Performance, Power, and Reliability Tradeoffs of STT-RAM Cell Subject to Architecture-Level Requirement

Hai Li[1], Xiaobin Wang[2], Zhong-Liang Ong[3], Weng-Fai Wong[3], Yaojun Zhang[4], Peiyuan Wang[4], and Yiran Chen[4]

[1]Department of Electrical and Computer Engineering, Polytechnic Institute of New York University, Brooklyn, NY 11201 USA
[2]Seagate Technology, Bloomington, MN 55435 USA
[3]School of Computing, National University of Singapore, Singapore 117417, Republic of Singapore
[4]Electrical and Computer Engineering Department, University of Pittsburgh, Pittsburgh, PA 15261 USA

Large switching current and long switching time have significantly limited the adoption of spin-transfer torque random access memory (STT-RAM). Technology scaling, moreover, makes it very challenging to reduce the switching current while maintaining the reliability of magnetic tunneling junction (MTJ) to be similar to that of the earlier generations. In this work, we shall exploit a key insight that in the most on-chip caches where STT-RAM is most likely to be deployed, the lifespan of the data stored in the memory cells is much shorter than the data retention time requirement assumed in STT-RAM development, namely, 4 ~ 10 years. We also quantitatively investigated the possibility of trading off MTJ nonvolatility for improved switching performance, e.g., the switching time and/or current, under architectural level guidance. We further proposed and evaluated a hybrid memory design technique that partitions the on-chip STT-RAM cache into two parts with different nonvolatility performances so as to better fit the diverse retention time requirements of different data sets.

Index Terms—**MRAM, nonvolatility, spin-torque, spintronic.**

I. Introduction

SPIN-TRANSFER torque random access memory (STT-RAM) is a promising memory technology. STT-RAM features zero standby power (nonvolatility), small memory cell area (~1/6 of static RAM (SRAM)'s [1]) and nanosecond read access time. However, our previous work has identified that the major technical obstacles for using STT-RAM are its long write access time and high write power dissipation [2]. When the MTJ (magnetic tunneling junction) switching time is reduced down to 10 ns and below, the switching current required has to increase exponentially [3]. Since the clock frequency of modern computing systems are in the multigiga Hz range, the write speed of STT-RAM cells, which is in the order of tens of nanoseconds, may significantly degrade the overall performance of the system.

The nonvolatility of an MTJ is quantitatively measured by the data retention time, which is the maximum time duration for which data can be stored in the MTJ. It is well known that there is a tradeoff between switching performance, i.e., switching time and the required switching current, and the data retention time of the MTJ. Specifically, relaxing the thermal stability of an MTJ can improve its switching performance while reducing its data retention time [3].

This work is based on a key observation: STT-RAMs are most likely to be deployed as on-chip processor caches, and the residency of data in the on-chip cache is much shorter than the default data retention time assumed in STT-RAM designs, namely, 4 ~ 10 years [3], [4]. Therefore, it is possible to shorten the data retention time of the MTJ so as to improve its write performance. Data that are not frequently updated can be redirected to a different part of the cache that is capable of retaining data for a longer period of time so that correctness of the cache's operation is not compromised [16].

II. Preliminary

A. MTJ Writing Performance and Retention Reliability

Competition between the short term dynamic writability and the long term thermal stability (nonvolatility) is the key for MTJ tradeoff between writing performance and data retention reliability. For an MTJ magnetic element, long term retention is determined by magnetization stability energy over thermal energy at ambient temperature

$$\Delta = \left(\frac{K_{\mathrm{eff}} V_{\mathrm{eff}}}{k_{\mathrm{B}} T}\right) = \left(\frac{M_{\mathrm{s}} H_{\mathrm{c}} V_{\mathrm{eff}}}{k_{\mathrm{B}} T}\right) \tag{1}$$

where K_{eff} is the effective anisotropy due to the element's shape and intrinsic crystal anisotropy. V_{eff} is the effective volume considering nonuniform magnetization reversal. M_{s} is the magnetization saturation. H_{c} is the coercivity field including magnetocrystalline anisotropy and shape anisotropy.

