DOVA: A Dynamic Overwriting Voltage Adjustment for STT-RAM L1 Cache

Jinbo Chen, Keren Liu, Xiaochen Guo, Patrick Girard, Yuanqing Cheng

DOVA: A Dynamic Overwriting Voltage Adjustment for STT-RAM L1 Cache

Jinbo Chen*, Keren Liu*, Xiaochen Guo†, Patrick Girard‡ and Yuanqing Cheng§
*School of Electrical and Information Engineering, Beihang University, Beijing, China 100191
†Department of Electrical & Computer Engineering, Lehigh University, West Bethlehem, PA, USA 18015
‡LIRMM, CNRS, Montpellier, France 34095
§School of Microelectronics, Beihang University, Beijing, China 100191
Email: yuanqing@ieee.org

***Abstract*—As device integration density increases exponentially predicted by Moore's law, power consumption becomes a bottleneck for system scaling. On the other hand, leakage power of on-chip cache occupies a large fraction of the total power budget. STT-RAM is a promising candidate to replace SRAM as on-chip cache due to its ultra-low leakage power, high integration density and non-volatility. However, building L1 cache with STT-RAM still faces severe challenges especially because of its high write latency and energy overheads. Moreover, intensive accesses in L1 cache accelerate oxide breakdown and threaten the lifetime of STT-RAM significantly. In this paper, we propose a Dynamic Overwriting Voltage Adjustment (DOVA) technique for STT-RAM L1 cache. A high write voltage is used for performance critical cache lines while a low write voltage is used for other cache lines to approach an optimal trade-off between reliability and performance. Experimental results show that the proposed technique can improve cache performance up to 18%, and 9% on average with almost the same reliability level as in the case when only the low write voltage is used.**

I. Introduction

The shrinking of transistor feature size enables fast switching speed and high on-chip integration density. However, the off-chip memory bandwidth cannot improve at the same pace. To fill the performance gap between processor and main memory, on-chip cache size increases quickly and introduces large area and power overhead. STT-RAM is a promising technique to replace SRAM as on-chip cache due to small cell size, fast access speed and ultra-low leakage power [1].

Most existing work focuses on using STT-RAM to replace SRAM as last level cache but few works focus on totally non-volatile cache hierarchy design. The main reason is that the write procedure of STT-RAM is time and power-consuming. L1 cache is usually write-intensive and affects processor performance dramatically.

Moreover, with the aggravating "Power Wall" issue, it is desirable to build a non-volatile memory hierarchy to save energy and keep data persistent to approach ultra low power design [2]. Recently, several non-volatile processor architectures have been proposed such as [3], which reflects this emerging trend.

In this work, we explore the possibility of using STT-RAM as L1 cache. Since the write speed of STT-RAM depends on the write voltage magnitude, we can increase the write speed by increasing the write voltage. However, as L1 cache is more write-intensive compared to lower level caches, high write voltage degrades STT-RAM lifetime significantly due to the TDDB (Time-Dependent Dielectric Breakdown) effect [4]. High voltage also incurs more write energy which may swallow the energy savings brought by STT-RAM. To deal with this problem, we propose a Dynamic Overwriting Voltage Adjustment (DOVA) technique,which classifies write operations in L1 cache as critical writes and non-critical writes. The critical writes adopt a high write voltage to reduce write latency while non-critical ones use a low voltage to save write energy and prolong STT-RAM lifetime. To the best of our knowledge, this is the first work to explore STT-RAM based L1 cache design considering reliability, performance and energy together. The main contributions are listed as follows:

- We observe that the write speed of STT-RAM depends on the write voltage and evaluate the dependency of STT-RAM lifetime on write voltage quantitatively. Then, we build a stochastic model to capture this relationship.
- To select write voltage effectively, L1 writes are classified into critical writes and non-critical writes, and a write voltage adjustment technique is proposed to identify them such that proper write voltages can be selected.
- Experimental results on SPEC2006 benchmarks running on a 4-core CPU show that DOVA can achieve up to 18% performance improvement and 9% improvement on average only incurring 1.74% average degradation for 5-year failure probability, 6.51% average energy consumption increase and negligible storage overhead compared to using the low write voltage only.

