Cache Revive: Exploiting Volatility of STT-RAM Caches for Enhanced Performance in CMPs.

Anonymous MICRO submission

Paper ID 761

Abstract

Spin-Transfer Torque RAM (STT-RAM) is a CMOS compatible emerging non-volatile memory (NVM) technology that has the potential to replace the conventional on-chip SRAM caches for designing a more efficient memory hierarchy for future multicore architectures. However, it's high write latency and dynamic write energy are major obstacles for being competitive with the SRAM-based cache hierarchy. On the other hand, STT-RAM technology has another adaptable feature that it is possible to reduce its write latency by reducing its retention time, thereby making it volatile. In this paper, we exploit this volatile property of the STT-RAM for designing an efficient L2 cache architecture. The paper addresses several critical design issues such as how do we decide a suitable retention time for last level cache, what is the relationship between retention time and write latency, and how do we architect the cache hierarchy with a volatile STT-RAM. Through an extensive execution driven analysis of the inter-write time of several PARSEC and SPEC 2006 banchmarks, we observe that retention time in the order of 10-40 ms is a good design point to handle most of the writes. Then for the rest of the cache blocks that have a higher inter-write time than the STT-RAM retention time, we propose an architectural solution to identify these blocks with a per block 2 bit counter, temporarily save a limited number of MRU blocks in a buffer, and writeback the rest of the dirty blocks to avoid any data loss. Our experiments with 4

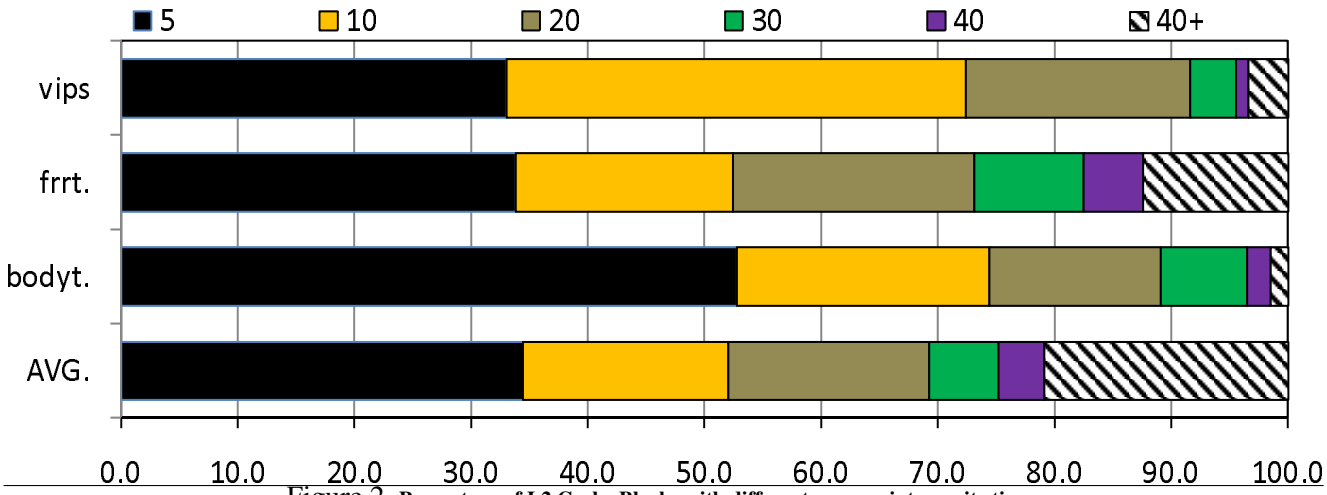

Figure 2. **Percentage of L2 Cache Blocks with different average inter-write times**

