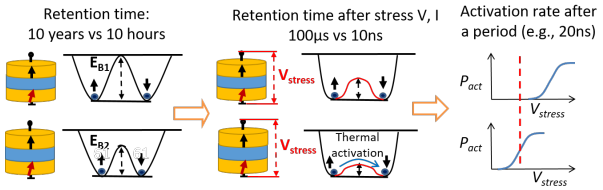

Fig. 7: The simplified illustration of the proposed sensing principle. Two MTJs with variations have different activation rate after voltage stressing.

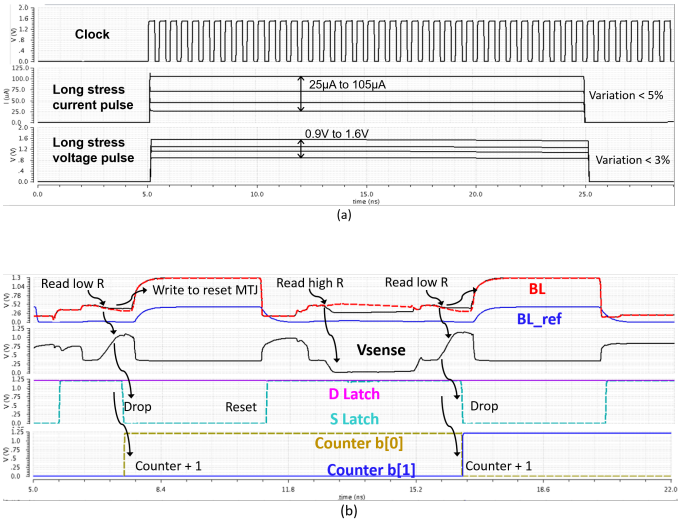

Fig. 8: (a) Different stress current/voltage in the proposed monitor. (b) Simulated waveforms of read, reset and counting operations.

with stress voltage V_s indicates MRAM arrays have average retention error rate over E_r. Multiple stress voltage levels may be introduced for other applications like the adaptive write in Section 6.

Table 1 shows the comparison between the proposed variation monitor and conventional thermal monitors. In conventional monitors, long latency and high energy are consumed by analog-to-digital blocks and sensing bipolar transistors. The proposed monitor is less accurate but faster with lower energy/sample and smaller area. Larger sensing array can improve the accuracy by reducing the standard deviation (σ) (Fig. 9) allowing for using finer granularity of stress levels at the expense of sensing energy and latency. In addition, the granularity of stress current/voltage is also

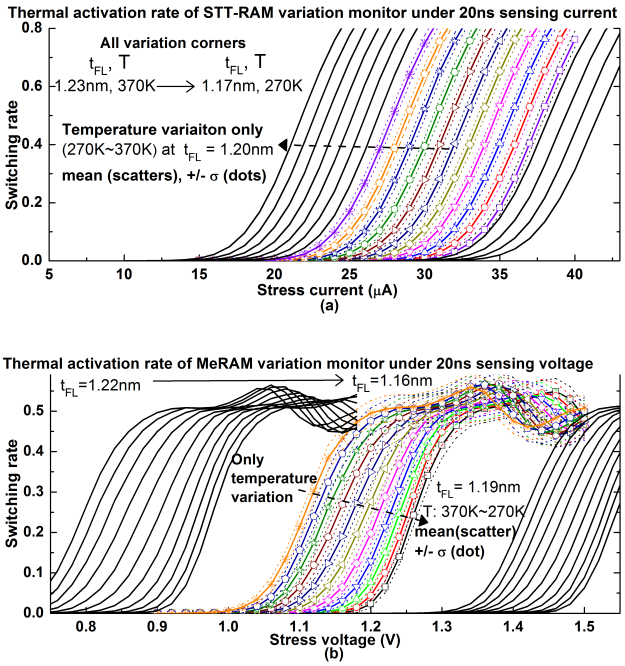

Fig. 9: Switching rate of (a) STT-MTJ- and (b) VC-MTJ-based variation monitor under different stress current and voltage respectively. The color lines are switching rate for only temperature variation (10^oC interval). The dot lines outline standard deviations (σ) of thermal activation rate (σ is caused by process variation and random thermal activation).

TABLE 1: Comparison between conventional thermal monitors and the proposed variation monitor. The proposed monitor uses 256 MTJs and 10 stress levels

Monitor	Latency	Accuracy	Energy	Area
S1 [47]	0.1ms	9^oC	$0.015\mu J$	$0.01mm^2$
S2 [48]	0.2ms	3^oC	$0.24\mu J$	$0.04mm^2$
S3 [49]	1ms	2^oC	$0.49\mu J$	$0.01mm^2$
S4 [50]	100ms	0.1^oC	$13.8\mu J$	$0.04mm^2$
this(STT)	1-10μs	10^oC	0.12-1.2nJ	$0.0005mm^2$
this(Me)	1-10μs	10^oC	0.27-2.7nJ	$0.0005mm^2$

constrained by process variation of CMOS circuit. Fortunately, the achieved accuracy is enough for selecting optimal write pulse and reliable read voltage for STT-MRAM and MeRAM (i.e., Sections 6.1 and 7 show that three stress levels are enough) indicating that the proposed monitor supports the proposed adaptive write and read schemes with less overhead. The area of the monitor is dominated by the 8-256 decoder (97.1% of total transistors). The area of 8-256 decoder was estimated through synthesize, place and route using commercial 65nm library.

