Variable-Energy Write STT-RAM Architecture with Bit-Wise Write-Completion Monitoring

Tianhao Zheng, Jaeyoung Park, Michael Orshansky, Mattan Erez
Dept. of Electrical and Computer Engineering
The University of Texas at Austin
Austin, Texas
Email: {thzheng, jypark22, orshansky, mattan.erez}@utexas.edu

***Abstract*—**
In this paper we demonstrate an energy-reduction strategy that relies on the stochastic long-tail nature of the STT-RAM write operation. To move away from the traditional worst-case approach, the per-cell write process is continuously monitored and is terminated as soon as each cell's state matches the written state. Since the average write duration is far shorter than the worst-case duration, the average write energy is significantly reduced by the proposed architecture. We developed a light-weight circuit for fast state change detection and bit-line shutdown and evaluated it using a compact STT-RAM model targeting an implementation in a 16nm technology node. Our analysis indicates that at the required write-error rate the proposed architecture reduces write energy by $87.3\% - 99.5\%$ depending on the write direction, and on average achieves 96.5% write energy saving in 16 SPEC CPU 2006 applications compared to conventional design. Compared to the best previously known architecture that exploits stochasticity (verify-on-write), we reduce write energy by approximately $6.5\times$.

I. Introduction

Spin-torque transfer memory (STT-RAM) is as a candidate for a universal memory technology that may be able to provide integration density close to DRAM, the non-volatility of Flash memory, fast read speed close to that of SRAM, and practically zero standby power. At the 16nm node, STT cell designs with very competitive characteristics – a read time of $1 - 5\text{ns}$ and an average write time of $5 - 10\text{ns}$ – are feasible. These characteristics are attractive both for replacing SRAM in large on-chip last-level caches and for replacing some or all of the off-chip DRAM. In addition to reducing static power and increasing density compared to SRAM, the non-volatility of STT-RAM opens new opportunities for improving processor power management. As a DRAM replacement, STT-RAM eliminates the need for refresh; refresh operations increasingly interfere with demand traffic and consume significant power in large installations.

Despite its potential, several issues stand in the way of wide-scale STT-RAM adoption. One such critical issue is the high energy required for reliable write operations. Writing an STT-RAM cell is done with a relatively high current, requiring significantly more power than reading the cell. More importantly, the STT write process is inherently stochastic and the actual time to complete a write varies dramatically, with the distribution having a very long tail. This stochasticity of switching time is temporal, leading to variation in transition time even for a single cell. As a result, conservatively guaranteeing a reliable write requires maintaining the write current for a duration much longer than that required for an average write to complete.

We propose a novel approach that exploits write stochasticity to significantly reduce the write energy in STT-RAM. With our *variable-energy writes* (VEWs), the write current of each individual cell is terminated once that cell's state matches its desired write value. This is in contrast to the conventional approach, which fixes the write duration of all writes to match the expected worst-case delay for a given level of reliability.

In order to replace SRAM and DRAM within the memory system, STT-RAM writes must be highly reliable. Even when considering the fact that write operations are to multiple bits at a time and that error protection techniques are available, the single bit error rate must be very low; the typical single bit error rate needed is 1.5×10^{-7} [1]. The difference between a pulse needed to achieve this error rate and the mean pulse duration is almost $20\times$. Even when targeting a very high error rate of 0.01, the difference between the pulse duration that guarantees this error rate and the mean pulse is still $4\times$. This means that substantial energy savings could be achieved for a large fraction of switching events if we have the ability to terminate the current pulse once the specific switching occurred. These energy savings can be achieved by moving away from the traditional worst-case approach towards a technique in which we can detect the write completion of each bit and turn off the switch current. To realize variable-energy writes, we developed a light-weight circuit that continuously monitors the state of each STT-RAM cell and disables its write current when it senses the desired state has been achieved.