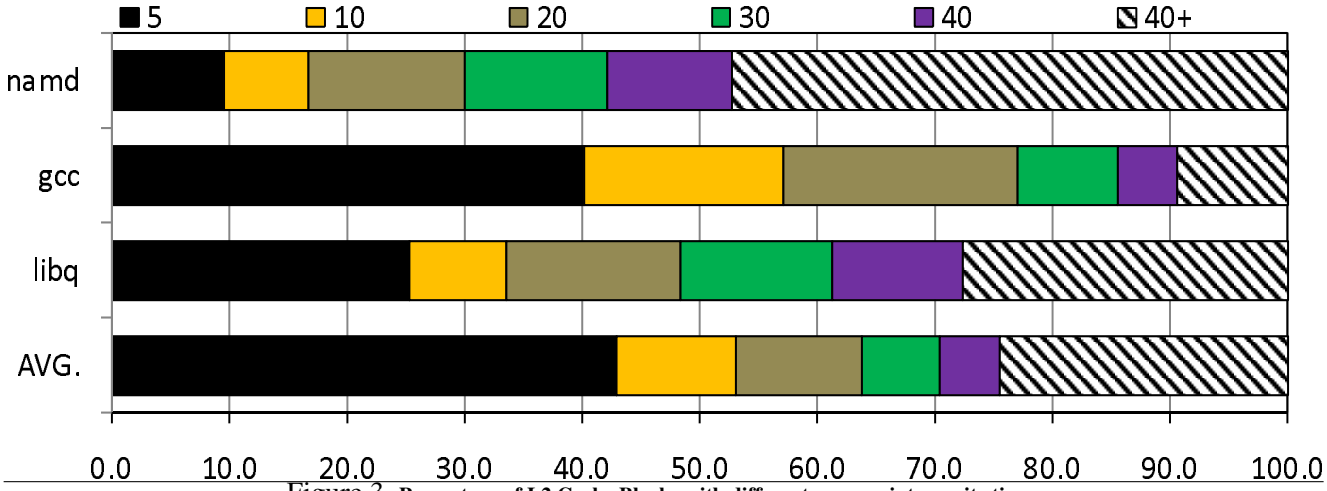

Figure 3. **Percentage of L2 Cache Blocks with different average inter-write times**

and 8-core architectures with an SRAM-based L1 cache and STT-RAM-based L2 cache indicate that not only we can eliminate the high write overhead of an NVM STT-RAM, but can provide on an average 10-12% improvement in IPC compared to the traditional SRAM-based design, while reducing the energy consumption significantly

