# MTJ Variation Monitor-assisted Adaptive MRAM Write

Shaodi Wang
shaodiwang@g.ucla.edu

Hochul Lee
chul0524@ucla.edu

Cecile Grezes
grezes.cecile@gmail.com

Pedram Khalili
pedramk@ucla.edu

Kang L. Wang
wang@ee.ucla.edu

Puneet Gupta
puneet@ee.ucla.edu

Department of Electrical Engineering, University of California, Los Angeles
Los Angeles, CA 90095

## ABSTRACT

Spin-transfer torque random access memory (STT-RAM) and magnetoelectric random access memory (MeRAM) are promising non-volatile memory technologies. But STT-RAM and MeRAM both suffer from high write error rate due to thermal fluctuation of magnetization. Temperature and wafer-level process variation significantly exacerbate these problems. In this paper, we propose a design that adaptively selects optimized write pulse for STT-RAM and MeRAM to overcome ambient process and temperature variation. To enable the adaptive write, we design specific 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. With adaptive write, the write latency of STT-RAM and MeRAM cache are reduced by up to 17% and 59% respectively, and application run time is improved by up to 41%.

## Keywords

MeRAM; STT-RAM; adaptive write; thermal monitor; process variation; temperature variation; MTJ

## 1. INTRODUCTION

Magnetoresistive random access memory (MRAM) [1] using magnetic tunnel junctions (MTJ)s is a promising data storage technology due to its non-volatility, zero leakage power, and high endurance. Spin-transfer torque RAM (STT-RAM) designed with MTJs switched by Spin-transfer torque (STT-MTJ) [2, 3] is identified as a possible replacement of current memory technologies, such as static RAM (SRAM) cache [4, 5] and Dynamic RAM (DRAM) memory [6]. The recent development of voltage-controlled MTJs (VC-MTJ)s with voltage-controlled magnetic anisotropy (VCMA) provides more promising performance [7–9]. This technology allows for precessional switching, a process which provides flipping of the magnetization upon a voltage pulse, irrespective of the initial state. It enables the use of minimum sized access transistors, as well as precessional switching to simultaneously achieve low energy ($1fJ/bit$), high density ($6F^2$) and high speed (<1ns of switching) magnetoelectric random access memory (MeRAM).



*DAC '16, June 05-09, 2016, Austin, TX, USA*



DOI: http://dx.doi.org/10.1145/2897937.2897979

However, both STT-RAM and MeRAM face the challenge of high write error rate (WER) due to thermal fluctuation. Increasing write current and time reduces the WER of STT-RAM at the expense of high write power, large access transistors, and long write latency. For MeRAM, there is no straightforward method to reduce WER.

Process and temperature variation further exacerbates the problems [10–13]. Local variation causing MTJ diameter and oxide tunnel barrier thickness changes leads to resistance change or MTJ failure [14]. Compared with local variation (standard deviation of MTJ resistance is 1.5% in a 4-Mb MRAM array [15]), wafer-level variations, including thickness variation of free layer and oxide tunnel barrier layer, more severely affect MTJ performance [1, 16]. The wafer-level free layer thickness variation can dramatically change energy barrier in free layer and thermal stability, especially for out-of-plane MTJs, which face less challenge of fabrication and switching energy compared with in-plane MTJs [17–19]. Temperature variation during operation also affects energy barrier, STT and VCMA effect. Temperature and process variation together can change the energy barrier by 200%, indicating that extreme high write energy is required if STT-RAM is designed for worst process and temperature corner. Differently from STT-RAM, MeRAM requires prise write voltage to achieve the least WER, but the voltage varies with energy barrier and hence is sensitive to process and temperature variation.

We propose an adaptive write scheme which selects optimized write pulse for STT-RAM and MeRAM to achieve faster write speed based on run-time variation sensing. We also design an MTJ-based variation monitor utilizing thermal activation and VCMA effect. The monitor enables in-situ process and temperature variation sensing. The monitor achieves remarkable area, power, and latency improvement compared with conventional on-chip thermal monitors.