The mechanism we exploit is orthogonal to most previous proposals for curbing write energy, which focused on tuning STT-RAM parameters for trading off lower write current with increased volatility [2], [3] or terminating write operations early for cells whose values do not change [4]. A related mechanism for exploiting STT-RAM write stochasticity, verify-on-write (VOW), was recently proposed by Bi et al. [5]. VOW is based on a similar insight to ours but has two important limitations that our work overcomes. First, VOW only functions correctly if the STT-RAM cell is designed with '0'/'1' write asymmetry where switching the cell in one direction requires an order of magnitude less time than

978-1-4799-1235-3/13/$31.00 © 2013 IEEE This is the authors̀ version of the work. The denitive version was published in the Proceedings of ISLPED̀13

the other direction (see Section II). Second, we introduce a novel circuit solution that allows bit-wise monitoring and write current termination. That allows us to control each bit independently. In VOW, bit-wise current termination is not possible forcing the write termination signal to wait for the longest-duration write in an entire word to complete. We show that by avoiding these limitations, VEWs are a significantly more attractive solution.

We evaluate the variable-energy write design using a compact STT-RAM cell model targeting an implementation in a 16nm technology node. Our detailed evaluation indicates that variable-energy writes reduce the overall average STT-RAM write energy by $87.3-99.5\%$ compared to conservative fixed-duration writes, depending on the write direction. We also analyzed the impact when considering actual main-memory traffic of industry-standard application benchmarks. We find that VEWs reduce overall write energy by 96.5% on average across these benchmarks compared to the conventional design. Compared to VOW we reduce write energy by approximately $6.5\times$.

II. Modeling Stochastic STT-RAM Writes

In this section, we describe the STT-RAM cell write process, explain the opportunity for reducing energy, and discuss the stochastic model we rely on for evaluation.

A. STT-RAM Write Process

The core component in an STT-RAM cell is its magnetic tunnel junction (MTJ) [6]. The MTJ consists of two layers of magnetic material separated by a dielectric layer. The two magnetic layers have their own spin directions, with one layer pinned to a fixed polarization and the second layer being free. The spin of the free layer can be switched from one orientation to its opposite by applying a current pulse through the MTJ. The MTJ as a unit can be in one of two states, *anti-parallel* (*AP*) and *parallel* (*P*). In the *AP* state the free and pinned layers have opposite spin directions and in the *P* state both have the same spin. Each of these two states exhibits a distinct resistance corresponding to storing a binary '0' or '1' (e.g., *P* and *AP*).

An access transistor is connected with the MTJ to control its operation. Write '0' and write '1' operations proceed by turning on the access transistor and injecting a relatively high write current in one of two directions (from source line to bit line or vice verse). A sense amplifier is connected to each bit line to detect the state in MTJ in a read operation that requires a much lower current and a shorter pulse duration compared to the write operation.

The change of MTJ state occurs when the current passing through the junction exceeds a certain minimum magnitude and is maintained for sufficient time. The process of MTJ write is a process of alignment of the magnetic orientation of the regions of the ferromagnetic layer. The total switching time of STT-RAM consists of the incubation time and transit time [7]. The incubation time is defined as the time needed for the electrons to climb up the potential barrier in an MTJ, while the transit time denotes the time for the electrons to descend the potential barrier to the other state.

There are two distinct physical mechanisms that govern MTJ switching, which depend on the magnitude of injected current: thermally activated switching and fast precessional switching. The thermally activated switching regime holds for currents at or below a certain *critical current* (I_{C0}) defined at zero Kelvin. The thermally activated switch process is relatively slow, with mean switch times of several nanoseconds to tens of nanoseconds, and is also highly stochastic. The average current required for switching (I_C^{therm}) depends on the write pulse duration (T_{wr}). The following deterministic model is often used to describe this relationship, despite the inherent stochasticity of the write process [8]:

$$I_C^{therm}\left(T_{wr}\right) = I_{C0}\left\{1 - \frac{1}{\Delta}\ln\left(\frac{T_{wr}}{\delta_0}\right)\right\} \tag{1}$$

where Δ is the thermal stability factor [8], [9].

The second mechanism, fast precession switching, is very rapid, typically occurring within 1ns, and shows less stochasticity. However, activating this switching process requires a current that is much larger than I_{C0}. The average current required in fast precessional switching also depends on the write pulse duration and can be described as:

$$I_C^{prec}\left(T_{wr}\right) = I_{C0} + \frac{C\ln\left(\pi/2\theta\right)}{T_{wr}} \tag{2}$$

where C and θ represent the relaxation time and initial angle between the free layer and reference layer, respectively [8].