1. Introduction

2. Motivation

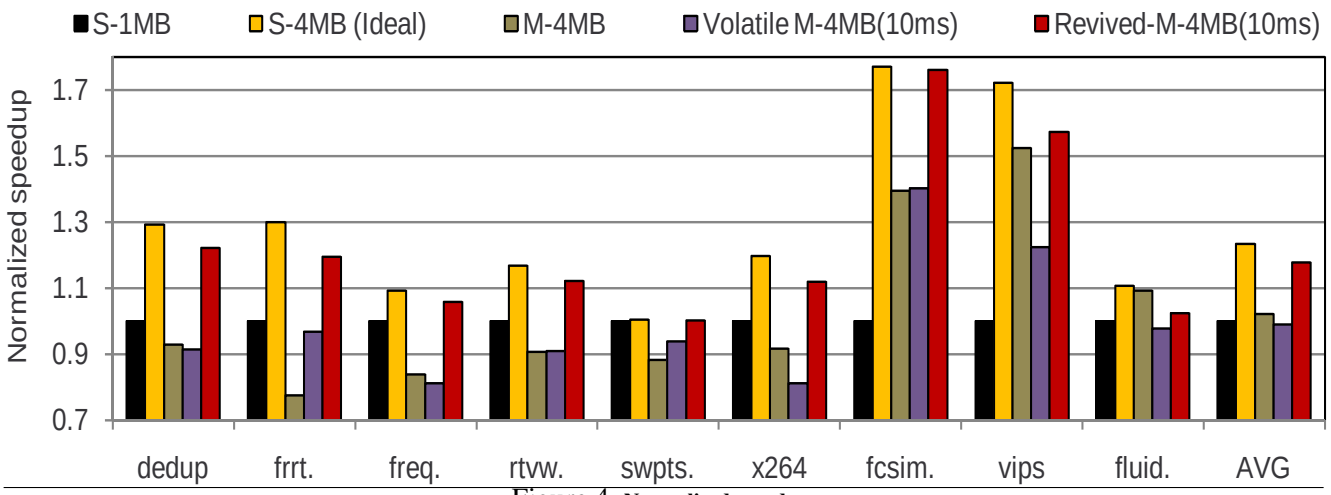

Figure 4. **Normalized speedup**

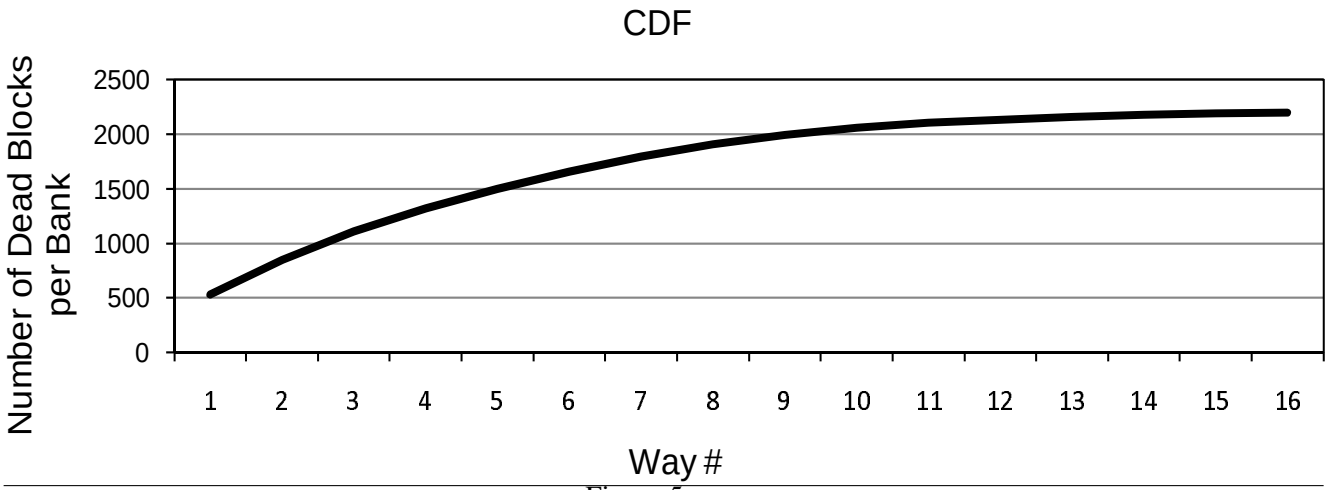

Figure 5. **CDF**

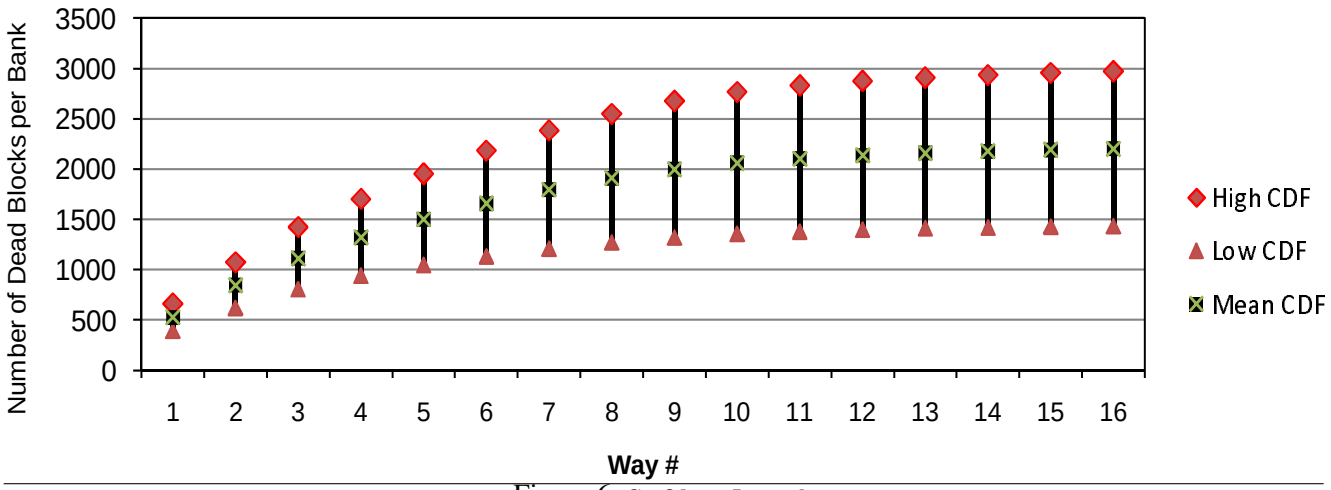

Figure 6. **Confidence Intervals**

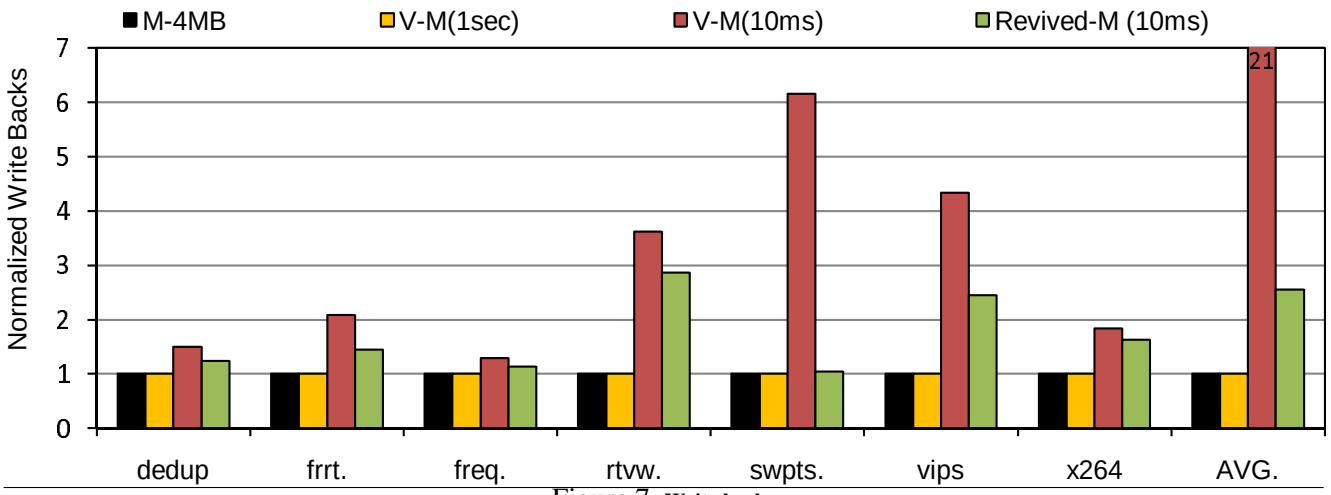

Figure 7. **Write backs**

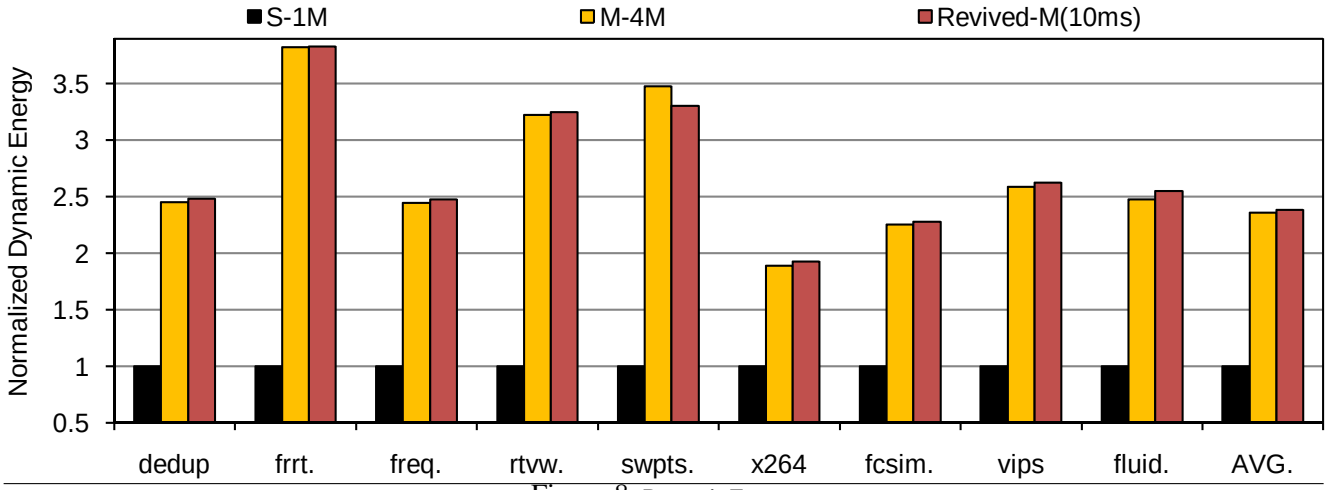

Figure 8. **Dynamic Energy**

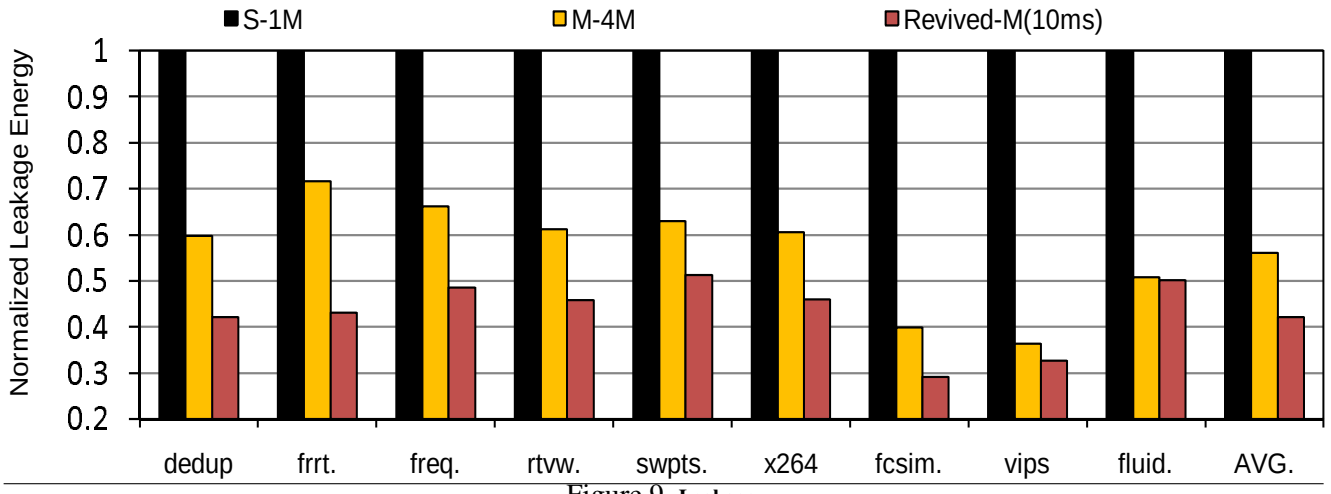

Figure 9. **Leakage**

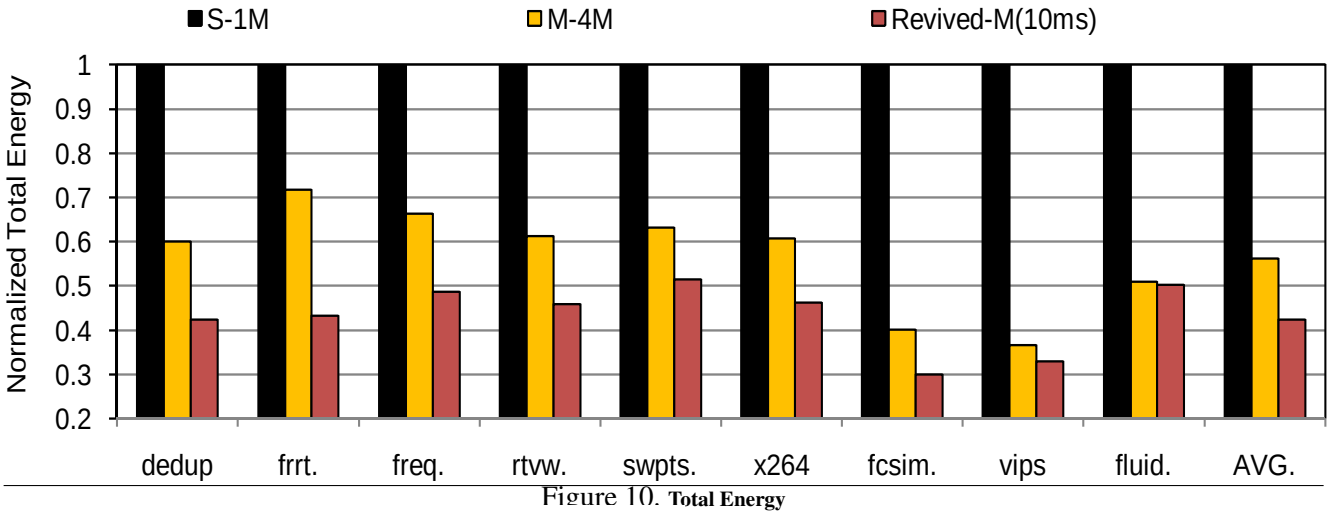

Figure 10. **Total Energy**

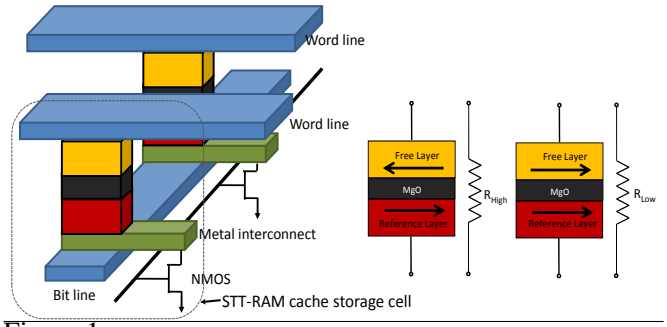

Figure 1. **(a) Structural view of of STT-RAM Cache Cell (b) Anti Space Parallel (High Resistance, Indicating "1" state (c) Parallel (Low Resistance, Indicating "0" state**

Architects always envision to have a cache hierarchy which not only has fast access times but also consume very less leakage and dynamic power.To bridge the gap between the current and the utopian cache hierarchy, exploration of new properties of emerging memory technologies is imperative. [**?**] et.al proposed that reducing the planar area of STT-RAM, reduces the retention time which in turn leads to lower write time. In current era, where the STT-RAM cell size is in the range of 3-5 F^2, there is not enough scope for