6 Adaptive Write

6.1 Adaptive Write Scheme

The adaptive write scheme is to dynamically select an optimized pulse width (voltage) for STT-MRAM (MeRAM) to minimize write latency according to ambient variation. Multiple pulse widths are simply implemented by delay circuits shared by multiple bit-lines, and hence introducing negligible overhead. Generating multiple write pulse voltages requires voltage regulators, which are also shared by

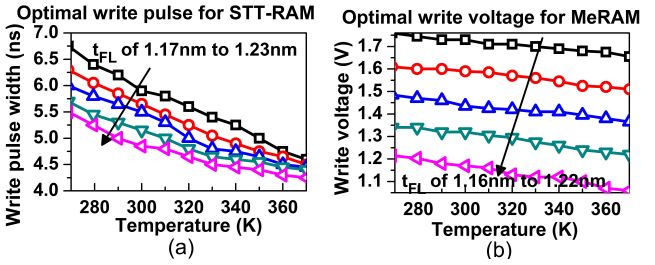

Fig. 10: Optimal write pulses for (a) STT-MRAM and (b) MeRAM under different t_{FL} and temperature corners.

the entire MRAM array. Local variations like temperature variation over MRAM array [15] can be captured by placing multiple proposed monitors. One such monitor only uses one bit-line with an area overhead of <0.005% (i.e., adding monitor circuits in MRAM boundary does not affect MRAM fabrication regularity). The monitor also consumes negligible power (i.e., 2.7nW for one variation sample per second) compared with power of MRAM array (>10 mW).

6.2 Adaptive Write using Variation Monitor

In this section, we evaluate the write scheme with the proposed variation monitor. The write circuit for MRAM is designed to enable program-and-verify [51] which performs a read check following a write (the writing data is pre-stored in D Latch in Fig. 6), and if a write error is detected, additional writes are performed to correct the error. With this, 0 WER is guaranteed for MeRAM and STT-MRAM irrespective of the single write pulse voltage/width. For STT-MRAM, shortening single write pulse leads to reduction in both latency and energy for a single write trial, As a trade-off, WER increases as well as the chance of additional writes, possibly adding overall latency and energy. There is an optimal single write pulse achieving minimum expected write latency. Such optimal pulse can reduce STT-MRAM's expected latency and energy by over 60% compared with conventional write design [10]. The optimal pulse widths (voltages) for minimum expected latency (including initial write, read checks, and additional writes) of STT-MRAM (MeRAM) are shown in Fig. 10. The pulse width for STT-MRAM spans from 4.25ns to 6.75ns mainly affected by temperature. The voltage range for MeRAM is from 1.05V to 1.75V affected by both temperature and t_{FL}.

In the following evaluation, the combined temperature and t_{FL} corners are divided into groups based on the variation monitor's outputs (stress levels reaching P_{SW} threshold). Each group has an optimized write pulse minimizing the maximum write latency in the group. More write pulse choices (equal to stress levels) result in shorter programming latency.

Our evaluation flow is illustrated in Fig. 11 (a). We simulate the peripheral circuit (see Fig. 6) with a bit-line size of 256 MTJs using 32nm commercial library and simulate the WER of MTJs with LLG-based numerical model. In the bit-line programming simulations, 10 temperature variation corners from 270K to 370K and three wafer-level free layer thickness variation corners of 0.06nm are enumerated. The 30 temperature-process variation-corner combinations are classified into groups according to the output levels from the proposed variation monitor. For each group of variation

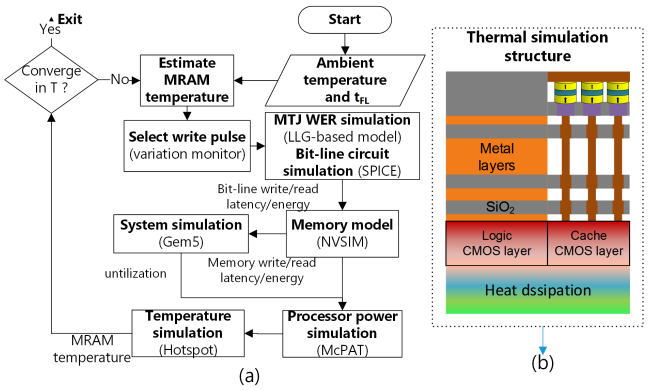

Fig. 11: (a) Evaluation flow of adaptive write in MRAM based system. (b) The cross-section structure for thermal simulations.

corners, the maximum write latency is minimized by selecting one optimal write pulse (pulse width for STT-MRAM and pulse voltage of MeRAM).

Bit-line-level results show that STT-MRAM has write latency variation from 5.5 ns to 7.5 ns and MeRAM has that from 4 ns to 10.1 ns. With the inputs of bit-line results, we use NVSIM [52] to obtain latency and energy of MRAM array (cache). In Fig. 13, the write latency of STT-MRAM L2 Cache with different t_{FL} corners is shown to decrease with increased number of pulse choices, and each point is the maximum or average latency over 10 temperature corners from 270K to 370K. The maximum write latency of STT-MRAM is improved by up to 17%. The maximum latency for t_{FL} corner of 1.17nm does not see improvement because the corner with 1.17nm t_F and 270K is always the worst one to be optimized no matter how many groups (pulse width choices) are used. MeRAM's write latency reduction is up to 59%, but there is a latency jump for t_{FL} of 1.19nm from one to two voltage choices. This is because when only one group (single write voltage) is used, the optimal voltage of 1.19nm t_{FL} corner is close to the optimal voltage for all corners (i.e., the voltage to minimize WER for all corners in Fig. 3b), but when two groups are used, the optimal voltage for 1.19nm corner gets farther from those for both groups. As seen, three choices are efficient enough for write latency improvement.

We modified gem5 [53] (i.e., original Gem5 only has fixed cache write time, we have added the support for varying cache write time, which is necessary for MRAM evaluation) to simulate two cases: 1) an x86 processor with one core and one single-level 8-MB MRAM data cache; 2) an x86 processor with two cores, two 1-Mb MRAM L2, and one 16-MB MRAM L3 caches (L1 uses default SRAM). We modified McPAT [54] to simulate processor power and used Hotspot [55] to simulate MRAM temperature with the structure shown in Fig. 11b.

We simulated one billion instructions of SPEC benchmarks using our evaluation flow. The application run time reduction with adaptive write is shown in Fig. 12. The processors with single-level MRAM see noticeable application speedup after using adaptive write, where up to 41% and 9% run time reduction are shown for MeRAM and STT-MRAM respectively. However, the improvement is much less for processors with MRAM L2 and L3 (up to 10% and 2% for MeRAM and STT-MRAM respectively), because cache write latency improvement is hidden by SRAM L1. This indicates that the adaptive write scheme may be more efficient for embedded applications with single-level MRAM cache. Compared with MeRAM, STT-MRAM write latency improvement is not significant.