The rest of the paper is organized as follows: Section II introduces background knowledge and the dielectric breakdown mechanism of STT-RAM. Section III firstly establishes an STT-RAM reliability model by investigating the relationship of dielectric breakdown and write voltage quantitatively. Then, a statistical analysis on write operations in L1 cache is presented to motivate our work. Section IV details the implementation of DOVA. Section V describes the experiment results in terms of write energy, performance and reliability. Section VI summarizes related work and Section VII concludes the paper.

Jinbo Chen and Keren Liu contributed equally to this work.

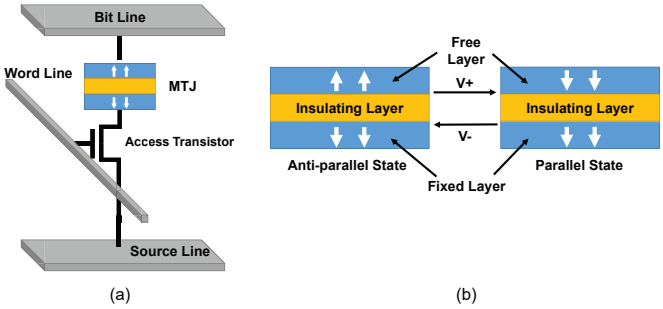

Fig. 1. The illustration of (a) the 1T-1MTJ structure and (b) the MTJ.

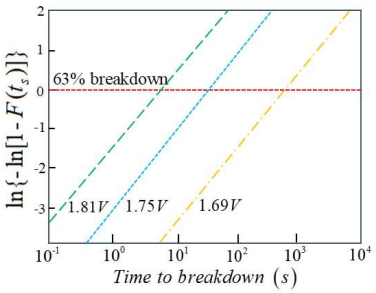

Fig. 2. TDDB effect of STT-RAM characterized by Weibull distribution (reproduced from [6]).

II. Background

A. Introduction to STT-RAM

The commonly used STT-RAM cell is made of an access transistor and an MTJ, which is shown in Fig. 1(a). The access transistor controls reads and writes of STT-RAM. The MTJ (Magnetic Tunnel Junction) is mainly composed of a fixed layer, a free layer and an insulating (oxide) layer. The fixed layer has fixed magnetization after fabrication. The magnetization of the free layer can be switched by injecting a spin polarized current. If the magnetization of the free layer is the same as that of the fixed layer, the MTJ has a low resistance (logic '0'). Conversely, the MTJ has a high resistance (logic '1') as shown in Fig. 1(b).

B. Time Dependent Dielectric Breakdown (TDDB) of STT-RAM

Experiments confirmed that the breakdown of insulating layer within a MTJ highly depends on the write voltage, and constant high voltage stress is the main cause of the failure of MTJ [5]. A certain duration of high voltage stress can cause oxide trapping in the insulating (oxide) layer. In the oxide breakdown process, factors such as oxide thickness and impurities in the oxide layer also accelerate the formation of oxide trapping. The procedure is irreversible, and may form a conductive path between the fixed layer and the free layer, making the MTJ malfunction.

The MTJ breakdown procedure mentioned above is a time-dependent probabilistic event. Weibull distribution is widely used for reliability analysis and lifetime prediction of semiconductor devices, which can be used to describe the statistical characteristics of the breakdown with the increase of write voltage [6].

III. Modeling of STT-RAM Lifetime Dependence on the Write Voltage

A. Stochastic Modeling of the Lifetime of a Single MTJ Due to the TDDB Effect

The MTJ breakdown mechanism can be described by Weibull distribution as follows:

$$F(t_s) = 1 - e^{(-\frac{t_s}{t_{63\%}})^{\beta}} \tag{1}$$

where t_s denotes the voltage stress duration, namely the accumulated switching time of an MTJ, $t_{63\%}$ is the specific accumulated switching time when the breakdown probability approaches 63%, and β represents the shape parameter of the Weibull distribution. For a specific MTJ, β is only related to oxide thickness, process variation and independent of the stress voltage [6].

Weibull distribution can be converted into the following linear form:

$$\ln\{-\ln[1-F(t_s)]\} = \beta(\ln t_s - \ln t_{63\%}) \tag{2}$$

where β can be directly obtained by interpolation of experimental measurements.

Our reliability model is built on the basis of Weibull distribution and MTJ parameters from [6] (refer to details in Section IV.A).

Assume that the thickness of the MgO oxide layer is 1.25nm [6]. Let $\ln\{-\ln[1-F(t_s)]\} = 0$ to obtain the corresponding voltage stress time $t_{63\%}$. The results are shown in Table I.