reducing write time by this technique. Figure shows how write time reduces when retention time is reduced by means of changing the write pulse duration. The 10 year retention time STT-RAM cache design is traditional non-volatile STT-RAM. As we observe from the graph that write time significantly reduces, as we tend towards volatile STT-RAM.

The reduction of retention time of cache blocks has opened plethora of challenges for the architects. The main challenges are (1) To avoid any data loss because of volatility of cache blocks. (2) To ensure that the data is correct and no random bit flips have happened. 3) To architect cache hierarchy, which deals with issues (1) and (2) and still reap performance and energy benefits of reduced write time. This paper systematically addresses these challenges by finding a ideal retention time which can lead to

minimum possible write latency with minimal architectural overheads.

By means of many experiments we came up with a term called revival time which is defined as the time between consecutive writes to the same block. We can say that after every revival time interval, physical cache block is refreshed and ready to be used again.

Figure **??** shows the percentage of L2 cache blocks binned into different average revival time slots for PARSEC and SPEC 2006 Benchmarks. We collected these results by modeling a 4MB STT-RAM L2 Unified Banked Cache using M5 Simulator with 2GHz processor consisting of 4 cores.Table contains the details of the configuration of the simulated system. While collecting results, we ensured that block is valid when the consecutive writes are performed on this block. If the block gets invalidated in between, we only consider the time between, when the block is last written and when the block gets invalidated.

We observe from the graph that there are significant number of blocks which gets refreshed frequently and also a good percentage of blocks which remain unrefreshed for longer time. This graph gives us the basis on which we can choose the optimal retention time. Reducing the retention time too much will make the cache too volatile and increasing it will aggravate the write time. We choose 10ms as the optimal retention time, as there are majority of blocks which gets refreshed within 10ms and hence there is no worry of them getting lost. There is a good probability that the hashed blocks in the figure can get kicked out from the cache by LRU replacement policy as they may not have been accessed in near past, and it is ok