There is an important difference in current values that are supplied by the access transistor to the MTJ in the course of a normal operation of the cell. Depending on the value being written to the cell the bitline voltage is set high and the source line voltage is set low, or vice versa, see Fig. 2. The effective V_{gs} of the access transistor is different in the two cases, resulting in significantly different current values of $2\times$ or more, unless word-line boosting is used in one of the two cases. The difference in the delivered current leads to a significant asymmetry of MTJ switching times. As we demonstrate in this paper, it is beneficial for the overall energy minimization to use word-line boosting to make the distributions symmetric.

B. Stochastic Switching Model

The time needed for the MTJ to switch is stochastic and the switch time distribution depends on the magnitude of the current. An essential limitation of the above deterministic models is that they fail to take into account the stochasticity in the switching process and describe the mean switching behavior rather than the entire distribution. The variable write energy technique is based on the probabilistic model of the switching behavior. Empirical observations and numerical simulations suggests that stochasticity of switching time is maximum in the thermal switching regime [10]. Therefore, it is crucial to be able to model stochastic behavior in this regime. The following probability model for the thermally activated switch

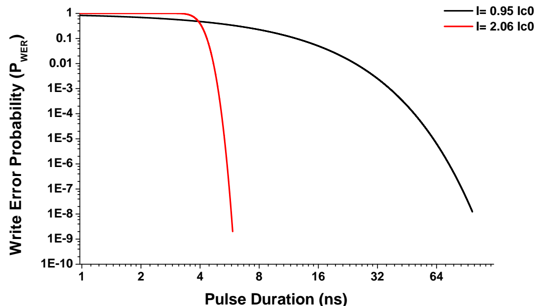


Fig. 1: Probabilistic model of write error probability as a function of write pulse duration; shown for several currents.

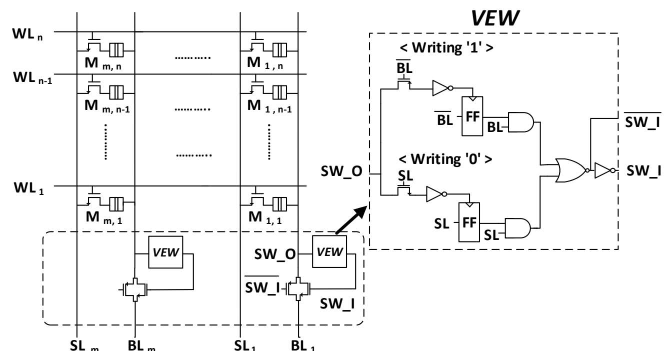


Fig. 2: Block diagram of a memory array with the proposed VEW circuit.

duration is derived by Diao et al. [9] using the Neel-Brown relaxation formula. The model describes the switch probability $P_{SW}(t, I)$, which is the probability of switching occurring for a pulse duration t at current I:

$$P_{SW} = 1 - \exp\left\{-\frac{t}{\tau_0}\exp\left[-\Delta\left(1 - \frac{I}{I_{C0}}\right)\right]\right\} \quad (3)$$

where τ_0 is the inverse of the thermal attempt frequency that has a typical value of 1ns.

In the fast precession switching regime, the stochasticity of switching time is lower. The ratio of the switch time standard deviation to its mean is in the range from 0.2 to 1 [10]. The same ratio has the value of 0.08 for moderately wide pulses in the fast precessional switching regime [9]. Unfortunately, in this regime, the exact closed-form model for the switching time probability distribution is not available. However, empirical measurements suggest that the form of the distribution is asymmetric Gaussian [10]. We therefore adopt this model for our experiments in the case of asymmetric writes where the current through the MTJ is $2.05I_{C0}$ in our design (Fig. 1).

Importantly, in the thermal switching regime (e.g., $I = 0.95I_{C0}$), we find that the model predicts that the pulse duration required to reach a high write success probability is much larger than the pulse duration corresponding to a switching probability of 50%. Fig. 1 shows the width of the distribution by plotting $1 - P_{SW}$, which can be thought of as the write error probability. We observe a strong case of a long-tail distribution as the difference between the mean pulse duration and the pulse corresponding to the very low error probability is very large.