As mentioned above, β in Weibull distribution formula is independent of the write voltage and only related to the thickness of the oxide layer and the process variation. Therefore, the approximation of β, i.e., $\bar{\beta}$, can be calculated by averaging multiple sets of experimental measurements. Then, we can get

TABLE I
$t_{63\%}$ AND β VALUES WITH DIFFERENT WRITE VOLTAGES

Write Voltage/V	$t_{63\%}$/s	β
1.81	9.802	1.278
1.75	49.457	1.402
1.69	264.023	1.455

$\bar{\beta}$ from the following expression:

$$\bar{\beta} = \frac{\beta_{1.81} + \beta_{1.75} + \beta_{1.69}}{3} = 1.3783 \tag{3}$$

Moreover, three sets of parameters shown in Table I can be used to calculate $t_{63\%}$ and β parameters under any write voltage with the "Voltage Power Law" [7]:

$$t_{63\%} = aV^{-N} \tag{4}$$

where a is a process-dependent parameter and N denotes the voltage acceleration factor.

According to the three pairs of $(t_{63\%}, V)$ obtained from Fig. 2, we can get the fitting results as follows:

$$a = 2.3 \times 10^{13} \qquad N = 48.01 \tag{5}$$

Thus we can obtain:

$$t_{63\%} = 2.3 \times 10^{13} \times V^{-48.01} \tag{6}$$

With $t_{63\%}$ and β, the Weibull distribution of the MTJ TDDB lifetime under different write voltages can be obtained. Taking the voltage stress duration as the input, the breakdown probability of an MTJ is obtained with the following formula:

$$F(t_s) = 1 - e^{-(\frac{t_s}{2.3\times10^{13}\times V^{-48.01}})^{1.3783}} \quad (7)$$

B. Reliability Model of One Cache Line Considering the TDDB Effect

Based on the cache raw lifetime metric proposed in [8] and the MTJ lifetime model built above, we propose a reliability model for a STT-RAM based cache line.

The read voltage of a MTJ is relatively low, and hence has little or no effect on the device failure. So only the impact of write voltage is considered in this work. From Weibull distribution, it is clear that the higher the write voltage is, the greater the MTJ failure probability is, under a given voltage stress duration.

Assume that the cache line size is 64 bytes, and each bit is implemented with the 1T-1MTJ cell structure. The cache raw lifetime is defined as the time of occurrence of the first MTJ failure in a cache line. So the cache line failure probability can be expressed as:

$$P_{cacheline} = 1 - (1 - P_{MTJ})^{512} \quad (8)$$

The failure probability of the whole cache is defined as the *average failure probability* of all cache lines, which is:

$$P_{cache} = \frac{1}{n}\sum P_{cacheline} \quad (9)$$

In the following work, we will use this model to evaluate the lifetime of STT-RAM based L1 cache.

IV. The Proposed Dynamic Overwriting Voltage Adjustment (DOVA) Technique

A. Motivation & Main Idea

In order to improve the performance of STT-RAM L1 cache, an intuitive method is to increase the write voltage since write latency decreases when write voltage increases. We adopted the MTJ model proposed in [9] to evaluate the dependency of write latency and write voltage. The MTJ model parameters are listed in Table II. We performed circuit simulations on 1T-1MTJ cell to obtain the write latency under different write voltages. Then we fed simulation results to NVSim [10] to get the cache line write latency under different voltages. NVSim results show that when write voltage increases from 1.18V to 1.81V, the write latency can be reduced from 6.78ns to 4.61ns, which is 32.01% reduction.

On the other hand, the increasing write voltage may aggravate the TDDB effect and lead to shortened STT-RAM lifetime according to our reliability model. Therefore, it is of great importance to find an acceptable trade-off between write performance and reliability.

There are some existing works focusing on optimizing the trade-off in non-volatile memories with adaptive write voltage. However, most of the existing work chooses to utilize customized circuits to track the critical-path writes in the instruction queue to apply adaptive voltage. These designs result in high hardware complexity and area overhead. On the contrary, this paper proposes a novel method from a statistical view focusing on the cache line access behavior, and manages to balance the trade-off considering performance, reliability and energy consumption together.