Our contributions are summarized as follows.

- We have designed an MTJ-based variation monitor to sense process and temperature variation. 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%.

## 2. RELATED WORK AND BACKGROUND

Two frameworks [10, 20] are proposed to minimize STT-RAM failures caused by process variation to improve yield. A MTJ-based sensor is proposed in [21] to sense magnetic field attacking to STT-RAM. However, this monitor requires more advanced MTJs with smaller dimension than STT-RAM array to be protected, which also introduces fabrication difficulty of printing different sized MTJs on single die. In [22], an early write termination methodology is proposed to complete STT-RAM 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 is not able to assist MeRAM due to its long sensing latency of over 0.5ns.

STT-MTJ and VC-MTJ are resistive memory devices and share a similar device structure, their resistance is determined by the magnetization directions of two ferromagnetic layers. The direction of one layer is fixed (referred to reference layer) while the other one can be switched (referred to 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 resistance difference is quantified by tunnel magnetoresistance ratio (TMR, defined as $(R_H - R_L)/R_L$), where TMR of over 300% [23] has been demonstrated. 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) [17–19].

STT-MTJ is switched by bidirectional current, while VC-MTJ is switched by one-directional voltage pulse. Fig. 1 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.

## 3. WRITE ERROR UNDER VARIATION

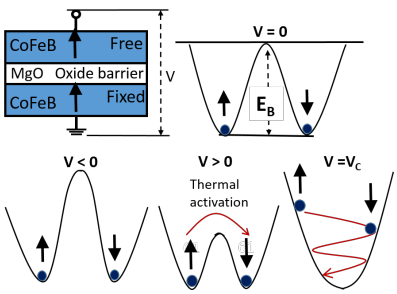


Figure 1: VCMA-induced precessional switching. A positive (negative) voltage on a MTJ reduces (increases) the energy barrier separating the two magnetization states. A full energy barrier reduction leads to precessional switching.

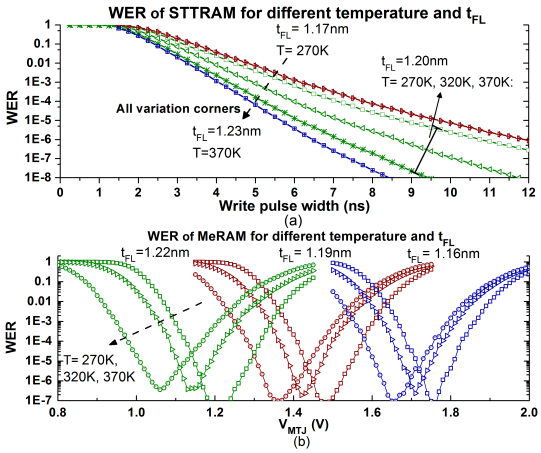


Figure 2: (a) The STT-RAM P-to-AP WER as a function of write pulse width under different $t_{FL}$ and temperature corners. In STT-RAM, P-to-AP switching is more difficult and dominates write latency. (b) The average AP-to-P and P-to-AP WER of MeRAM as a function of write voltage.

The switching behavior of STT-RAM and MeRAM are affected by temperature and free layer thickness ($t_{FL}$) [11, 24]. We simulate the WER of STT-RAM and MeRAM under different $t_{FL}$ and temperature corners using an LLG-based numerical model[1] including temperature dependence, VCMA effect, STT effect, and thermal fluctuation, which has been verified against experimental data in [12]. The $t_{FL}$ variation are assumed to be within 5% across wafer [16]. The temperature varies from 270K to 370K. Resistance variation (due to MTJ shape change) has limited impact on write behavior (i.e., STT-MTJ has low resistance, and its write current is mainly determined by access transistors, while the high resistance of VC-MTJ drops over 95% supply voltage with negligible variation) and is simply treated as random Gaussian variation in the simulations together with variation of access transistors [25] due to line edge roughness, random doping fluctuation, and non-rectangular gate effect.