6.3 Cache Power Saving

In the adaptive write proposed in this paper, we aim to improve write latency for both STT-MRAM and MeRAM regardless of the power. Fortunately, the cache power is also reduced with increased number of write pulse choices as an additional benefit of the adaptive write. For STT-MRAM, more pulse choices lead to shorter overall programming time and possibly less MTJ switching time indicating energy reduction. This is because driving current to switch MTJ dominates power consumption, and less programming time usually leads to less energy. For MeRAM, the adaptive write chooses appropriate write voltage to reduce WER, indicating less additional program-and-verify cycles. The energy of MeRAM is dominated by repeated bit-line charging and discharging and hence less cycles give rise to energy reduction. Fig. 14 shows that the maximum and average power of L3 Cache over different variation corners decrease with increased pulse choices. Again, the adaptive write in this paper is designed for latency reduction, but it can also be designed for power reduction alternatively, which will achieve even more energy reduction than Fig. 14.

7 Adaptive Read

To improve the STT-MRAM read reliability and efficiency, MTJ sensing margin should be maximized while maintaining a read disturbance rate below the error-correcting-code's (ECC) tolerable rate [56]. This is non-trivial because of the tradeoff between sensing margin and read disturbance. To improve sensing margin, a large sensing current is required to create more voltage difference, which however increases read disturbance rate. Moreover, the sensing margin and read disturbance rate are also severely affected by process and temperature variation. Simply designing for the worst variation corner will lead to insufficient reliability margin. To resolve this issue, we propose an adaptive read scheme which dynamicaly control sensing circuits according to process and temperature variations. This scheme can improve sensing margin without sacrificing read disturbance.

Read disturbance rate depends on STT-MTJ thermal stability, which varies with sensing current amplitude, free layer thickness and temperature. On the other hand, sensing margin also depends on sensing current amplitude and temperature. This is because STT-MTJ resistance, which strongly affects sensing margin, has strong dependence on temperature, especially for the AP resistance as illustrated in Fig. 15. Therefore temperature variation is important to both read disturbance and sensing margin, and fortunately it can be monitored. Together with temperature, wafer-level free layer thickness variation, which affects read disturbance, can be monitored by the proposed variation monitor. Therefore, according to outputs from a conventional temperature monitor and the proposed variation monitor, the proposed adaptive read is able to select between two

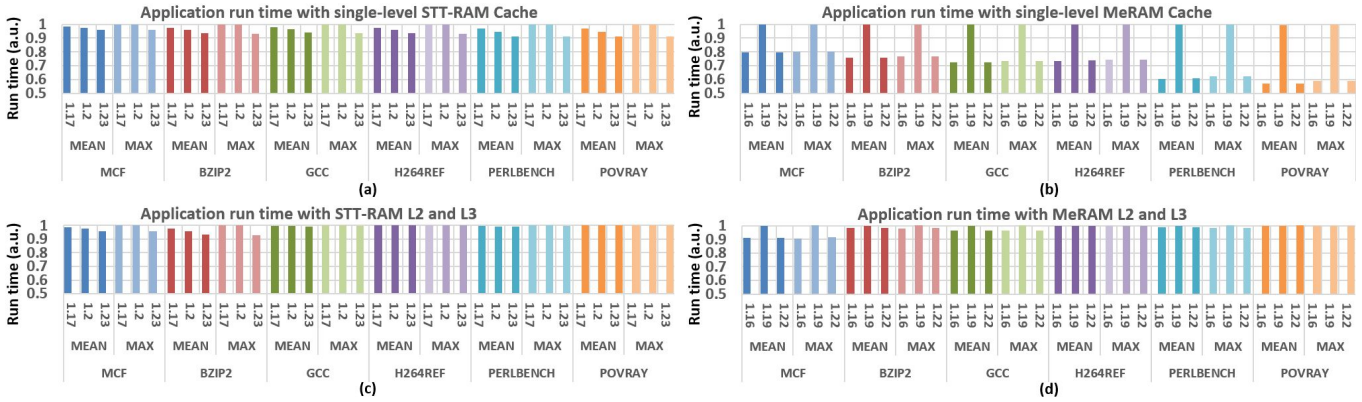

Fig. 12: The average/maximum run time of SPEC benchmarks using adaptive write (with three write pulse choices) for (a) one-core processor with single-level 8-MB STT-MRAM cache and (b) single-level 8-MB MeRAM MeRAM cache, a dual-core processor with (c) 1-MB STT-MRAM L2 and 16-MB STTRAM L3, and (d) 1-MB MeRAM L2 and 16-MB MeRAM L3 over temperature corners (270K to 370K). Run time is normalized to the maximum run time for processors without adaptive write (one write pulse choice) for each benchmark.

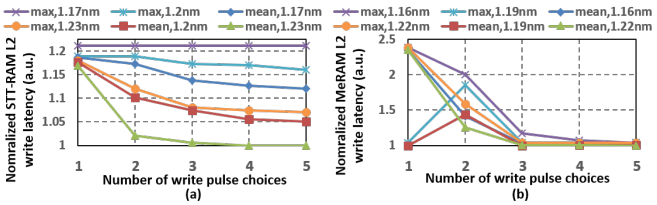

Fig. 13: The maximum and average write latency in (a) 1MB STT-MRAM L2 and (b) MeRAM L2 from 270K to 370K under different t_{FL} corners with different number of write pulse choices.