III. Variable-Energy Writes

We introduce the variable energy write architecture as an effective way to exploit the wide distribution of write time. Write energy is reduced by utilizing a mechanism that (1) monitors the instantaneous state (resistance) of the MTJ, and (2) deactivates the write current once the correct value being written has registered. In addition to saving energy when the write switches the MTJ state, our design inherently shuts down the write current when the value being written equals the value already stored. Below we describe the monitoring circuit that compares the MTJ resistance to a reference value and can thus determine the stored value and sense the change of resistance on a switch.

A. Implementation Details

Fig. 2 shows a schematic view of the proposed circuit comprised of the monitoring sub-circuit and the shutdown sub-circuit. The monitoring sub-circuit, whose schematic is further expanded in an inset, tracks the bitline voltage at node SW_O. Depending on the direction of the write, the voltages corresponding to the *AP* and *P* states are significantly different. Therefore, to detect both transitions, the monitoring sub-circuit contains two comparators whose thresholds are set to 630mV for $P \rightarrow AP$ and to 390mV for $AP \rightarrow P$. For both transitions, the voltage difference between the *AP* and *P* states for the assumed STT-RAM cell parameters (Tab. I) is around 80mV. Because of the magnitude of the difference, the resolution of the comparator does not have to be very high. We found that sufficient resolution can be achieved by using a single-stage CMOS inverter as the comparator. The switching threshold voltage of the inverter can be set to the midpoint between the *P* and *AP* write voltages by properly selecting the PMOS/NMOS sizing ratio.

Consider a $P \rightarrow AP$ transition. At the start of the write operation the voltage is raised on the wordline (WL) and the bitline (BL), and lowered on the source line (SL). The access transistor turns on and the current flows through the MTJ and the access transistor. If the current supplied by the NMOS is sufficiently large, the MTJ undergoes a state change. A state change modifies the MTJ resistance, which leads to a voltage change at the output node of the access transistor. The voltage change on node SW_O is detected by the comparator (the tuned inverter), which then turns the shutdown transmission gate off, terminating the pulse. The AND gate, which is used in the shutdown circuit, ensures that this monitoring circuit is only active when writing a logical '1'. The flip-flop in the VEW circuit works as a delay element to prevent an unwanted feedback loop between the monitoring and reset circuits.

The area overhead of the above design is 2 inverters, 2 AND gates, 1 OR gate per column, and 2 flip-flops. We estimate the area to be $1860F^2$. The relative area overhead will be

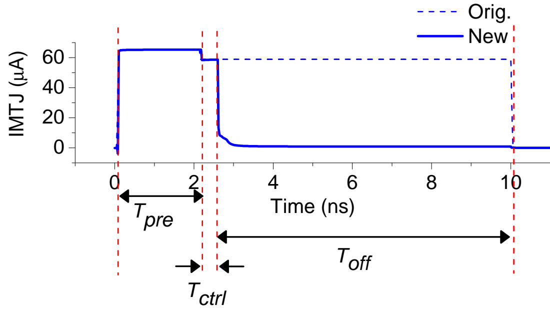


Fig. 3: Simulated current of a $P \rightarrow AP$ transition for the conventional (original) and VEW techniques.

reduced with column multiplexing, which is often used to deal with pitch-matching of sense amplifier. For comparison, the area overhead reported in the previously proposed early-write termination (EWT) [4] is $6637F^2$.

Note that the proposed write completion circuit controls each bit independently which maximizes write efficiency. This is in contrast to the VOW architecture of Bi et al. [5], which terminates writes at the granularity of an entire word (Fig. 5). Therefore, VEWs save significantly more energy, as we discuss in Section IV.

We also note that the presence of significant process variations may make our current VEW design sub-optimal. We plan to utilize existing design strategies for dealing with variability, such as post-silicon tuning and self-calibration. In addition, memory array islands can help deal with correlated as well as uncorrelated variations between memory regions [11], and error checking and correcting (ECC) can be used, in part, to correct errors due to uncorrelated variations [12], [13]. Investigating a variation-tolerant VEW architecture will be a focus of our future work.

B. Circuit Validation

We designed the proposed circuit using the 16nm Predictive Technology Model (PTM) MOSFET model and a compact MTJ model [14], [15], [16]. The MTJ parameters are derived from the 17nm MTJ manufactured by Samsung [15], as detailed in Tab. I. Simulation results of the proposed circuit are shown in Fig. 3, for a $P \rightarrow AP$ transition. While the conventional 1-bit cell without the write completion circuit consumes substantial power until the end of the clock period, the 1-bit cell with the VEW circuit minimizes power after the switch occurs at 3ns.