The WER of STT-RAM and MeRAM under different temperature and $t_{FL}$ corners are shown in Fig. 2. The variation can shift WER by over 1,000X. The WER of STT-RAM is mainly affected by temperature, while MeRAM is strongly affected by both $t_{FL}$ and temperature. WER reduction requires to choose appropriate write pulse adaptively for MRAM array according to its temperature and process variation. One conventional solution is exhausted chip variation test and in-situ temperature monitor [26–29] placement in MRAMs.

## 4. VARIATION MONITOR

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

### 4.1 Sensing principle

Monitoring variation through directly measuring WER 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.

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

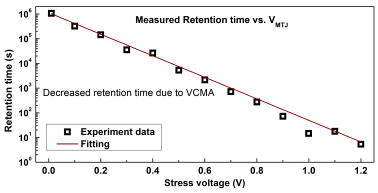


Figure 3: The experimentally measured retention time as a function of stress voltage on MTJs.

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

As described by (1) [30, 31], 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 ($\Delta$, proportional to energy barrier), critical current of STT-MTJs ($I_C(\Delta)$), and critical voltage of VC-MTJs ($V_C(\Delta)$). The write pulse width (determined by ($I_C(\Delta)$ and $\Delta$) and voltage ($V_C(\Delta)$) of STT-MTJs and VC-MTJs also depend on $\Delta$. This indicates that knowing the $t_{R,STT}$ and $t_{R,VC}$ change due to temperature and process variation can predict the MRAM write behavior change. Retention time of MTJs is too long to be measured directly. Fortunately, as illustrated by the Eqn. (1), applying current/voltage on MTJs reduces retention time exponentially giving rise to a possible way of measurement. We utilize this observation in the proposed variation monitor and call such applied voltage/current stress voltage/current for simplicity. This observation is demonstrated in experiment measurement, where retention time decreases exponentially with increasing stress voltage due to VCMA effect in Fig. 3.

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

When the retention time reduces to sub-$\mu s$, the MTJ switching rate ($P_{SW}$) due to thermal activation during under 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.

## 4.2 Circuit implementation and simulation

The principle of the proposed MTJ-based variation monitor is to obtain switching rate of an MTJ array after a stress operation (applying a stress voltage and current for 20ns). If the switching rate reaches preset 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. 4. In a stress operation, all MTJs in the monitor are in high resistance state initially. The write control circuit applies a stress current (for STT-RAM) 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) in the stress voltage selection array. 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

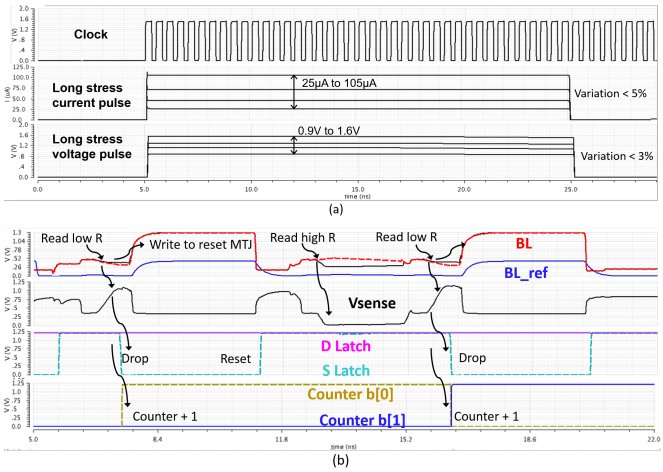


Figure 5: (a) Different stress current/voltage in the proposed monitor. (b) Simulated waveforms of read, reset and counting operations.

each MTJ one by one and reads its state. In the read, the bit-line ($BL$) and reference bit-line ($BL\_ref$) are precharged 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 rises 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 commercial library. The stress pulses are shown in Fig. 5 (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. 5 (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. The second read is on a non-switched MTJ, and hence no action is taken after the read. If the counted number reaches preset 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 preset 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 switching rate and standard deviation ($\sigma$) of a 256-MTJ variation monitor with different stress levels and variation corners as shown in Fig. 6. In these curves, if we select a preset threshold between 10% to 30%, the voltages to reach the threshold under 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. Therefore, for a given constant $t_{FL}$, ten stress levels can achieve accuracy of $10^oC$.