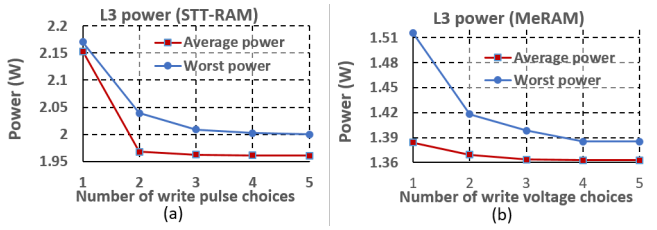

Fig. 14: The maximum and average write power for (a) 16MB STT-MRAM L3 and (b) MeRAM L3 over temperature (from 270K to 370K) and t_{FL} (0.06nm change) corners with different number of write pulse choices.

reference resistors and two read voltages to improve MRAM read reliability. Read voltage selection is to maintain read disturbance rate within ECC's capability [56], where the selection is based on MRAM thermal stability monitored by the proposed monitor. Reference resistor selection is to improve sensing margin according to MTJ temperature-related resistance change assisted by a conventional temperature monitor.

7.1 Adaptive Sensing Circuit using Multiple Reference Resistance

As stated in Section 4, STT-MTJs with low thermal stability are susceptible to read disturbance. where high read voltage, high temperature and low free layer thickness usually result

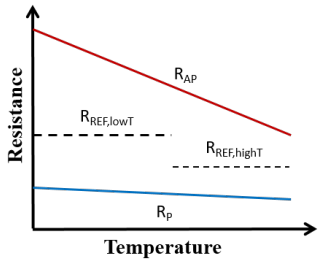

Fig. 15: An illustration for using two reference resistors for STT-MTJ state sensing at low and high temperature.

in low thermal stability. The level of temperature and free layer variation are obtained using the proposed monitor (Section 5). To select a read voltage in the adaptive read, one threshold stress current level is set in the variation monitor: when the monitor's output (variation-induced thermal stability change) is below the threshold, A high read voltage is selected, and vice versa. For temperature dependence of STT-MTJ resistance, the AP resistance changes dramatically with temperature [20], and the change is approximately linear, while the P resistance is more stable. Therefore low TMR and low sensing margin presents at high temperature. Experiment data [57] shows that TMR drops from 192% at 4.2K to 90% at room temperature, and the TMR will further drop at higher temperature like chip operating temperature (e.g., over 80°C). To improve the sensing margin, a low and a high reference resistors are selected at high and low temperature respectively as illustrated in Fig. 15. The resistor selection is controlled by an on-chip temperature monitor. Again, one threshold temperature is needed for the selection, which can be obtained from experimental data or empirical models like in [20].

The proposed sensing circuit is shown in Fig. 16. High and low read voltages are selected by the signal "Low Δ", which is an output from the proposed variation monitor. Reference resistors are selected by the signal "Low temperature", which is an output from an on-chip temperature monitor.

We conduct an example evaluation. First we fitted an

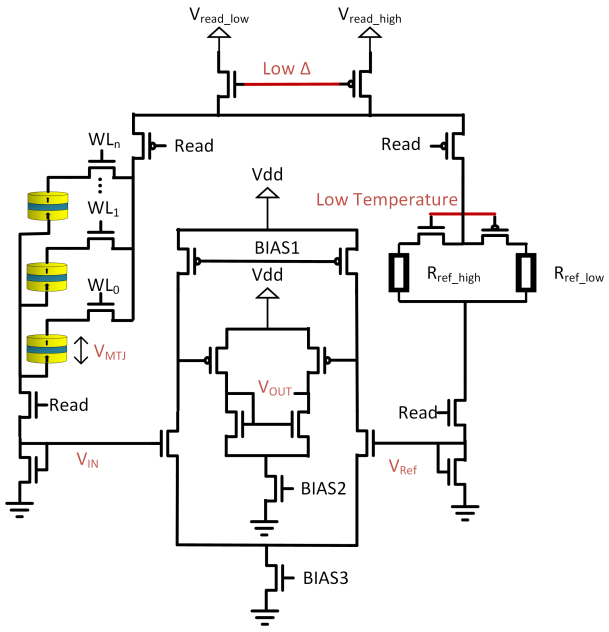

Fig. 16: Sensing circuit used in the adaptive read. The switch of two reference resistors are controlled by temperature monitor. The switch of read voltage is controlled by the proposed variation monitor.

MTJ resistance model from [20]. Then we simulated and fitted a read disturbance model based on an MTJ switching model from [10]. In the models, MTJ AP resistance drops from 5k Ω (15 ^oC) to 3.33k Ω (120 ^oC), and P resistance drops from 2k Ω to 1.8k Ω. In our sensing circuit, the current direction was chosen to the direction of P-to-AP switching current for the reason that P-to-AP switching is more resistant than AP-to-P switching, hence our selection gives lower read disturbance rate. Therefore, only P MTJs are possibly disturbed in a read operation. In the evaluation, the tolerable read disturbance rate by ECC is 10^{-9} [56]. According to Fig. 4, we used 0.66V and 0.78V as low and high read voltages, giving rise to 100 mV and 150 mV voltage drops across P MTJ respectively. In addition to temperature variation, we also considered another 10% resistance variation due to process variation which is not monitored. Hence the reference resistance should be designed to sense 90% $R_{AP}(T)$ and 110% $R_P(T)$, where T is temperature. To maximize sensing margin, 3.25k Ω and 2.85k Ω were chosen as the high and low reference resistances. CMOS sensing circuit were simulated using SPICE with a Verilog-A MTJ model [10] and 32nm commercial library (temperature models included).

The STT-MRAM will normally work at room temperature with low process variation, where the adaptive read scheme selected the high reference resistor and the high read voltage. As a comparison, the conventional non-adaptive read design has to be designed for the worst variation corner (high temperature and strong process variation), which uses a low read voltage and a low reference resistor. We performed circuit simulations of the proposed adaptive read design and the conventional non-adaptive read design. The sensing waveforms are shown in Fig. 17. The sensing margin (V_{in} - V_{ref}) was improved from 26.8 mV of non-adaptive read to 37.8 mV of adaptive read. In the meantime, the read disturbance rates for both designs are controlled below 10^{-9} all the time.