TABLE II
MTJ Model Parameters

Symbol	Parameter Name	Value
L	MTJ diameter	40 nm
W	MTJ thickness	100 nm
K_u	magnetic anisotropy	$5 \times 10^5 A/m$
α	magnetic damping constant	0.03
M_s at $25°C$	saturation magnetization	$3.68 \times 10^3 A/m$
t_{ox}	oxide barrier thickness	0.85 nm
t_F	free layer thickness	33.55 nm
γ	gyromagnetic ratio	$1.76 \times 10^7 rad/(s \cdot T)$

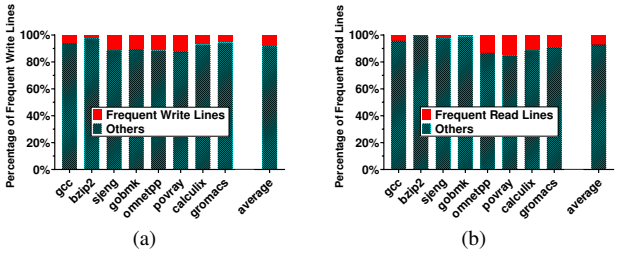

Fig. 3. (a). The percentages of frequent write lines for some SPEC2006 benchmarks. (b). The percentages of frequent read lines for some SPEC2006 benchmarks.

By investigating the access pattern of L1 cache on SPEC2006 benchmarks, we found that the number of read and write access times among different cache lines are extremely uneven (refer to the detailed experiment setup in Section V).

When running one benchmark, we firstly sorted all cache lines in descending order with the number of read accesses and write accesses, respectively. Then, we chose cache lines in order accounting for 80% of the overall read accesses and write accesses respectively. The percentage of these frequent write cache lines and frequent read cache lines are shown in Fig. 3. On average, a small fraction of cache lines in L1 cache accounts for 80% of total write accesses. Read accesses of cache lines follow the similar distributions. Therefore, we can focus on these *read-intensive lines* and *write-intensive lines* for performance improvement.

Intuitively, We can use high write voltage for all read-intensive and write-intensive lines to get better performance. It's worth noting that there are some cache lines that are both read-intensive and write-intensive lines. However, for write-intensive lines, high write voltage will seriously accelerate the TDDB effect, and reduce L1 cache lifetime dramatically. To solve the problem, we can exclude all write-intensive lines from cache lines with the high write voltage and only apply the high write voltage to the rest read-intensive lines to obtain an optimal trade-off between performance and lifetime considering the TDDB effect.

B. Implementation of DOVA

Based on the previous description, this paper proposes a Dynamic Overwriting Voltage Adjustment (DOVA) technique,

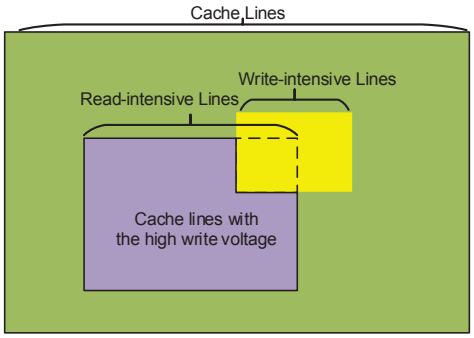

Fig. 4. The Venn diagram of the generation of critical cache lines with the high write voltage.

which aims to improve the performance of STT-RAM based L1 cache while maintain its reliability without increasing energy consumption significantly. The detailed workflow of DOVA is illustrated in Fig. 5 and presented as follows:

1) Record the number of read and write accesses of every cache line in L1 cache during the program profiling stage;
2) Sort all cache lines in descending order by their read access times. Then choose the cache lines, whose read access times fall into the top $m\%$ of the whole cache lines, and these cache lines are taken as read-intensive lines.
3) Sort all cache lines in descending order by their number of write access times. Then choose the cache lines, whose write access times fall into the top $n\%$ of the whole cache lines, as write-intensive lines. Note that m and n are variables depending on different applications;
4) Exclude write-intensive lines from read-intensive lines, and view the rest read-intensive lines as critical lines requiring the high write voltage, as shown in Fig. 4. Generate the Table of Critical Lines (TCL) to store these critical lines;
5) During the runtime, before every write operation to L1 cache, the TCL is firstly checked to identify if the L1 cache line is in the TCL. If the L1 cache line is in the TCL, the high write voltage is used. Otherwise, the low write voltage is applied.