Table 1 shows the comparison between the proposed variation monitor with conventional thermal monitors. The conventional monitors target on high precision, where long latency and high energy are consumed by analog-to-digital blocks and bipolar sensing transistors. The proposed monitor has less accuracy but faster speed, lower energy/sample,

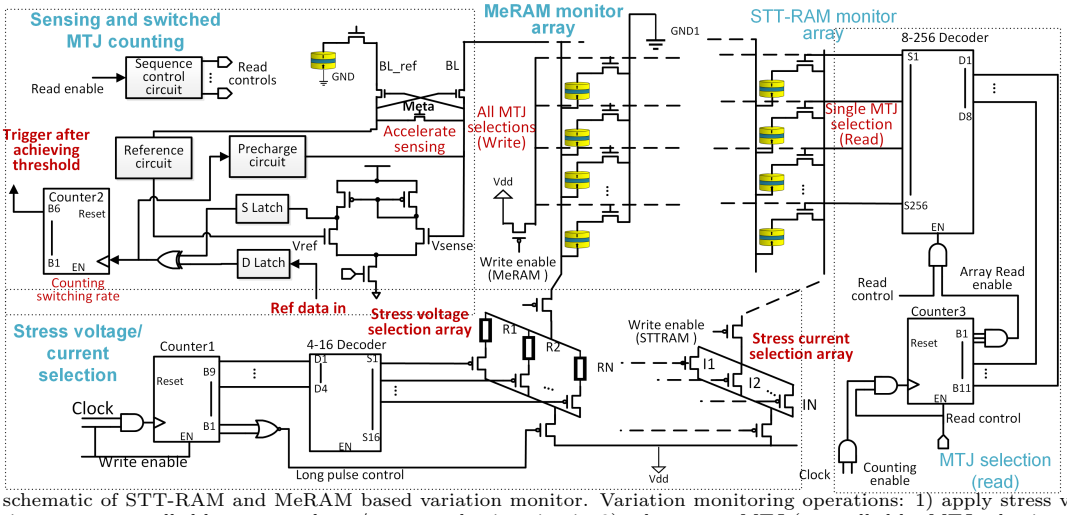


Figure 4: The schematic of STT-RAM and MeRAM based variation monitor. Variation monitoring operations: 1) apply stress voltage/current on MRAM monitor array controlled by stress voltage/current selection circuit; 2) select every MTJ (controlled by MTJ selection circuit) one by one to read and count MTJ switching rate (controlled by sensing and switched MTJ counting circuit).

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 [26] | 0.1ms | $9^oC$ | $0.015\mu J$ | $0.01mm^2$ |
| S2 [27] | 0.2ms | $3^oC$ | $0.24\mu J$ | $0.04mm^2$ |
| S3 [28] | 1ms | $2^oC$ | $0.49\mu J$ | $0.01mm^2$ |
| S4 [29] | 100ms | $0.1^oC$ | $13.8\mu J$ | $0.04mm^2$ |
| this(STT) | 1-10$\mu s$ | $10^oC$ | 0.12-1.2$nJ$ | $0.0005mm^2$ |
| this(Me) | 1-10$\mu s$ | $10^oC$ | 0.27-2.7$nJ$ | $0.0005mm^2$ |

and smaller area. Its accuracy can be improved by using more MTJs to reduce $\sigma$ of curves in Fig. 6 as well as using finer grids of stress levels in the monitor, which quadratically increases sensing energy and latency. In addition, finer grids of stress current/voltage require less process variation in circuit, which is also the accuracy limitation. Fortunately, selecting optimal write pulse for STT-RAM and MeRAM does not require high accuracy (i.e., Section 5.1 shows that three stress levels are enough) indicating that the proposed monitor is well suited to the adaptive write selection with the least overhead. The area of the monitor is dominated by