to loose them unless they are dirty. There could be some blocks in the same region which are frequently read and not written. In section 5, we describe how these types of blocks are dealt with. We also propose a scheme in Section 5 to handle the blocks which are in the region 10-40ms.

3. STT-RAM Design

3.1. STT-RAM Cell Basics

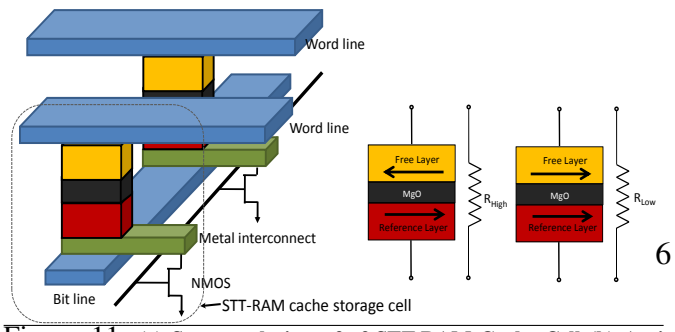

Figure 11. **(a) Structural view of of STT-RAM Cache Cell (b) Anti Space Parallel (High Resistance, Indicating "1" state (c) Parallel (Low**

STT-RAM uses Magnetic Tunnel Junction (MTJ) as the memory storage and leverages the difference in magnetic directions to represent the memory bit. As shown in Fig. 11,

MTJ contains two ferromagnetic layers. One ferromagnetic layer is has fixed magnetization direction and it is called the reference layer, while the other layer has a free magnetization direction that can be changed by passing a write current and it is called the free layer. The relative magnetization direction of two ferromagnetic layers determines the resistance of MTJ. If two ferromagnetic layers have the same directions, the resistance of MTJ is low, indicating a "1" state; if two layers have different directions, the resistance of MTJ is high, indicating a "0" state.