In step 2 and 3, it is worth noting that there is a trade-off about determining the values of m and n. Different percentages lead to different optimization results. Based on experiment results of SPEC2006, we choose 30% for $m\%$ and 10% for $n\%$, which can cover more than 90% read access times and 80% write access times as shown in Table III. With these settings, DOVA can improve performance and maintain high reliability over different SPEC2006 benchmarks.

V. Experiment Results

A. Experiment Setup

Firstly, we present the experimental methodology for evaluating DOVA technique. We used NVSim [10] to get parameters like cache access latency, access energy, and leakage power at 22nm technology node. Afterwards, our proposed strategy was implemented in a system-level simulator - gem5 [11], and the detailed gem5 simulation configuration is shown in Table IV. Additionally, gem5 is modified to accommodate the asymmetry of read and write latencies. We also implemented the proposed Dynamic Overwriting Voltage Adjustment (DOVA) with non-blocking gem5 configuration. 12 benchmarks from SPEC2006 [12] were used for performance evaluations. Additionally, the energy consumption was obtained by L1 cache access statistics, and the expected lifetime was derived based on cache line access statistics produced by gem5 and the reliability model presented in Section III. TCL was obtained by 1 million instruction profiling for every benchmark. In the simulations, each benchmark was executed for 1 billion instructions after 100 million warming-up instructions.

TABLE III
The Coverage of Fractions of Total Read/Write Access Times for SPEC2006 Benchmarks When $m\% = 30\%$ and $n\% = 10\%$

Benchmark	Fraction of read accesses	Fraction of write accesses
bzip2	94.44%	85.89%
gcc	98.56%	63.55%
gromacs	98.10%	94.78%
calculix	92.87%	88.79%
average	96.00%	83.25%

TABLE IV
gem5 Simulation Configurations

Component	Configuration
CPU	quad-core, 2GHz, X86
L1 Cache STT-RAM	private, split I/D caches, 32KB 64Bytes block size, write-back policy 4-way set-associativity, LRU write latency 10 cycles (low write voltage) 14 cycles (high write voltage) read latency 3 cycles
L2 Cache STT-RAM	shared, 2MB, 64Bytes block size 8-way set-associativity read latency 7 cycles, write latency 15 cycles LRU, write-back policy
Main memory	8GB, DDR3
Protocol	MI_example

Considering the dynamic write voltage selection, we chose 1.18V as the low write voltage, which is identical to that in the most recent work [13]. The high write voltage was set to 1.81V after considering the trade-off between performance, energy consumption and expected lifetime. We used 5-year cache failure probability (the failure probability of the cache which works continuously for 5 years) to analyze the lifetime.

B. Experiment Results and Analyses

We evaluated the proposed strategy against two baseline settings: only using the low write voltage 1.18V and only using the high write voltage 1.81V.

*1) **Performance Evaluations and Overhead Analysis**:* Fig. 6 shows the normalized performance results of DOVA and 1.81V cases compared to the 1.18V baseline. The figure indicates that the performance improvement of DOVA can be up to 18%, and approximately 9% on average compared to the 1.18V