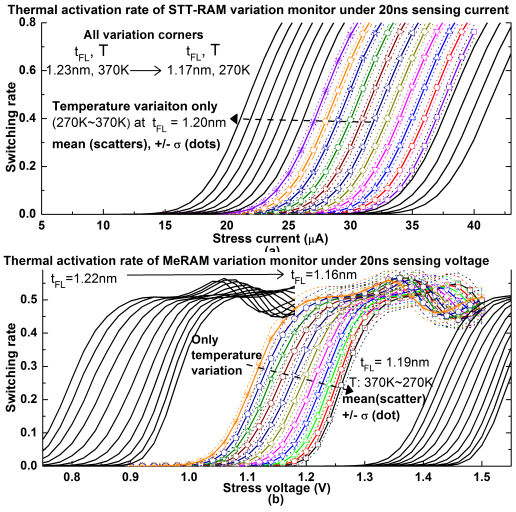


Figure 6: Switching rate of (a) STT-MTJ- and (b) VC-MTJ-based variation monitor under different stress current and voltage respectively. The color lines are switching rate for only temperature variation ($10^oC$ interval). The dot lines outline standard deviations ($\sigma$) of thermal activation rate ($\sigma$ is caused by process variation and random thermal activation).

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.

Though the wafer-level resistance variation of STT-RAM is not considered in the simulation, but can also be partially monitored because the stress voltage/current shift induced by resistance variation is proportional to write voltage/current shift.

# 5. ADAPTIVE WRITE

## 5.1 Adaptive write scheme

The adaptive write scheme is to dynamically select an optimized pulse width (voltage) for STT-RAM (MeRAM) out of multiple voltage (current) choices to minimize write latency according to ambient variation. Creating multiple pulse widths uses simple delay circuits, which is shared by multiple bit-lines with negligible overhead. Multiple write pulse voltage requires multiple voltage regulators, and the regulators can be shared by the entire MRAM array. Temperature variation over MRAM array [13] can be captured by placing multiple proposed monitors to monitor local variation. One such monitor only uses one bit-line in MRAM boundary with an area overhead of <0.005% (i.e., adding monitor control 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).

Schemes to make optimized write pulse selections with and without the proposed variation monitor are shown in Fig. 7. With the variation monitor, write pulse is selected according to output variation level. Without the variation monitor, exhaustively memory chip test is required for each chip to obtain and store optimized pulses for different tem-

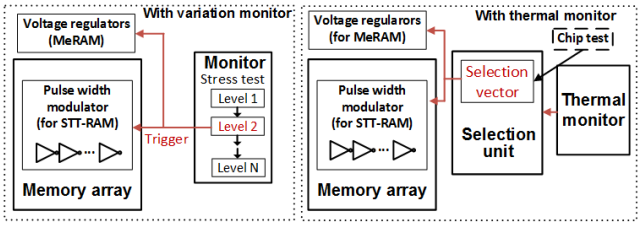


Figure 7: Adaptive write scheme using the MTJ-based variation monitor or conventional thermal monitors.

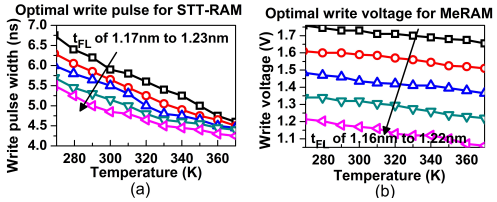


Figure 8: Optimal write pulses for (a) STT-RAM and (b) MeRAM under different $t_{FL}$ and temperature corners.

perature, and a conventional thermal monitor is required to make dynamic pulse selection

## 5.2 Adaptive write using variation monitor

In this section, we evaluate the write scheme with the proposed variation monitor. The write circuit for MRAM is implemented with read check function [32] which performs a read check following a write (the writing data is pre-stored in D Latch in Fig. 4), and a write error gives rise to additional writes until all errors are fixed. With this, WER of 0 is guaranteed for MeRAM and STT-RAM irrespective of the single write pulse voltage/width. For STT-RAM, shortening single write pulse reduces latency and energy, as a trade-off, WER of the write and chance of additional writes increase, which add overall latency and energy. Hence, there is an optimal single write pulse achieving minimum expected latency, and it can be found given a WER function of pulse voltage/width. Such optimal pulse can reduce STT-RAM's expected latency and energy by over 60% compared with conventional write circuit [12]. The optimal pulse width (voltage) for minimum expected latency (including initial write, read checks, and additional writes) of STT-RAM (MeRAM) are shown in Fig. 8. The pulse width for STT-RAM 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 output (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 write latency.