When the MTJ switching time is longer than 10 ns, thermally activated process dominates. The relationship between the switching current density and the switching time of the MTJ can be modeled as [3], [5]

$$J_{\mathrm{c}}(T_{\mathrm{sw}}) = J_{\mathrm{c0}}\left(1 - \frac{1}{\Delta}\ln\left(\frac{T_{\mathrm{sw}}}{\tau_0}\right)\right) \tag{2}$$

where T_{sw} is the MTJ switching time. τ_0 is the relaxation time. J_{c0} is the switching threshold current density that causes a spin flip in the absence of any external magnetic fields at 0 K as

$$J_{\mathrm{c0}} = \left(\frac{2e}{\hbar}\right)\left(\frac{\alpha}{\eta}\right)(t_{\mathrm{F}} M_{\mathrm{s}})(H_{\mathrm{k}} \pm H_{\mathrm{ext}} + 2\pi M_{\mathrm{s}}). \tag{3}$$

Here e is the electron charge, α is the damping constant, t_{F} is the magnetic element thickness, $\hbar$ is the reduced Planck's constant, and η is spin current polarization efficiency.

The short term magnetization switching is dominated by magnetization dynamics. The switching time could not be approximated by exponential dependence of energy barrier

Manuscript received February 21, 2011; accepted May 31, 2011. Date of current version September 23, 2011. Corresponding author: H. Li (e-mail: hli@poly.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMAG.2011.2159262

[6], [7]. Moreover, for spin torque induced magnetization switching, the magnetization stability barrier and the critical switching current depends significantly on the magnetization dynamics symmetry [8], [9].

To study tradeoffs between improved write performance and longer data retention, the switching current versus dynamic switching time are calculated for the whole time range, from dynamic magnetization switching in nanosecond time scale to thermal magnetization reversal in the time scale of years. The calculation is based upon solving the stochastic magnetization dynamics equation describing spin torque-induced magnetization motion at finite temperature [7], [10]. The calculations in this manuscript assume the baseline parameters of the 45 nm technology node in [11].

B. Tradeoffs Between MTJ Performance and Nonvolatility

Given a certain switching current, some possible approaches to improve the MTJ switching time include increasing η, reducing α, M_{s}, H_{k}, and t_{F} [11], and applying the assisting magnetic field H_{ext} [12].

The MTJ data retention time is mainly determined by the magnetization stability energy barrier height Δ

$$t_{\mathrm{store}} = \frac{1}{f_0} e^{\Delta}. \tag{4}$$

Here f_0 is the thermal attempt frequency, which is of the order of 1 GHz for storage purposes [3]. For long-term data storage applications, e.g., standalone and mobile memory, a Δ value of $40 \sim 60$ is required for a data-retention time of $4 \sim 10$ years [3], [4].

It should be pointed out that (4) is the mean switching time for a magnetic element with a given magnetic energy barrier over thermal agitation energy Δ. For magnetic memory design, we need to consider not only the probabilistic nature of thermal magnetization switching, but also the energy barrier distributions due to device process variations. The bit error probability of a memory element can be derived after integrating the magnetization switching probability and the energy barrier distributions. The detailed derivation can be found in Appendix I of [13]. The corresponding bit error probability for memory magnetic elements can be linked to the average $\overline{M}(t)$ over M_s as (13) in [13] shows. For magnetic elements with log normal energy barrier distribution

$$\rho(\Delta) = \frac{1}{\beta\Delta\sqrt{\pi}} \exp\left(-\frac{1}{\beta}\ln\frac{\Delta}{\Delta_0}\right). \tag{5}$$

The probability of error is

$$error = \frac{1}{2}\left[1 - erf\left(\frac{1}{\beta}\ln\frac{\Delta_0\exp(\beta^2/2)}{\ln(2f_0 t/\ln 2)}\right)\right]. \tag{6}$$

For other energy barrier distributions, we can derive similar formula.

We note that some approaches for improving MTJ switching performance, such as reducing $M_{\mathrm{s}}, H_{\mathrm{k}}, t_{\mathrm{F}}, V$, will lower Δ and degrade the data retention time of the MTJ. Also, as (1) and (4) show, the data retention time of an MTJ exponentially decreases when its working temperature T rises.

Fig. 1 shows the relationship between the switching current and the switching time for a 45 × 90 nm elliptical MTJ. The MTJ data retention time is estimated as the switching time when the switching current is zero. When the working temperature is increased from 275 K to 350 K, the MTJ data retention time decreased from 6.7×10^6 years to 4.27 years.