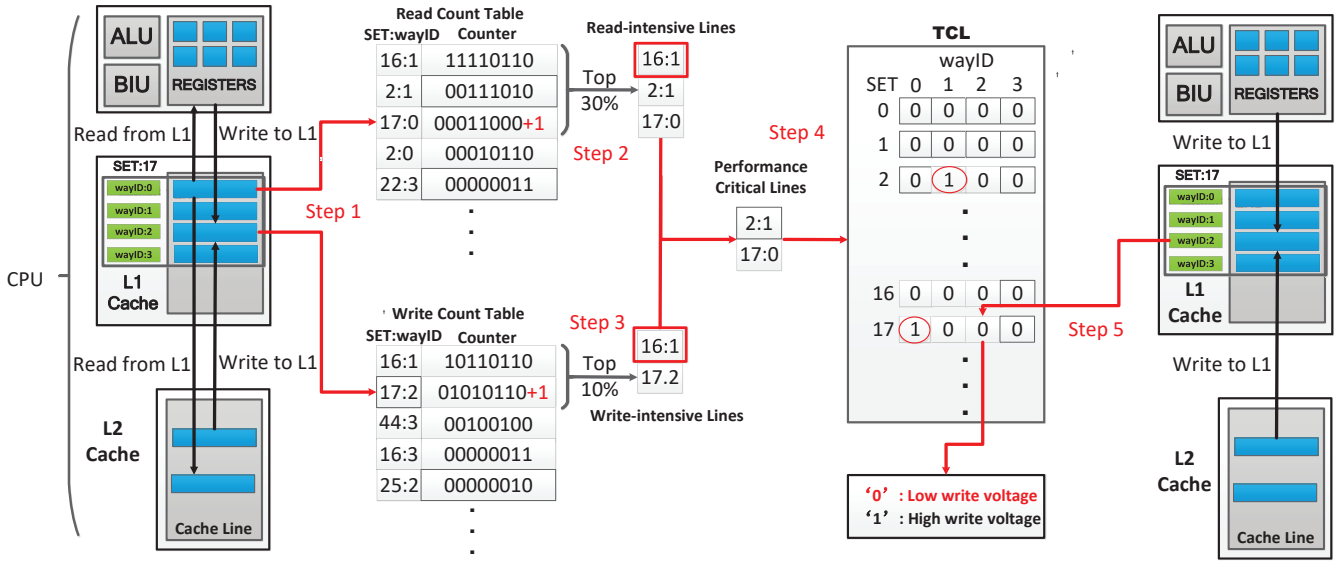

Fig. 5. The work flow of DOVA technique. Step (1) record read/write access times (profiling). Step (2) identify read-intensive lines. Step (3) identify write-intensive lines. Step (4) identify critical lines with the high write voltage and generate a table of critical lines (TCL) to store them. Step (5) check TCL and apply write voltage accordingly.

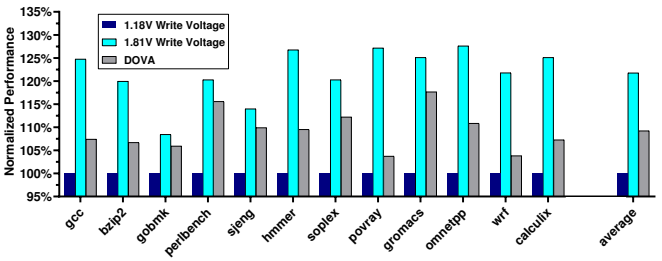

Fig. 6. Normalized performance improvements when running SPEC2K6 benchmarks on a quad-core processor.

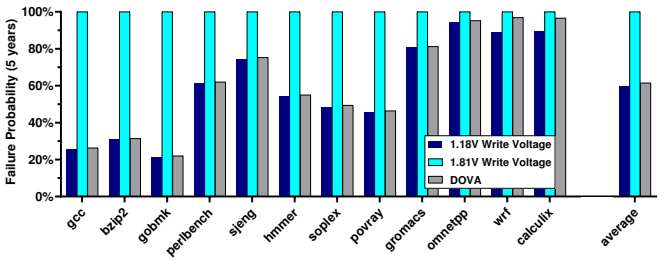

Fig. 7. 5-year cache failure probability comparisons.

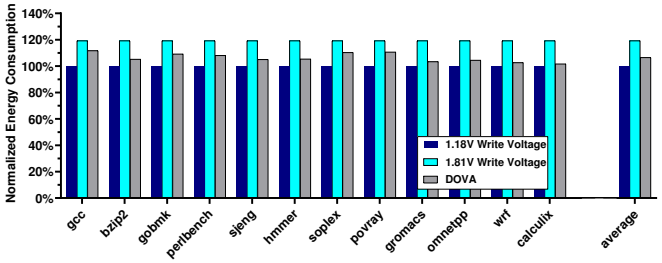

Fig. 8. Normalized energy consumptions.

baseline. Although the improvement can be higher if 1.81V is applied, the expected lifetime and energy consumption are seriously deteriorated, which will be evaluated in the following discussions.

Since the TCL can be directly accessed by c_i(the set of the line) and r_j(the index of the cache line in the set), the time latency of looking up TCL is negligible compared to cache access latency due to its very small size.

*2) **Expected Cache Lifetime**:* Fig. 7 presents 5-year cache failure probability of DOVA compared with other two cases. From the comparison results, we can observe that DOVA leads to only 1.74% average degradation of failure probability compared to the 1.18V case, and is 38.51% lower than that of 1.81V case. This is because DOVA protects the majority of write-intensive cache lines with low write voltage and only applies high write voltage to cache lines that are read-intensive but not write-intensive.