As shown in Fig. 11, when writing "0" state into STT-RAM cells, positive voltage difference is established between SL and BL; when writing "1" state, vice versa. The current amplitude required to reverse the direction of the free ferromagnetic layer is determined by the size and aspect ratio of MTJ and the write pulse duration.

3.2. Write current versus write pulse width trade-off

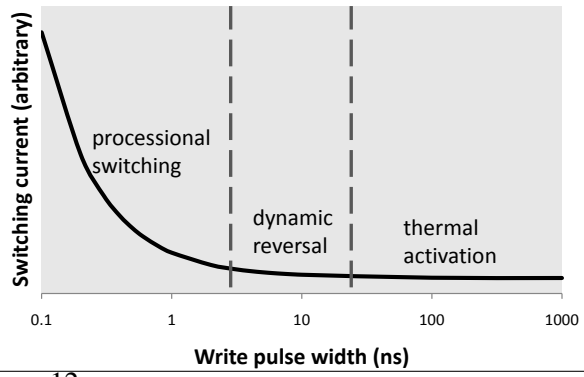

Figure 12. **(a) Structural view of of STT-RAM Cache Cell (b) Anti Space Parallel (High Resistance, Indicating "1" state (c) Parallel (Low Resistance, Indicating "0" state**

The current amplitude required to reverse the direction of the free ferromagnetic layer is determined by a lot of factors such as material property, device geometry and importantly the write pulse duration. Generally, the longer the write pulse is applied, the less the switching current is needed to switch the MTJ state. Three distinct switching modes were identified [3] according to the operating range of switching pulse width τ: thermal activation ($\tau > 20ns$), processional switching ($\tau < 3ns$) and dynamic reversal ($3ns < \tau < 20ns$).

The relationship between switching current density J_c and write pulse width τ was characterized by

an analytical model in [8]. The equations are listed as follows,

$$J_{c,TA}(\tau) = J_{c0}\{1 - (\frac{k_B T}{E_b})ln(\frac{\tau}{\tau_0})\} \quad (1)$$

$$J_{c,PS}(\tau) = J_{c0} + \frac{C}{\tau^{\gamma}} \quad (2)$$

$$J_{c,DR}(\tau) = \frac{J_{c,TA}(\tau) + J_{c,PS}(\tau)e^{-k(\tau-\tau_c)}}{1 + e^{-k(\tau-\tau_c)}} \quad (3)$$

where $J_{c,TA}$, $J_{c,PS}$, $J_{c,DR}$ are the switching current densities for thermal activation, precessional switching and dynamic reversal respectively. J_{c0} is the critical switching current density, k_B is the Boltzmann constant, T is the temperature, E_b is the thermal barrier, and τ_0 is inverse of the attempt frequency. C, γ, k, and τ_c are fitting constants. Based on the observation from Fig. 12 and analysis of the analytical model, we found very different switching characteristics in the three switching modes. For example, in thermal activation mode, the required switching current increases very slowly even we decrease the write pulse width by orders of magnitude, thus short write pulse width is more favorable in this regime because reducing write pulse can reduce both write latency and energy without much penalty on read latency and energy. While in processional switching, write current goes up rapidly if we further reduce write pulse width, therefore minimum write energy of the MTJ is achieved at some particular write pulse width in this regime. Consequently, this paper will focus on the exploration of write pulse width in processional switching and dynamic reversal to optimize for different design goals.