Our evaluation flow is illustrated in Fig. 9 (a). We simulate the peripheral circuit (see Fig. 4) with a bit-line size of 256 MTJs using 32nm commercial library and simulate the WER of MTJs with LLG-based numerical model. The bit-line-level write latency varies from 5.5ns to 7.5ns for STT-RAM and 4 to 10.1ns for MeRAM for all variation corners and number for write pulses (1 to 5). With the inputs of bit-line results, we use NVSIM [33] to obtain latency and energy of MRAM array (cache). In Fig. 10, the write latency of

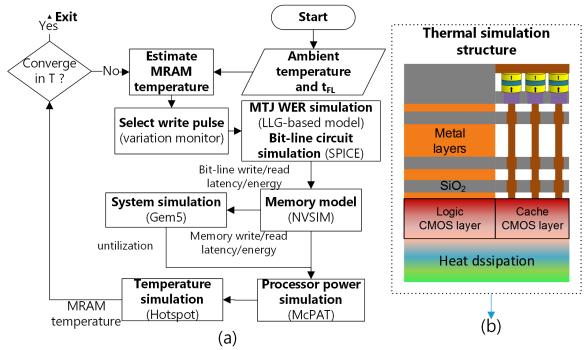


Figure 9: (a) Evaluation flow of adaptive write in MRAM based system. (b) The cross-section structure for thermal simulations.

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 of temperature corners of 270K to 370K. MeRAM's write latency reduction is up to 59%. There is a latency increase for $t_{FL}$ of 1.19nm using from one to two voltage choices, because that 1.19nm $t_{FL}$ corner is closer to optimized voltage when only one write voltage is used (see Fig. 2b). The write latency of STT-RAM is improved by up to 17%. The maximum latency for $t_{FL}$ corner of 1.17nm is not seen improvement because the corner with 1.17nm $t_F$ and 270K is always the worst corner to be optimized in its variation corner group no matter how many choices is adapted. As seen, three choices are efficient enoughfor write latency improment.

We modified gem5 [34] 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 [35] to simulate processor power and used Hotspot [36] to simulate MRAM temperature with the structure shown in Fig. 9b.

We simulated one billion instructions of SPEC benchmarks using our evaluation flow. The application run time reduction with adaptive write are shown in Fig. 11. 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-RAM respectively. However, the improvement are much less for processors with MRAM L2 and L3 (up to 10% and 2% for MeRAM and STT-RAM 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-RAM write latency improvement is not significant. Actually, the write energy is more crucial issue for high-speed STT-RAM cache (e.g., write latency within 3 ns), where large write current is required and sensitive to variation. Our future work will evaluate the adaptive write scheme in STT-RAM energy reduction.

# 6. CONCLUSION

We design an MTJ-based variation monitors 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 the write latency of STT-RAM and MeRAM according to ambient process and temperature variation. The write latency of STT-RAM and MeRAM cache is reduced up to 17% and 59% respectively, while simulated application run time is shown up to 1.7X improvement. We expect this technique to significantly speedup embedded processors with MeRAM memory, or to reduce energy dramatically for processors with high-speed STT-RAM. Our future work is looking at these applications.

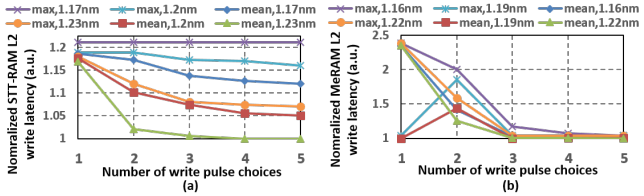


Figure 10: The maximum and average write latency in (a) 1MB STT-RAM L2 and (b) MeRAM L2 from 270K to 370K under different $t_{FL}$ corners with different number of write pulse choices.