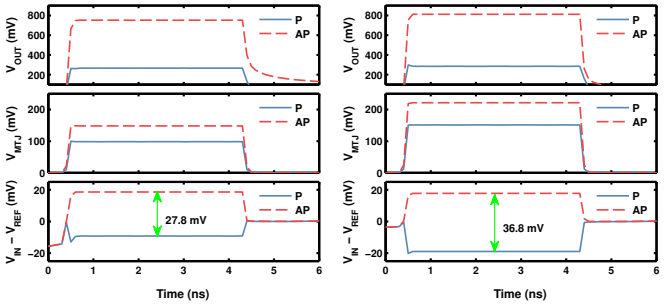

Fig. 17: (a) Simulated sensing waveforms for a conventional non-adaptive read design with single read voltage and single reference resistor. (b) Simulated sensing waveforms for the adaptive STT-MRAM read scheme with two read voltage and two reference resistors, where the high read voltage and high reference resistor are selected at normal condition (room temperature and MTJs with high thermal stability).

7.2 Adaptive read for lower disturbance rate

At some temperature and process variation corners, an MTJ is easily disturbed by high read current, creating high read disturbance rate. With the proposed monitor, an adaptive read scheme is proposed to dynamically lower read voltage at such corners to reduce read disturbance rate, leading to long service time before a failure. To evaluate its benefits on system reliability, we simulate the failure rates of MRAM systems with or without adaptive read using an memory reliability simulator MEMRES [56]. The simulator enables fast memory reliability simulation with system-level reliability management including ECC, page retirement, memory mirroring, memory scrubbing and rank sparing.

TABLE 2: Architecture of a 8-GB DRAM DIMM.

Ranks	Chips	banks	Mats	Rows	Columns	Access-Rate
1	16+2	8	128	512	8192	1e12/hour

TABLE 3: Fault FIT rates for STT-MRAM. The read disturbance error rates are for STT-MRAM ($t_{FL} = 1.2nm$) under 320K and 370K using adaptive read and non-adaptive read.

Fault types	Transient FIT	Permanent FIT	Cover-Rate
Single-word	1.4	0.3	1
Single-column	1.4	5.6	0.02
Single-row	0.2	8.2	0.002
Single-bank	0.8	10	0.002
multi-banks	0.3	1.4	0.002
single-lane	0.9	2.8	0.002
Read disturbance error rate	non-adaptive read	5.37e-7(370K), 3.43e-8(320K)	
	adaptive read	1e-9(370K), 1.8e-10(320K)	

The tested 8-GB STT-MRAM configuration is shown in Table 2. In the simulation, a single-error correction-and-double-error detection (SECDED) [58] is enabled to correct any single-bit error in a 72-bit word (64 data bits and 8 parity bits). SECDED is very efficient to correct read disturbance error for that MRAM read disturbance causes one bit flip in a word. The MRAM also enables scrubbing function, which periodically scans entire memory and fixes

all detected soft errors, e.g., MRAM retention error, read disturbance error. The two methods are most cost-effective for read disturbance error. Table 3 shows injected fault FITs in the simulation. All fault types and FIT are obtained from DRAM field studies [59, 60] except the read disturbance rate, because MRAM and DRAM share similar peripheral circuits, and those faults are mostly caused by peripheral circuit failures.

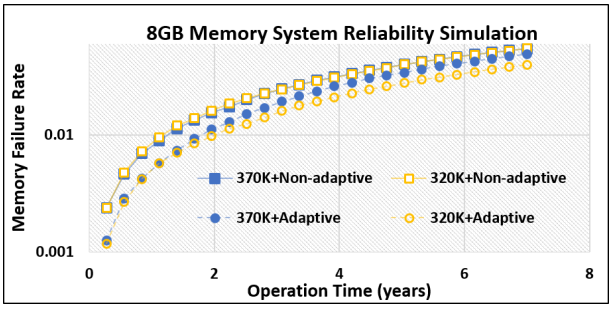

Fig. 18: Failure rates (accumulated failure probability) of STT-MRAM in a 7-year operation. Adaptive read can obviously reduce memory failure rate and lower service cost. The example variation corners are $T_{FL} = 1.20nm$ with temperatures of 320K and 370K.

We performed 50,000 simulations of 7-year long STT-MRAM operation. The MRAM system fails only when the ECC cannot correct faults. Though read disturbance error itself can be corrected by SECDED, the coincidence of read disturbance error and other faults, e.g., column fault, in a single word can result in an ECC failure. The failure rates (accumulated failure probability) are plotted in Fig. 18. As seen from the results, adaptive read can relative reduce system failure rates and extend memory service time by about half to one year. Among all failures, read disturbance only accounts for about 22% to 24% failures for non-adaptive read cases, and 5% to 10% for adaptive read cases. If we focus on the read disturbance related failures, adaptive read reduces them by 84% and 59% respectively for 320K and 370K cases. This demonstrates the effectiveness of adaptive read.

8 Conclusion

We design an MTJ-based variation monitor to sense process and temperature variation. At the same accuracy, the variation monitor achieves 20X smaller area, 10X faster speed, and 5X less energy. We propose an adaptive write scheme to minimize write latency of STT-MRAM and MeRAM according to ambient process and temperature variation. The write latency of STT-MRAM and MeRAM cache is reduced by up to 17% and 59% respectively, while simulated application run time shows up to 1.7X improvement. We also propose an adaptive read design to improve sensing margin while maintaining read disturbance rate with ECC's capability. It dynamically selects between two read voltages and two reference resistors according to chip temperature and process variations. This scheme can improve the sensing margin by 1.4X against non-adaptive read. To further mitigate read disturbance impact on memory system, adaptive read can dynamically lower read voltage according to the proposed monitor result. It can extend memory service time by haft to one year, and reduce read disturbance induced memory failure by 59% to 84%.

References

[1] C Heide. "Spin currents in magnetic films". *Phys. Rev. Lett.* 87.19 (2001), p. 197201.