*3) **Write Energy and Hardware Overhead Analysis**:* Fig. 8 shows the normalized write energy of DOVA and the 1.81V case compared to the 1.18V case. DOVA only incurs 6.51% energy overhead on average (1.71% minimum), while 1.81V incurs 19.27% energy overhead on average because it uses high voltage for all write accesses, which is unnecessary since most write accesses only marginally impact performance as mentioned in Section IV.

In order to implement DOVA, a cache access counter and a TCL are required. Each cache line has a 4-byte counter, adding up to 2 KB for a 32KB L1 cache with 4-way set-associativity. In terms of TCL, assume the number of sets in L1 is x and the associativity is y. TCL is a $x \times y$ matrix with each entry storing 1 bit to show whether the corresponding cache line should be written with high voltage or not. Therefore, TCL storage overhead is 64 bytes. The total storage overhead of DOVA is 2112 bytes which is negligible compared with the typical cache size.

VI. Related Work

Most of the state-of-art STT-RAM based L1 caches are implemented with a buffer to cope with high write latency

issue, such as SRAM buffer in [14] and Very Wide Buffer in [3]. Although adding a buffer is helpful, it induces hybrid CMOS/MTJ cache hierarchy which contradicts with the pure non-volatile cache hierarchy investigated in this paper. Conventional non-volatile cache lifetime extension techniques can be mainly divided into two categories. The first category mainly focuses on balancing cache intra-set and inter-set access variations, aiming at improving the lifetime of on-chip non-volatile caches [8] [15] [16]. Further techniques based on the concept have also been proposed in [17] and [18]. The second category is about error correction techniques. On one hand, typical ECC is implemented in some STT-RAM cache architecture like [19]. On the other hand, novel ECC strategy, for example, Interleaved Single Error Correction-Double Error Detection (SEC-DED) has been proposed in [20] to improve the lifetime of L2 and L3 caches. Different from existing research work, the proposed DOVA supports pure non-volatile STT-RAM cache hierarchy and is both reliable and performance friendly.

VII. Conclusion

STT-RAM is a competitive candidate to build caches to replace SRAM because of its advantages like high density, long durability, non-volatile storage, etc. However, it suffers from the problem of high write latency, and it is imperative to accelerate write speed to enable STT-RAM based L1 cache. Increasing write voltage is an effective method. However, this high write voltage may reduce STT-RAM lifetime due to the TDDB effect. Thus it is critical to make an optimal trade-off between write performance and lifetime of STT-RAM L1 cache. This paper proposes a Dynamic Overwriting Voltage Adjustment technique to write different types of cache lines with different write voltages. Experimental results showed that DOVA can improve cache speed performance by 18% in maximum, and 9% on average. In the meantime, the average degradation of cache failure probability caused by DOVA is only 1.74%, which is 38.51% lower than the 1.81V high write voltage case. Moreover, compared with the 1.18V case, the write energy consumption increase of DOVA is 6.51% on average (1.71% in minimum), which is much lower than the 19.27% write energy consumption increase of the 1.81V case.

Acknowledgment

This work is partially supported by Supported by Beijing Natural Science Foundation under grant No. 4192035, Science, Technology and Innovation Commission of Shenzhen Municipality under grant No. JCYJ20180307123657364.

References

[1] W. Zhao, Z. Wang, S. Peng, L. Wang, L. Chang, and Y. Zhang. Recent progresses in spin transfer torque-based magnetoresistive random access memory (STT-RAM). *SCIENTIA SINICA Physica, Mechanica & Astronomica*, 46(10):107306, 2016.

[2] S. P. Park, S. Gupta, N. Mojumder, A. Raghunathan, and K. Roy. Future cache design using STT MRAMs for improved energy efficiency: Devices, circuits and architecture. In *Proc. of Design Automation Conference (DAC)*, pages 492–497, June 2012.

[3] M. P. Komalan, C. Tenllado, J. I. G. Perez, F. T. Fernandez, and F. Catthoor. System level exploration of a STT-RAM based level 1 data-cache. In *Proc. of Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 1311–1316, March 2015.

[4] K. Kim, C. Choi, Y. Oh, H. Sukegawa, S. Mitani, and Y. Song. Time-dependent dielectric breakdown of MgO magnetic tunnel junctions and novel test method. *Japanese Journal of Applied Physics*,56(4S), 04CN02, 2017.