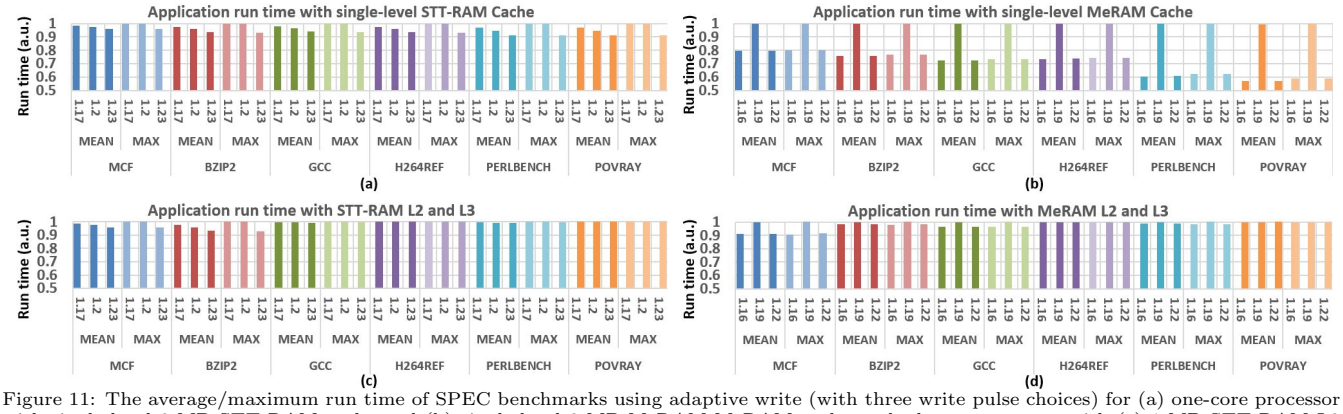


Figure 11: The average/maximum run time of SPEC benchmarks using adaptive write (with three write pulse choices) for (a) one-core processor with single-level 8-MB STT-RAM cache and (b) single-level 8-MB MeRAM MeRAM cache, a dual-core processor with (c) 1-MB STT-RAM L2 and 16-MB STTRAM L3, and (d) 1-MB 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.

## Acknowledgment

The authors would like to thank Yang Zhang and Mark Gottscho for their help in evaluation experiment setup.

## References

[1] S. Tehrani et al. "Progress and outlook for MRAM technology". *TMAG* (1999).

[2] C. Heide. "Spin currents in magnetic films". *Phys. Rev. Lett.* (2001).

[3] D. Worledge et al. "Spin torque switching of perpendicular Ta| CoFeB| MgO-based magnetic tunnel junctions". *Appl. Phys. Lett.* (2011).

[4] C. W. Smullen et al. "Relaxing non-volatility for fast and energy-efficient STT-RAM caches". *HPCA*. IEEE. 2011.

[5] A. Jog et al. "Cache revive: architecting volatile STT-RAM caches for enhanced performance in CMPs". *Proc. DAC*. ACM. 2012.

[6] E. Kultursay et al. "Evaluating STT-RAM as an energy-efficient main memory alternative". *ISPASS*. IEEE. 2013.

[7] S. Kanai et al. "Electric field-induced magnetization reversal in a perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction". *Appl. Phys. Lett.* (2012).

[8] Y. Shiota et al. "Induction of coherent magnetization switching in a few atomic layers of FeCo using voltage pulses". *Nature materials* (2012).

[9] W.-G. Wang et al. "Electric-field-assisted switching in magnetic tunnel junctions". *Nature materials* (2012).

[10] J. Li et al. "Variation-tolerant Spin-Torque Transfer (STT) MRAM array for yield enhancement". *Proc. CICC*. IEEE. 2008.