Fig. 4 shows simulation results for writing a '0', when a '0' is already stored in a cell ($P \rightarrow P$). The current through the MTJ (IMTJ) for the baseline circuit (dashed line) is kept high for the entire pulse duration (10ns in this simulation), while the MTJ current controlled by the proposed write-completion circuit (WCC) drops to zero within 1ns, which is the response time of the monitoring and control circuits. We validate that our circuit design behaves correctly in the array configuration as well. We use wire resistance and capacitance as specified by ITRS for a 16nm process (resistivity: 22 μΩ-cm; capacitance

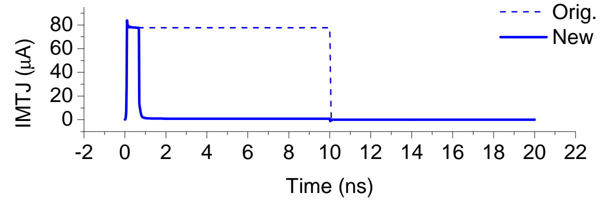


Fig. 4: Simulation results of a 1-bit cell with the proposed write completion circuit writing '0' while already in the '0' state ($P \rightarrow P$).

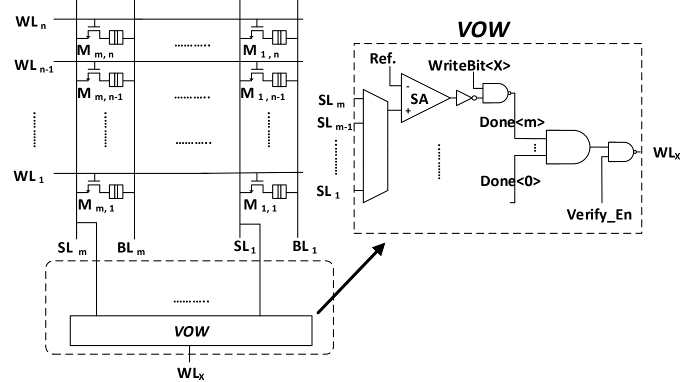


Fig. 5: Block diagram of a memory array with VOW [5].

per unit length: 1.6pF/cm) [17]. The wire delay, simulated with Cadence Spectre, is less than 10ps.

IV. Variable Energy Write Technique: Evaluation

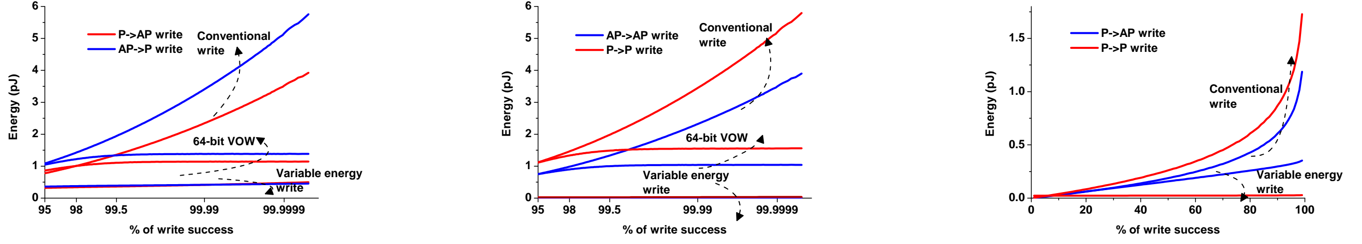
In this section, we evaluate the energy saving achieved by applying our variable-energy write technique.

A. Methodology

To evaluate the energy consumption of VEWs, we consider each write operation as a three-stage process (depicted in Fig. 3). The first stage, T_{pre}, is the time after the write starts until the MTJ switch occurs (previous value held in the cell). The second stage, T_{ctrl}, is the time between the switch and the write current is terminated; this is the reaction time of our monitoring, control, and delay circuits. The third stage, T_{off}, is the time after shutdown until the end of the pre-determined pulse duration, which equals the time required to meet a given maximum error rate target. Without VEWs, there

TABLE I: Key parameters of 17nm MTJ [14], [15], [18].