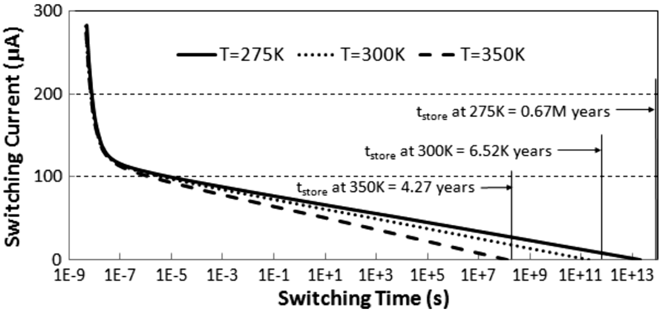


Fig. 1. Relationship between the switching current and the switching time of “base” MTJ design.

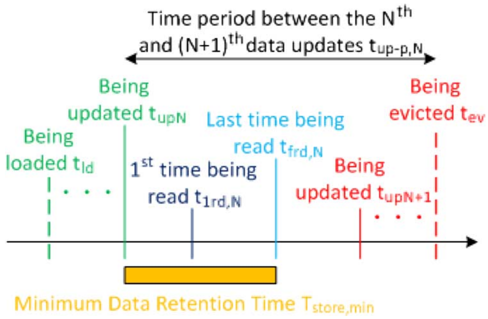

Time period between the N th and (N+1) th data updates t up-p,N
Being updated t upN
Being loaded t ld
1 st time being read t 1rd,N
Last time being read t frd,N
Being updated t upN+1
Being evicted t ev
Minimum Data Retention Time T store,min

Fig. 2. Lifespan of a data in the cache.

C. On-Chip Data Storage Patterns

Unlike standalone and mobile memories that may have to store data for long time periods, the data stored in the on-chip cache of a computing system are frequently updated. Fig. 2 shows the typical lifespan of a cache data. After a data block at address A1 is loaded into the cache at CA1 at time t_{ld}, it may be updated many times before it is evicted from the cache. We use t_{upN} to denote the time at which the block is updated for the Nth times where $\mathrm{N} = 0$ denotes the time at which the block is first loaded into the cache. To ensure the stored data are stable when being accessed, the minimum data retention time $t_{\mathrm{store,min}}$ must be longer than the maximum time duration between the time the block is updated (or loaded in) and the last time that it is read, or $max(t_{\mathrm{frd,N}} - t_{\mathrm{upN}})$. If the data block is evicted from the cache at t_{ev} after $t_{\mathrm{frd,N}}$, any attempt to access it will cause a cache miss.

However, $max(t_{\mathrm{frd,N}} - t_{\mathrm{upN}})$ varies significantly from block to block, and from program run to program run. In some extreme cases, some reserved data, for example, a block of instructions, may be stored in the cache for a very long time without being updated but is nevertheless frequently accessed. In other cases, data blocks may be frequently updated without ever being read between updates (for example, some sort of counters). Choosing $max(t_{\mathrm{frd,N}} - t_{\mathrm{upN}})$ to be one or the other extreme may end up as being too optimistic, or too pessimistic.

III. Magnetic and Architectural Evaluations

A. Tradeoff in MTJ Designs

To evaluate the impact of relaxing the nonvolatility of MTJ devices on switching performance, we simulated the required

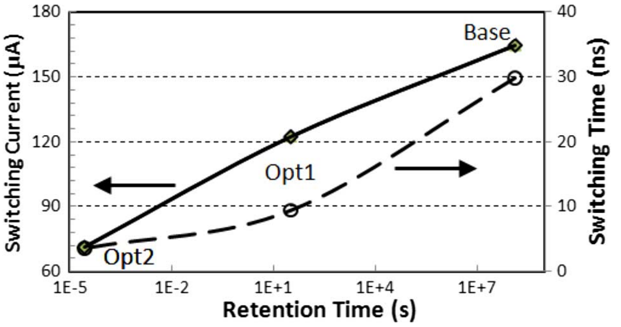


Fig. 3. Critical switching current improvement when relaxing the data retention time for 10 ns switching time at 350 K; and the switching time improvement when relaxing the data retention time under a 125 μA switching current at 350 K.