[2] DC Worledge, G Hu, David W Abraham, JZ Sun, PL Trouilloud, J Nowak, S Brown, MC Gaidis, EJ OSullivan, and RP Robertazzi. "Spin torque switching of perpendicular Ta— CoFeB— MgO-based magnetic tunnel junctions". *Appl. Phys. Lett.* 98.2 (2011), pp. 022501–022501.

[3] Clinton W Smullen, Vidyabhushan Mohan, Anurag Nigam, Sudhanva Gurumurthi, and Mircea R Stan. "Relaxing non-volatility for fast and energy-efficient STT-RAM caches". *HPCA*. IEEE. 2011, pp. 50–61.

[4] Adwait Jog, Asit K Mishra, Cong Xu, Yuan Xie, Vijaykrishnan Narayanan, Ravishankar Iyer, and Chita R Das. "Cache revive: architecting volatile STT-RAM caches for enhanced performance in CMPs". *Proc. DAC*. ACM. 2012, pp. 243–252.

[5] Emre Kultursay, Mahmut Kandemir, Anand Sivasubramaniam, and Onur Mutlu. "Evaluating STT-RAM as an energy-efficient main memory alternative". *ISPASS*. IEEE. 2013, pp. 256–267.

[6] S Kanai, M Yamanouchi, S Ikeda, Y Nakatani, F Matsukura, and H Ohno. "Electric field-induced magnetization reversal in a perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction". *Appl. Phys. Lett.* 101.12 (2012), p. 122403.

[7] Yoichi Shiota, Takayuki Nozaki, Frédéric Bonell, Shinichi Murakami, Teruya Shinjo, and Yoshishige Suzuki. "Induction of coherent magnetization switching in a few atomic layers of FeCo using voltage pulses". *Nature materials* 11.1 (2012), pp. 39–43.

[8] Wei-Gang Wang, Mingen Li, Stephen Hageman, and CL Chien. "Electric-field-assisted switching in magnetic tunnel junctions". *Nature materials* 11.1 (2012), pp. 64–68.

[9] S. Wang and S. Pal and T. Li and A. Pan and C. Grezes and P. Khalili-Amiri and K. L. Wang and P. Gupta. "Hybrid VC-MTJ/CMOS non-volatile stochastic logic for efficient computing". *Design, Automation Test in Europe Conference Exhibition (DATE), 2017*. Mar. 2017, pp. 1438–1443. DOI: 10.23919/DATE.2017.7927218.

[10] Shaodi Wang, Hochul Lee, Farbod Ebrahimi, P. Khalili Amiri, Kang L. Wang, and Puneet Gupta. "Comparative Evaluation of Spin-Transfer-Torque and Magnetoelectric Random Access Memory". *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 6 (2016).

[11] Shaodi Wang. "Design, Evaluation and Co-optimization of Emerging Devices and Circuits". PhD thesis. University of California, Los Angeles, 2017.

[12] Cecile Grezes, Hochul Lee, Albert Lee, Shaodi Wang, Farbod Ebrahimi, Xiang Li, Kin Wong, Jordan A Katine, Berthold Ocker, Juergen Langer, et al. "Write Error Rate and Read Disturbance in Electric-Field-Controlled MRAM". *IEEE Magnetics Letters* 8 (2016).

[13] Jing Li, Haixin Liu, Sayeef Salahuddin, and Kaushik Roy. "Variation-tolerant Spin-Torque Transfer (STT) MRAM array for yield enhancement". *Proc. CICC*. IEEE. 2008, pp. 193–196.

[14] Peiyuan Wang, Wei Zhang, Rajiv Joshi, Rouwaida Kanj, and Yiran Chen. "A thermal and process variation aware MTJ switching model and its applications in soft error analysis". *Proc. ICCAD*. IEEE. 2012, pp. 720–727.

[15] Y Eckert, Nuwan Jayasena, and G Loh. "Thermal feasibility of die-stacked processing in memory". *Proceedings of the 2nd Workshop on Near-Data Processing*. 2014.

[16] Jong-Yoon Park, Se-Koo Kang, Min-Hwan Jeon, Myung S Jhon, and Geun-Young Yeom. "Etching of CoFeB Using CO/ NH3 in an Inductively Coupled Plasma Etching System". *J. Electrochem. Soc* 158.1 (2011), H1–H4.

[17] Said Tehrani, JM Slaughter, E Chen, M Durlam, J Shi, and M DeHerren. "Progress and outlook for MRAM technology". *TMAG* 35.5 (1999), pp. 2814–2819.

[18] JM Slaughter, EY Chen, R Whig, BN Engel, J Janesky, and S Tehrani. "Magnetic tunnel junction materials for electronic applications". *JOM(USA)* 52.6 (2000), p. 11.

[19] Yanfeng Jiang, Yisong Zhang, Angeline Klemm, and Jian-Ping Wang. "Fast Spintronic Thermal Sensor for IC Power Driver Cooling Down". *Proc. IEDM*. 2016.

[20] Volker Drewello, J Schmalhorst, Andy Thomas, and Günter Reiss. "Evidence for strong magnon contribution to the TMR temperature dependence in MgO based tunnel junctions". *Physical Review B* 77.1 (2008), p. 014440.

[21] Shaodi Wang, Hochul Lee, Cecile Grezes, Pedram Khalili, Kang L Wang, and Puneet Gupta. "MTJ variation monitor-assisted adaptive MRAM write". *Proceedings of the 53rd Annual Design Automation Conference*. ACM. 2016, p. 169.

[22] Y. J. Song, J. H. Lee, and et. al. "Highly Functional and Reliable 8Mb STT-MRAM Embedded in 28nm Logic". *Proc. IEDM*. 2016.

[23] R Sbiaa, SYH Lua, R Law, H Meng, R Lye, and HK Tan. "Reduction of switching current by spin transfer torque effect in perpendicular anisotropy magnetoresistive devices". *J. Appl. Phys.* 109.7 (2011), p. 07C707.

[24] Yue Zhang, Weisheng Zhao, Yahya Lakys, J-O Klein, Joo-Von Kim, Dafiné Ravelosona, and Claude Chappert. "Compact modeling of perpendicular-anisotropy CoFeB/MgO magnetic tunnel junctions". *TED* 59.3 (2012), pp. 819–826.