[5] S. Amara-Dababi, H. Bea, R. Sousa, K. Mackay, and B. Dieny. Modelling of time-dependent dielectric barrier breakdown mechanisms in MgO-based magnetic tunnel junctions. *Journal of Physics D: Applied Physics*, 45(29):295002, Jul 2012.

[6] Y. Wang, H. Cai, L. A. d. B. Naviner, Y. Zhang, X. Zhao, E. Deng, J. Klein, and W. Zhao. Compact model of dielectric breakdown in spin-transfer torque magnetic tunnel junction. *IEEE Transactions on Electron Devices*, 63(4):1762–1767, April 2016.

[7] S. Van Beek, K. Martens, P. Roussel, G. Donadio, J. Swerts, S. Mertens, A. Thean, G. Kar, A. Furnemont, and G. Groeseneken. Voltage acceleration and pulse dependence of barrier breakdown in mgo based magnetic tunnel junctions. In *Proc. of IEEE International Reliability Physics Symposium (IRPS)*, pages MY–4–1–MY–4–4, April 2016.

[8] J. Wang, X. Dong, Y. Xie, and N. P. Jouppi. i2WAP: Improving non-volatile cache lifetime by reducing inter- and intra-set write variations. In *Proc. of IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*, pages 234–245, Feb 2013.

[9] Fong, X., Choday, S. H., Georgios, P., Augustine, C., Roy, K. (2014). Purdue Nanoelectronics Research Laboratory Magnetic Tunnel Junction Model. nanoHUB. doi:10.4231/D33R0PV04

[10] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi. Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 31(7):994–1007, July 2012.

[11] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood. The gem5 simulator. *SIGARCH Comput. Archit. News*, 39(2):1–7, August 2011.

[12] J. L. Henning. SPEC CPU2006 benchmark descriptions. *ACM SIGARCH Computer Architecture News*, 34(4):1–17, 2006.

[13] L. Wei, J. G. Alzate, U. Arslan, J. Brockman, N. Das, K. Fischer, T. Ghani, O. Golonzka, P. Hentges, R. Jahan, P. Jain, B. Lin, M. Meterelliyoz, J. O' Donnell, C. Puls, P. Quintero, T. Sahu, M. Sekhar, A. Vangapaty, C. Wiegand, and F. Hamzaoglu. A 7mb STT-RAM in 22FFL FINFET technology with 4ns read sensing time at 0.9v using write-verify-write scheme and offset-cancellation sensing technique. In *Proc. of IEEE International Solid- State Circuits Conference (ISSCC)*, pages 214–216, Feb 2019.

[14] H. Sun, C. Liu, W. Xu, J. Zhao, N. Zheng, and T. Zhang. Using magnetic RAM to build low-power and soft error-resilient L1 cache. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 20(1):19–28, Jan 2012.

[15] S. Mittal. Using cache-coloring to mitigate inter-set write variation in non-volatile caches. *arXiv preprint arXiv:1310.8494*, 2013.

[16] S. Mittal and J. S. Vetter. EqualWrites: Reducing intra-set write variations for enhancing lifetime of non-volatile caches. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(1), 103-114, 2015.

[17] S. Agarwal and H. K. Kapoor. Targeting inter set write variation to improve the lifetime of non-volatile cache using fellow sets. In *Proc. of IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*, pages 1-6, Oct 2017.

[18] S. Agarwal and H. K. Kapoor. Enhancing the lifetime of non-volatile caches by exploiting module-wise write restriction. In *Proc. of Great Lakes Symposium on VLSI(GLSVLSI)*, pages 213-218, New York, USA, 2019.

[19] H. Noguchi, K. Ikegami, N. Shimomura, T. Tetsufumi, J. Ito, and S. Fujita. Highly reliable and low-power nonvolatile cache memory with advanced perpendicular STT-RAM for high-performance CPU. In *Proc. of Symposium on VLSI Circuits (VLSI)*, pages 1–2, June 2014.

[20] H. Farbeh, H. Kim, S. G. Miremadi, and S. Kim. Floating-ecc: Dynamic repositioning of error correcting code bits for extending the lifetime of STT-RAM caches. *IEEE Transactions on Computers*, 65(12):3661–3675, Dec 2016.