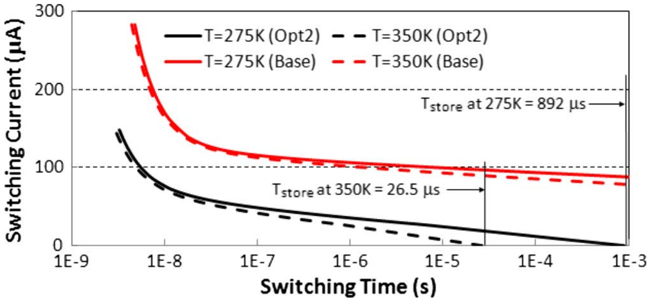
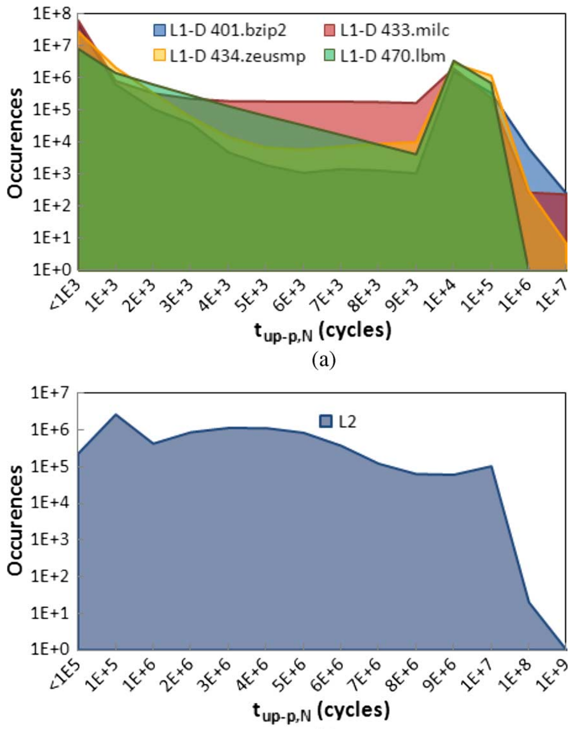


Fig. 4. MTJ switching performances at different MTJ data retention time specifications.



(a)

(b)

Fig. 5. Statistics of the duration between the time the data is updated (or loaded in) and the last time the data are read (a) L1-Data Cache. (b) L2 cache.

TABLE I
CONFIGURATIONS OF THE SIMULATED L1-D AND L2 CACHES

	L1-D Cache	L2 Cache
Size (Byte)	32768	4194304
Associativity	8	16
Cache block size (Byte)	64	64
Read Latency (cycles)	3	14

switching current of three MTJ designs with a 45 × 90 nm ellipse shape. The "baseline" MTJ design ensures a data retention time up to 4.27 years at 350 K while the data-retention data of the other two designs ("Opt1" and "Opt2") are optimized for switching performance. As shown in Fig. 3, when the MTJ data retention time is scaled from 4.27 years to 265 μs, the required MTJ switching current decreases from 164.5 μA to 71.4 μA for a 10 ns switching time at 350 K. At an MTJ switching current of 125 μA, the corresponding switching times of all three MTJ designs vary from 29.8 ns to 3.6 ns, an improvement of more than 8X, as shown in Fig. 3.

B. Temperature Dependency

The scaling of the magnetization stability energy barrier height Δ leads to an increased temperature sensitivity of the MTJ switching performance, as shown in Fig. 4. The shift of the switching performance curve under the same temperature difference, i.e., from 275 K to 350 K, increases when the MTJ design is changed from "baseline" to "Opt2."

C. Statistics of Cache Access Patterns

An on-chip cache is made up of many cache blocks, which is the unit of reads and writes. Set associative caches were introduced to alleviate the collision between multiple memory blocks that maps to one cache block [14]. In an M-way set associativity cache hierarchy, every memory block can be loaded into one of the M possible cache blocks.