[25] KJ Lee, Olivier Redon, and Bernard Dieny. "Analytical investigation of spin-transfer dynamics using a perpendicular-to-plane polarizer". *Appl. Phys. Lett.* 86.2 (2005), p. 022505.

[26] Jae-Won Jang, Jongsun Park, Swaroop Ghosh, and Swarup Bhunia. "Self-correcting STTRAM under magnetic field attacks". *DAC*. IEEE. 2015, pp. 1–6.

[27] Ping Zhou, Bo Zhao, Jun Yang, and Youtao Zhang. "Energy reduction for STT-RAM using early write termination". *ICCAD*. IEEE. 2009, pp. 264–268.

[28] Shaodi Wang, Andrew Pan, Chi On Chui, and Puneet Gupta. "Tunneling Negative Differential Resistance-Assisted STT-RAM for Efficient Read and Write Operations". *IEEE Transactions on Electron Devices* 64.1 (2017), pp. 121–129.

[29] S. Wang, A. Pan, C. Grezes, P. Khalili Amiri, K. L. Wang, C. O. Chui, and P. Gupta. "Leveraging nMOS Negative Differential Resistance for Low Power, High Reliability Magnetic Memory". *IEEE Transactions on Electron Devices* 64.10 (Oct. 2017), pp. 4084–4090. ISSN: 0018-9383. DOI: 10.1109/TED.2017.2742500.

[30] Seyedhamidreza Motaman, Swaroop Ghosh, and Nitin Rathi. "Impact of process-variations in STTRAM and adaptive boosting for robustness". *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*. EDA Consortium. 2015, pp. 1431–1436.

[31] Tianhao Zheng, Jaeyoung Park, Michael Orshansky, and Mattan Erez. "Variable-energy write STT-RAM architecture with bit-wise write-completion monitoring". *Proceedings of the 2013 International Symposium on Low Power Electronics and Design*. IEEE Press. 2013, pp. 229–234.

[32] Rajendra Bishnoi, Fabian Oboril, Mojtaba Ebrahimi, and Mehdi B Tahoori. "Self-timed read and write operations in STT-MRAM". *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 24.5 (2016), pp. 1783–1793.

[33] Daisuke Suzuki, Masanori Natsui, Akira Mochizuki, and Takahiro Hanyu. "Cost-efficient self-terminated write driver for spin-transfer-torque RAM and logic". *IEEE Transactions on Magnetics* 50.11 (2014), pp. 1–4.

[34] Rajendra Bishnoi, Mojtaba Ebrahimi, Fabian Oboril, and Mehdi B Tahoori. "Improving write performance for STT-MRAM". *IEEE Transactions on Magnetics* 52.8 (2016), pp. 1–11.

[35] Wang Kang, Zheng Li, Jacques-Olivier Klein, Yuanqing Chen, Youguang Zhang, Dafiné Ravelosona, Claude Chappert, and Weisheng Zhao. "Variation-tolerant and disturbance-free sensing circuit for deep nanometer STT-MRAM". *IEEE Transactions on Nanotechnology* 13.6 (2014), pp. 1088–1092.

[36] Safeen Huda and Ali Sheikholeslami. "A novel STT-MRAM cell with disturbance-free read operation". *IEEE Transactions on Circuits and Systems I: Regular Papers* 60.6 (2013), pp. 1534–1547.

[37] Sparsh Mittal. "Mitigating Read-disturbance Errors in STT-RAM Caches by Using Data Compression". *arXiv preprint arXiv:1711.06790* (2017).

[38] Rajendra Bishnoi, Mojtaba Ebrahimi, Fabian Oboril, and Mehdi B Tahoori. "Read disturb fault detection in STT-MRAM". *Test Conference (ITC), 2014 IEEE International*. IEEE. 2014, pp. 1–7.

[39] Taemin Lee and Sungjoo Yoo. "Selective refresh to avoid read disturb errors in STT-RAM main memory". *SoC Design Conference (ISOCC), 2016 International*. IEEE. 2016, pp. 315–316.

[40] Arijit Raychowdhury. "Pulsed READ in spin transfer torque (STT) memory bitcell for lower READ disturb". *Nanoscale Architectures (NANOARCH), 2013 IEEE/ACM International Symposium on*. IEEE. 2013, pp. 34–35.

[41] Hochul Lee, Cécile Grezes, Shaodi Wang, Farbod Ebrahimi, Puneet Gupta, Pedram Khalili Amiri, and Kang L Wang. "Source line sensing in magneto-electric random-access memory to reduce read disturbance and improve sensing margin". *IEEE Magnetics Letters* 7 (2016), pp. 1–5.

[42] Juan G Alzate, Pedram Khalili Amiri, Guoqiang Yu, Pramey Upadhyaya, Jordan A Katine, Juergen Langer, Berthold Ocker, Ilya N Krivorotov, and Kang L Wang. "Temperature dependence of the voltage-controlled perpendicular anisotropy in nanoscale MgO— CoFeB— Ta magnetic tunnel junctions". *Appl. Phys. Lett.* 104.11 (2014), p. 112410.

[43] Shaodi Wang, Greg Leung, Andrew Pan, Chi On Chui, and Puneet Gupta. "Evaluation of digital circuit-level variability in inversion-mode and junctionless FinFET technologies". *TED* 60.7 (2013), pp. 2186–2193.

[44] K Watanabe, B Jinnai, S Fukami, H Sato, and H Ohno. "Shape anisotropy revisited in single-digit nanometer magnetic tunnel junctions". *Nature communications* 9.1 (2018), p. 663.

[45] P Khalili Amiri, P Upadhyaya, JG Alzate, and KL Wang. "Electric-field-induced thermally assisted switching of monodomain magnetic bits". *J. Appl. Phys.* 113.1 (2013), p. 013912.