Term	Definition	Value	Unit
e	Electron charge	1.6E-19	C
$\hbar$	Reduced Planck constant	1.05E-34	Js
α	Magnetic damping constant	0.027	
η	Thickness of the oxide barrier	1.3	nm
t_F	Thickness of the free layer	0.9	nm
W	Width of MTJ	40.0	nm
L	Length of MTJ	17.0	nm
H_d	Out-of-plane magnetic anisotropy	1.3	T
P	Percentage of tunnel current	0.56	
I_{c0}	Critical current at zero Kelvin	68.9	μA
Δ	Thermal stability factor	34	



(a) Per-bit energy of $P \rightarrow AP$ and $AP \rightarrow P$ writes.

(b) Per-bit energy of $P \rightarrow P$ and $AP \rightarrow AP$ value-maintaining writes.

(c) Per-bit energy of $P \rightarrow AP$ and $P \rightarrow P$ writes across large range of write success probabilities.

Fig. 6: Summary of energy per-bit for conventional and VEWs across a range of write-success probabilities.

are two stages, T_{pre} as before and T_{post}, which is the time after the switch with the current still on.

We then calculate the write energy using the power consumed in each stage, as shown in Eq. 4 and Eq. 5 for baseline and VEWs, respectively.

$$E_{baseline} = V_{DD}(T_{pre}I_{pre} + T_{post}I_{post}) \tag{4}$$

$$\begin{aligned} E_{VEW} = V_{DD}(T_{pre}I_{pre} + T_{ctrl}I_{ctrl} + T_{off}I_{off}) \\ + P_{wcc}(T_{pre} + T_{ctrl} + T_{off}) \end{aligned} \tag{5}$$

P_{wcc} in Eq. 5 is the power of the write completion circuit (static and dynamic), which is the overhead of our technique. We obtain the circuit parameters, including its power consumption and reaction time (T_{ctrl}) using SPICE simulation ($V_{DD} = 1.05$V for the 16nm process with boosting). We compute the overall pulse duration to match a specific desired error probability as determined by the probabilistic model shown in Eq. 3 (Section II). We calculate the expected durations T_{pre}, T_{off}, and T_{post} by applying the model of Eq. 3 again.

We also compare VEWs with the related verify-on-write (VOW) [5] technique and with early-write termination (EWT) [4]. To directly compare all three techniques within the same context, we use the same MTJ parameters as in Tab. I. We also use the same basic components as our write-completion circuit, but modify their application to mimic VOW and EWT. To mimic VOW, we remove the boosting write circuit (VOW was designed assuming asymmetric writes and no boosting) and also shut down the pulse to an entire 64-bit word at a time, rather than each individual bit. We use Monte Carlo simulation to obtain the expected word-completion time, which is the expected longest duration write in each word (we measure the actual number of $P \rightarrow AP$ writes in each word). We use the write-duration distributions for $0.95I_{C0}$ and $2.05I_{C0}$ as described in Section II (without boosting the circuit required $V_{DD} = 1.73$V). We also simulate an augmented version of VOW that can work with boosting (symmetric writes) by utilizing our monitoring circuit and using the Monte Carlo methodology assuming 50% of the bits in a word switch state. To mimic EWT, we only utilize our monitoring and shutdown circuits for those bits that do not change their state in each word-granularity write.

B. Results

In addition to the validation results presented in Section III-B, we now present the expected energy savings of using VEWs. We compare the energy of conventional, VOW and VEWs across a range of write error probabilities: from an unacceptably high error rate of 5% to our target error rate of 1.5×10^{-7} [1]. These results are summarized in Fig. 6a and Fig. 6b, which show the expected energy per bit for the cases of switching writes and writes that maintain already stored values ($P \rightarrow P$ and $AP \rightarrow AP$), respectively. Note that the horizontal axis in each figure represents success probability, rather than error rate. For the value-maintaining writes, the error probability is zero, however, we use the same horizontal scale, where each success rate point corresponds to a certain pulse duration.