We note that the cache access patterns, e.g., the duration between the time a block is updated (or loaded into the cache), and the time it was last read $(t_{\mathrm{frd,N}} - t_{\mathrm{upN}})$, varies from cache blocks to blocks, from program to program, or even from time to time. Fig. 5 depicts our simulated statistics on the $(t_{\mathrm{frd,N}} - t_{\mathrm{upN}})$ of cache blocks of the four L1 data (D) caches and the shared L2 cache in a quad-core microprocessor. Different SPEC benchmarks, namely 401.bzip2, 433.milc, 434.zeusmp, and 470.lbm, were executed on each CPU core, respectively [15]. The configurations of L1-D cache and L2 cache are shown in Table I. A total of two billion instructions were simulated. For the L1-D cache, the majority of the cache data, more than 95%, are accessed within the first 10^5 clock cycles after they are loaded or updated. In some benchmarks, i.e., 401.bzip2, this number can be as high as 99%. Similar scenario exists in L2 cache. However, we do observe that a small portion of cache data can be active for a long time: in benchmark 401.bzip2, there are about 5.1×10^6 instances where the time between a write and the last time the data are read exceeded 10^6 clock cycles. However, it is this small portion of cache block that may be active for a very long time that determines the minimum MTJ data retention time where we can have the tradeoffs to improve switching performance.

IV. Hybrid Associative Cache Design

We proposed a hybrid associative cache design that exploits the relaxation of MTJ nonvolatility that was originally limited by "long-life" cache data. The idea is shown in Fig. 6. A number of the ways of an N-way set-associative cache are implemented using STT-RAM cells with a scaled data-retention time and improved switching performance. The remaining ways are built

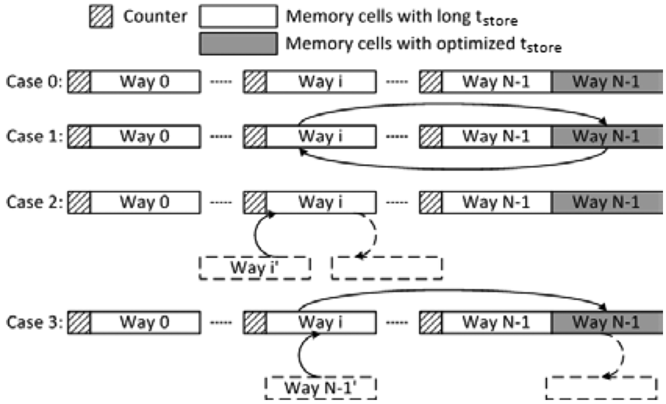


Fig. 6. Hybrid associative cache design.

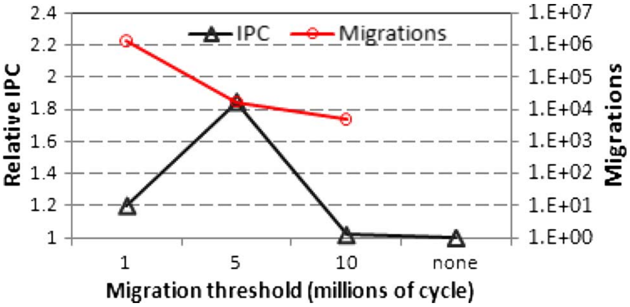


Fig. 7. Performance of hybrid associative cache design.

from STT-RAM cells with a higher retention time, and hence a relatively slower write speed. Every optimized cache block is augmented with a counter that stores the time when the cache block is updated. Since the granularity of MTJ data retention time is much larger than the clock cycles, the bit number of counters can be significantly reduced by incrementing the counter at a low frequency, say once every 10^3 clock cycles.

During a write access that updates an existing block, the hybrid cache works exactly like a normal cache. During a read access to an optimized cache block, the data last updated time are first checked. If the counter shows that the data stored in the memory block are stable and still far from the data retention time limit (Case 0 in Fig. 6), the cache block is returned just like a nonoptimized cache block. However, if the check reveals that the data are approaching the end of its lifetime, then besides reading the data (Case 1 in Fig. 6), the data is migrated into the non-optimized portion of the cache. Since most of the data will be updated before they reach the data-retention time threshold, Case 1 occurs infrequently. If a new block is loaded into an optimized cache block which is still far from the end of its lifetime, a normal data replacement operation is performed (Case 2 in Fig. 6). However, if the data are being loaded into the optimized cache block, we will search for a cache block with the oldest update time that is beyond the data-retention time threshold within the same cache set. If such a cache block is found, say, cache block i, it will be moved to the nonoptimized cache block and the new data are written into cache block i (Case 3 in Fig. 6). Otherwise, the nonoptimized cache block will be replaced directly. Just like the case where the cache block fails to be updated before its lifetime expires, the block will be reloaded from the lower level cache or the main memory. Here we are assuming a write-through cache.