[46] Y Higo, K Yamane, K Ohba, H Narisawa, K Bessho, M Hosomi, and H Kano. "Thermal activation effect on spin transfer switching in magnetic tunnel junctions". *Appl. Phys. Lett.* 87.8 (2005), pp. 082502–082502.

[47] Ching-Che Chung and Cheng-Ruei Yang. "An autocalibrated all-digital temperature sensor for on-chip thermal monitoring". *TCS* 58.2 (2011), pp. 105–109.

[48] Kyoungho Woo, Scott Meninger, Thucydides Xanthopoulos, Ethan Crain, Dongwan Ha, and Donhee Ham. "Dual-DLL-based CMOS all-digital temperature sensor for microprocessor thermal monitoring". *ISSCC*. IEEE. 2009, pp. 68–69.

[49] Poki Chen, Chun-Chi Chen, Yu-Han Peng, Kai-Ming Wang, and Yu-Shin Wang. "A time-domain SAR smart temperature sensor with curvature compensation and a 3σ inaccuracy of- 0.4 C + 0.6 C over a 0 C to 90 C range". *JSSC* 45.3 (2010), pp. 600–609.

[50] André L Aita, Michiel AP Pertijs, Kofi AA Makinwa, and Johan H Huijsing. "A CMOS smart temperature sensor with a batch-calibrated inaccuracy of$\pm$0.25 C (3σ) from-70 C to 130 C". *ISSCC*. IEEE. 2009, pp. 342–343.

[51] H. Lee, J.G. Alzate, R. Dorrance, X.Q. Cai, D. Markovic, P. Khalili Amiri, and K.L. wang. "Design of a Fast and Low-Power Sense Amplifier and Writing Circuit for High-Speed MRAM". *TMAG* 51.5 (May 2015), pp. 1–7.

[52] Xiangyu Dong, Cong Xu, Yuan Xie, and Norman P Jouppi. "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory". *ICCAD* 31.7 (2012), pp. 994–1007.

[53] Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R Hower, Tushar Krishna, Somayeh Sardashti, et al. "The gem5 simulator". *ACM SIGARCH Computer Architecture News* 39.2 (2011), pp. 1–7.

[54] Sheng Li, Jung Ho Ahn, Richard D Strong, Jay B Brockman, Dean M Tullsen, and Norman P Jouppi. "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures". *MICRO*. IEEE. 2009, pp. 469–480.

[55] Wei Huang, Shougata Ghosh, Siva Velusamy, Karthik Sankaranarayanan, Kevin Skadron, and Mircea R Stan. "HotSpot: A compact thermal modeling methodology for early-stage VLSI design". *TVLSI* 14.5 (2006), pp. 501–513.

[56] Shaodi Wang, Henry (Chaohong) Hu, Hongzhong Zheng, and Puneet Gupta. "MEMRES: A Fast Memory System Reliability Simulator". *IEEE Transactions on Reliability* 65.4 (2016), pp. 1783–1797.

[57] T Ishikawa, T Marukame, H Kijima, K-I Matsuda, T Uemura, M Arita, and M Yamamoto. "Spin-dependent tunneling characteristics of fully epitaxial magnetic tunneling junctions with a full-Heusler alloy Co 2 Mn Si thin film and a MgO tunnel barrier". *Applied physics letters* 89.19 (2006), p. 192505.

[58] Mario Blaum, Rodney Goodman, and Robert McEliece. "The reliability of single-error protected computer memories". *Computers, IEEE Transactions on* 37.1 (1988), pp. 114–119.

[59] V. Sridharan and D. Liberty. "A study of DRAM failures in the field". *High Performance Computing, Networking, Storage and Analysis (SC), 2012 International Conference for*. Nov. 2012, pp. 1–11. DOI: 10.1109/SC.2012.13.

[60] Vilas Sridharan, Nathan DeBardeleben, Sean Blanchard, Kurt B Ferreira, Jon Stearley, John Shalf, and Sudhanva Gurumurthi. "Memory Errors in Modern Systems: The Good, The Bad, and The Ugly". *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM. 2015, pp. 297–310.

Shaodi Wang (S'12) is currently a researcher in the NanoCAD lab at Department of Electrical Engineering, UCLA. Shaodi received the Ph.D. degree in electrical engineering from UCLA, in 2017, and the B.S. degree from Peking University in 2011.

Hochul Lee (S'13) received his B.S. in Electrical Engineering from Korea University, Seoul, South Korea in February 2005. In September 2005, he joined Semiconductor Material Device Lab (SMDL) in Seoul National University (SNU) to pursue his M.S degree. After graduation, he had worked for Samsung Electronics Flash memory circuit design team until July 2012. He joined in UCLA DRL and is currently a Ph.D. candidate exploring MTJs based hybrid CMOS circuit.

Cecile Grezes (M'15) received the B.Sc. degree in Physics and Mathematics from the Université Joseph Fourier, Grenoble, in 2008, the M.Sc. in Physics from the Ecole Normale Supérieure, Paris, in 2011, and the Ph.D. degree (cum laude) in physics from CEA Saclay/Université Pierre et Marie Curie, Paris in 2014.

Pedram Khalili Amiri (M'05) received the B.Sc. degree from the Sharif University of Technology in 2004 and the Ph.D. degree (cum laude) in electrical engineering from Delft University of Technology in 2008. He is an Assistant Adjunct Professor at the EE Dept. of University of California at Los Angeles since 2009.

Kang L. Wang (F'92) received the B.S. degree from the National Cheng Kung University, Taiwan, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently a Distinguished Professor and holds the Raytheon Chair Professor in physical science and electronics with the Electrical Engineering Department, University of California at Los Angeles, Los Angeles, CA, USA.

Puneet Gupta (M'07-SM'16) is currently a faculty member of the Electri-cal Engineering Department at UCLA. He received the B.Tech degree in Electrical Engineering from Indian Institute of Technology, Delhi in 2000 and Ph.D. in 2007 from University of California, San Diego. He co-founded Blaze DFM Inc. (acquired by Tela Inc.) in 2004 and served as its product architect till 2007.