[11] P. Wang et al. "A thermal and process variation aware MTJ switching model and its applications in soft error analysis". *Proc. ICCAD*. IEEE. 2012.

[12] S. Wang et al. "Comparative Evaluation of Spin-Transfer-Torque and Magnetoelectric Random Access Memory". *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* (2016).

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

[14] J.-Y. Park et al. "Etching of CoFeB Using CO/ NH3 in an Inductively Coupled Plasma Etching System". *J. Electrochem. Soc* (2011).

[15] R. W. Dave et al. "MgO-based tunnel junction material for high-speed toggle magnetic random access memory". *TMAG* (2006).

[16] J. Slaughter et al. "Magnetic tunnel junction materials for electronic applications". *JOM(USA)* (2000).

[17] R. Sbiaa et al. "Reduction of switching current by spin transfer torque effect in perpendicular anisotropy magnetoresistive devices". *J. Appl. Phys.* (2011).

[18] Y. Zhang et al. "Compact modeling of perpendicular-anisotropy CoFeB/MgO magnetic tunnel junctions". *TED* (2012).

[19] K. Lee, O. Redon, and B. Dieny. "Analytical investigation of spin-transfer dynamics using a perpendicular-to-plane polarizer". *Appl. Phys. Lett.* (2005).

[20] Y. Zhang, X. Wang, and Y. Chen. "STT-RAM cell design optimization for persistent and non-persistent error rate reduction: a statistical design view". *Proc. ICCAD*. IEEE. 2011.

[21] J.-W. Jang et al. "Self-correcting STTRAM under magnetic field attacks". *DAC*. IEEE. 2015.

[22] P. Zhou et al. "Energy reduction for STT-RAM using early write termination". *ICCAD*. IEEE. 2009.

[23] Y. M. Lee et al. "Giant tunnel magnetoresistance and high annealing stability in CoFeB/MgO/CoFeB magnetic tunnel junctions with synthetic pinned layer". *arXiv preprint cond-mat/0606503* (2006).

[24] J. G. Alzate et al. "Temperature dependence of the voltage-controlled perpendicular anisotropy in nanoscale MgO| CoFeB| Ta magnetic tunnel junctions". *Appl. Phys. Lett.* (2014).

[25] S. Wang et al. "Evaluation of digital circuit-level variability in inversion-mode and junctionless FinFET technologies". *TED* (2013).

[26] C.-C. Chung and C.-R. Yang. "An autocalibrated all-digital temperature sensor for on-chip thermal monitoring". *TCS* (2011).

[27] K. Woo et al. "Dual-DLL-based CMOS all-digital temperature sensor for microprocessor thermal monitoring". *ISSCC*. IEEE. 2009.

[28] P. Chen et al. "A time-domain SAR smart temperature sensor with curvature compensation and a $3\sigma$ inaccuracy of- 0.4 C + 0.6 C over a 0 C to 90 C range". *JSSC* (2010).

[29] A. L. Aita et al. "A CMOS smart temperature sensor with a batch-calibrated inaccuracy of$\pm$0.25 C ($3\sigma$) from- 70 C to 130 C". *ISSCC*. IEEE. 2009.

[30] P. K. Amiri et al. "Electric-field-induced thermally assisted switching of monodomain magnetic bits". *J. Appl. Phys.* (2013).

[31] Y. Higo et al. "Thermal activation effect on spin transfer switching in magnetic tunnel junctions". *Appl. Phys. Lett.* (2005).

[32] H. Lee et al. "Design of a Fast and Low-Power Sense Amplifier and Writing Circuit for High-Speed MRAM". *TMAG* (2015).

[33] X. Dong et al. "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory". *ICCAD* (2012).

[34] N. Binkert et al. "The gem5 simulator". *ACM SIGARCH Computer Architecture News* (2011).

[35] S. Li et al. "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures". *MICRO*. IEEE. 2009.

[36] W. Huang et al. "HotSpot: A compact thermal modeling methodology for early-stage VLSI design". *TVLSI* (2006).