To evaluate the efficiency of our proposed hybrid cache design, we compare the performance of using the uniform nonoptimized STT-RAM cache hierarchy and the hybrid STT-RAM cache hierarchy. The results are shown in Fig. 7. We use the optimized “Opt2” MTJ design in Section III-A to implement half of the ways in a 16-way L2 cache. Three data-retention time (migration) threshold of 1×10^6, 5×10^6, and 10×10^6 clock cycles were simulated. On a 3 GHz processor, the last value would correspond to 3.3 ms. Using a migration threshold nearing the average L2 cache block lifetime, our hybrid cache showed up to 80% performance improvement in terms of the instructions processed per cycle (IPC) compared to the nonoptimized design (with a relative $\mathrm{IPC} = 1$) for the four simulated benchmarks.

V. Conclusion

In this work, we analyzed the access patterns for on-chip caches, and evaluated the possibility of relaxing the MTJ data retention time so as to improve switching performance. Our results show that the majority of cache data stay active for much shorter time duration than the data-retention time assumed in current STT-RAM designs. By introducing a hybrid design, the nonvolatility of most STT-RAM cells in an on-chip cache can be aggressively reduced in return for significant switching performance and power improvements.

References

[1] X. Dong, “Circuit and microarchitecture evaluation of 3D stacking magnetic RAM (MRAM) as a universal memory replacement,” in *Proc. 45th Design Automation Conf.*, Jun. 2008, pp. 554–559.

[2] G. Sun *et al.*, “A novel architecture of the 3D stacked MRAM L2 cache for CMPs,” in *IEEE 15th Int. Symp. High Perform. Comput. Architect.*, Feb. 2009, pp. 239–249.

[3] Z. Diao *et al.*, “Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory,” *J. Phys.: Cond. Matter*, vol. 19, no. 16, p. 165209, 2007.

[4] D. Weller and A. Moser, “Thermal effect limits in ultrahigh-density magnetic recording,” *IEEE Trans. Magn.*, vol. 35, no. 6, pp. 4423–4439, Nov. 1999.

[5] A. Raychowdhury *et al.*, “Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances,” in *IEEE Int. Electron Devices Meeting*, Dec. 2009, pp. 1–4.

[6] N. H. Bertram *et al.*, “Dynamic-thermal effects in thin film media,” *IEEE Trans. Magn.*, vol. 37, no. 4, pp. 1521–1527, Jul. 2001.

[7] X. Wang, Y. Zheng, H. Xi, and D. Dimitrov, “Thermal fluctuation effects on spin torque induced switching: Mean and variations,” *J. Appl. Phys.*, vol. 103, p. 034507, 2008.

[8] X. Wang, W. Zhu, H. Xi, and D. Dimitrov, “Relationship between symmetry and scaling of spin torque thermal switching barrier,” *Appl. Phys. Lett.*, vol. 93, p. 102508, 2008.

[9] X. Wang, “Magnetization dynamics symmetry in spin torque induced magnetization switching,” *Symmetry*, vol. 2, no. 2, p. 99, 2010.

[10] X. Wang *et al.*, “Spin torque induced magnetization switching variations,” *IEEE Trans. Magn.*, vol. 45, no. 4, p. 2038, 2009.

[11] X. Wang, Y. Chen, H. Li, D. Dimitrov, and H. Liu, “Spin torque random access memory down to 22 nm technology,” *IEEE Trans. Magn.*, vol. 44, no. 11, pp. 2479–2482, Nov. 2008.

[12] Y. Chen *et al.*, “Combined magnetic- and circuit-level enhancements for the nondestructive self-reference scheme of STT-RAM,” in *Proc. 16th Int. Symp. Low Power Electron. Design*, Aug. 2010, pp. 1–6.

[13] N. H. Bertram, X. Wang, and V. L. Safonov, “Dynamic-thermal effects in thin film media,” *IEEE Trans. Magn.*, vol. 37, no. 4, pp. 1521–1527, 2001.

[14] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, 3rd ed. San Mateo, CA: Morgan Kaufmann, 2002.

[15] D. Weaver, Pre-Compiled Little-Endian Alpha ISA SPEC2000 Binaries [Online]. Available: http://www.eecs.umich.edu/~chriswea/benchmarks/spec2000.html

[16] C. Smullen, IV *et al.*, “Relaxing non-volatility for fast and energy-efficient STT-RAM caches,” in *Proc. 2011 HPCA*, 2011.