In all cases and across this entire error probability range, VEWs dramatically improve write energy compare with both the conventional baseline and a VOW with 64-bit sub-block(same as [5]). We assume 50% bits within one sub-block will change in each write in VOW. We do not include VEW here because VEW only save energy in value-maintaining writes. For the target error rate (1.5×10^{-7}), the savings are 87.3% for a $P \rightarrow AP$ switch and 92.0% for a $AP \rightarrow P$ switch, while VOW saves 71.0% and 76.0% respectively.The energy savings are even more significant when no switching occurs and are 99.3% and 99.5% for "writing" a $P \rightarrow P$ and $AP \rightarrow AP$, respectively (no state change). For VOW, the savings are 73.4% and 73.2%.

To have a global view of how VEWs compare with the conventional baseline write technique of fixed pulse duration, we show the expected write energy per bit across a wider range of write-success probabilities in Fig. 6c.

Finally, to put the energy savings into the context of a memory system, we apply our energy model to the main memory write traffic of 16 SPEC CPU 2006 benchmarks. We get memory trace form the 8 integer and 8 floating-point applications with the largest number of memory accesses with PIN [19], assuming an on-chip memory hierarchy with a single X86 Out-of-Order core, 32KB L1, 256KB L2 and 1MB 16-way set-associative last-level cache. Fig. 7 shows the relative energy of VEWs and the previously proposed early write

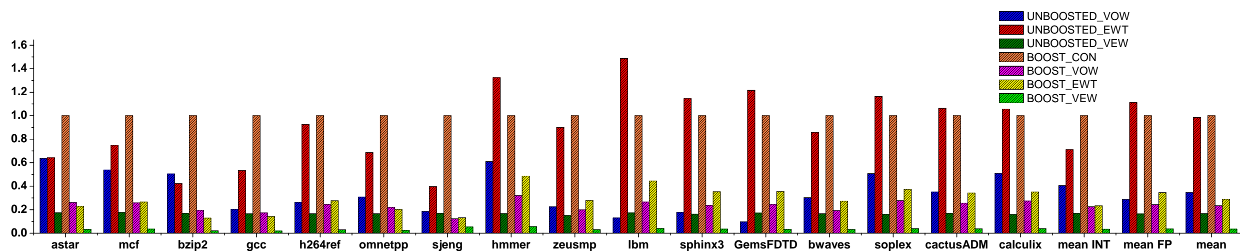


Fig. 7: Overall write energy of VEWs, VOW, and EWT with and without write boosting (normalized to conservative with boosting) for the main-memory traffic of 8 INT and 8 FP SPEC CPU 2006 benchmarks.

termination [4] and verify-on-write [5] design.

Early write termination is only able to reduce the energy of value-maintaining writes, and the word granularity shut down limited the benefit from VOW. As a result VEWs improve energy saving much more: VEWs decrease total write energy consumption by 96.5% on average, while VOW and EWT only achieve 76.6% and 71.0% reductions, respectively. This is because VEWs are very robust as they account for all 4 possible state combinations and monitoring and shutdown is applied for every bit independently. It is also important to note that VEWs work with and without boosting and that the symmetric writes enabled by boosting are the most energy-efficient technique overall; without boosting, VEWs are still the best option, but consume 4.76 times more energy than when boosting is enabled.

V. Conclusions

This paper proposes a novel variable-energy write STT-RAM architecture with a write-completion monitoring. In the proposed architecture, the write process is continuously monitored and is terminated as soon as the MTJ reached the required state. We also developed a light-weight circuit for fast state change detection and evaluated it using a compact MTJ model targeting an implementation in a 16nm technology node. The proposed technique has no significant area overhead and is easy to integrate in memory array. Our analysis indicates that at the required write-error rate the proposed architecture reduces write energy by $87.3\% - 99.5\%$ depending on the write direction. An important direction for future work is validating the performance of the scheme in the presence of process variability.

VI. Acknowledgments

We sincerely thank all anonymous reviewers for their feedback on the earlier drafts of this paper. This work was funded by Samsung Global Outreach Research Program.

References

[1] DC Worledge, G Hu, PL Trouilloud, DW Abraham, S Brown, MC Gaidis, J Nowak, EJ O'Sullivan, RP Robertazzi, JZ Sun, et al. Switching distributions and write reliability of perpendicular spin torque mram. In *IEDM*, 2010.

[2] Clinton W Smullen, Vidyabhushan Mohan, Anurag Nigam, Sudhanva Gurumurthi, and Mircea R Stan. Relaxing non-volatility for fast and energy-efficient stt-ram caches. In *HPCA*, 2011.

[3] Zhenyu Sun, Xiuyuan Bi, Hai Helen Li, Weng-Fai Wong, Zhong-Liang Ong, Xiaochun Zhu, and Wenqing Wu. Multi retention level stt-ram cache designs with a dynamic refresh scheme. In *MICRO*, 2011.

[4] Ping Zhou, Bo Zhao, Jun Yang, and Youtao Zhang. Energy reduction for stt-ram using early write termination. In *ICCAD*, 2009.

[5] Xiuyuan Bi, Zhenyu Sun, Hai Li, and Wenqing Wu. Probabilistic design methodology to improve run-time stability and performance of stt-ram caches. In *ICCAD*, 2012.

[6] M Hosomi, H Yamagishi, T Yamamoto, K Bessho, Y Higo, K Yamane, H Yamada, M Shoji, H Hachino, C Fukumoto, et al. A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-ram. In *IEDM*, 2005.

[7] Fumitaka Iga, Yasuhiro Yoshida, Shoji Ikeda, Takahiro Hanyu, Hideo Ohno, and Tetsuo Endoh. Time-resolved switching characteristic in magnetic tunnel junction with spin transfer torque write scheme. *Japanese Journal of Applied Physics*, 51(2):02BM02, 2012.

[8] Arijit Raychowdhury, Dinesh Somasekhar, Tanay Karnik, and Vivek De. Design space and scalability exploration of 1t-1stt mtj memory arrays in the presence of variability and disturbances. In *IEDM*, 2009.

[9] Zhitao Diao, Zhanjie Li, Shengyuang Wang, Yunfei Ding, Alex Panchula, Eugene Chen, Lien-Chang Wang, and Yiming Huai. Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory. *Journal of Physics: Condensed Matter*, 19(16):165209, 2007.

[10] Xiaobin Wang, Yuankai Zheng, Haiwen Xi, and Dimitar Dimitrov. Thermal fluctuation effects on spin torque induced switching: Mean and variations. *Journal of Applied Physics*, 103(3):034507–034507, 2008.

[11] Ashish K Singh, Ku He, Constantine Caramanis, and Michael Orshansky. Mitigation of intra-array sram variability using adaptive voltage architecture. In *ICCAD*, pages 637–644. ACM, 2009.

[12] R. Gabrys, E. Yaakobi, L. Grupp, S. Swanson, and L. Dolecek. Tackling intracell variability in tlc flash through tensor product codes. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 1000–1004, 2012.

[13] A. Agarwal, B.C. Paul, S. Mukhopadhyay, and K. Roy. Process variation in embedded memories: failure analysis and variation aware architecture. *Solid-State Circuits, IEEE Journal of*, 40(9):1804–1814, 2005.

[14] Yue Zhang, Weisheng Zhao, Yahya Lakys, J Klein, Joo-Von Kim, Dafiné Ravelosona, and Claude Chappert. Compact modeling of perpendicular-anisotropy cofeb/mgo magnetic tunnel junctions. *IEEE Trans. Electron Devices*, 59(3):819–826, 2012.

[15] Woojin Kim, JH Jeong, Y Kim, WC Lim, JH Kim, JH Park, HJ Shin, YS Park, KS Kim, SH Park, et al. Extended scalability of perpendicular stt-mram towards sub-20nm mtj node. In *IEDM*, 2011.

[16] Wei Zhao and Yu Cao. Predictive technology model for nano-cmos design exploration. *ACM J. on Emerging Techn. in Comp. Sys. (JETC)*, 3(1):1, 2007.

[17] ITRS. International technology roadmap for semiconductors. URL: http://www.itrs.net/Links/2012ITRS/Home2012.htm, 2012.

[18] S Ikeda, K Miura, H Yamamoto, K Mizunuma, HD Gan, M Endo, S Kanai, J Hayakawa, F Matsukura, and H Ohno. A perpendicular-anisotropy cofeb–mgo magnetic tunnel junction. *Nature Materials*, 9(9):721–724, 2010.

[19] Chi-Keung Luk, Robert Cohn, Robert Muth, Harish Patil, Artur Klauser, Geoff Lowney, Steven Wallace, Vijay Janapa Reddi, and Kim Hazelwood. Pin: building customized program analysis tools with dynamic instrumentation. In *PLDI*, 2005.