3.3. STT-RAM Cell Area Modeling

To simulate the performance of STT-RAM cache, it is important to estimate its cell area first. As mentioned before, each 1T1J STT-RAM cell is composed of one NMOS and one MTJ. The NMOS access device is connected in series with the MTJ. The size of NMOS is constrained by both SET and RESET current, which are inversely proportional to the writing pulse width. In order to estimate the current driving ability of MOSFET devices, a small test circuit using HSPICE with PTM 45nm HP model [10] is simulated. The BL-to-SL current and SL-to-BL current are obtained by assuming typical TMR (120%) and LRS ($3k\Omega$) value [7] and bursting wordline voltage to be 1.5V (the optimal value

is extracted from [1]). And we over size the access transistor width to guarantee enough write current provided to MTJ using the methodology discussed in [9]. To achieve high cell density, we model the STT-RAM cell area by referring to DRAM design rules [6]. As a result, the cell size of a STT-RAM cell is calculated as follows,

$$\text{Area}_{\text{cell}} = 3\,(W/L + 1)(F^2) \tag{4}$$

3.4. Impact of MTJ Retention Time on STT-RAM

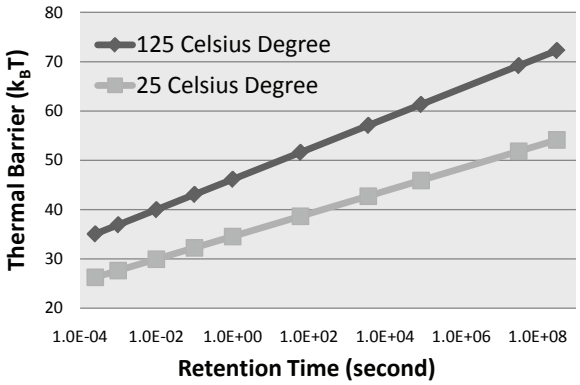

Figure 13. **MTJ thermal stability requirement for different retention time**

The retention time of a MTJ is largely determined by the thermal stability of the MTJ. The relation between retention time and thermal barrier is captured in Figure 13, which can be modeled as $t = C \times e^{k\Delta}$, where t is the retention time and Δ is the thermal barrier while C and k are fitting constants. Thermal stability of the free layer in an MTJ does not only have impact on retention time of STT-RAM memory cell but also on the write current. It was found in **??** that the switching current of MTJ increases almost linearly with thermal barrier when thermal barrier is $< 70k_BT$, where k_B is the Boltzman constant and T is temperature. Here we combine this observation with the write current versus write time trade-off described in Section 3.2, which essentially means that once the thermal barrier of a MTJ is lowered we are able to achieve faster write speed or/and smaller write current/energy. The most straightforward way to reduce thermal barrier is to tune device geometry such as planar area, thickness of free layer and aspect ratio of the elliptic MTJ.

3.5. STT-RAM Cache Simulation Setup

We simulate SRAM-based caches and STT-RAM-based caches with a tool called NVsim [4], which is a circuit-level performance, energy, and area simulator based on CACTI for emerging non-volatile memories. All the models described in this Section has been integrated in NVsim. The simulation

Table 1. 16-way L2 Cache Simulation Results

			Area (mm^2)	Read Latency (ns)	Write Latency (ns)	Leakage Power (mW)
$1MB$ SRAM			2.612	1.012	1.012	4542
$4MB$ STT-RAM	$t = 10yr$	Leakage Opt.	2.628	2.434	4.919	1399
		Latency Opt.	3.003	0.998	10.61	2524
	$t = 1s$	Leakage Opt.	2.203	2.044	3.552	1388
		Latency Opt.	2.904	0.973	5.571	2235
	$t = 100ms$	Leakage Opt.	2.181	1.994	3.432	1250
		Latency Opt.	2.902	0.963	3.002	2230
	$t = 10ms$	Leakage Opt.	2.167	1.956	3.390	1151
		Latency Opt.	2.901	0.959	2.598	2227

results are listed in Table 1. We can see that the leakage-optimized $4MB$ non-volatile STT-RAM cache has almost the same area with $1MB$ SRAM. This is consistent with previous work [5]. By relaxing retention time of STT-RAM with lower thermal barrier, the leakage-optimized STT-RAM cache can have smaller area, faster write latency and less leaky peripheral circuity. However, the read latency of leakage-optimized $4MB$ STT-RAM cache is significantly larger than $1MB$ SRAM cache because sensing the state of STT-RAM cell takes longer and fast SRAM sensing. Thus, we reduce the array size to improve the latency of STT-RAM cache. As can be seen in Table 1, the latency-optimized STT-RAM cache has noticeable better delay with $14\% - 34\%$ area overhead compared to leakage-optimized STT-RAM cache with the same retention time.

4. Architecting Volatile STT-RAM

5. Experimental Evaluation

6. Results

7. Prior Work

8. Conclusions

Sample bibliography [2]".

References

[1] S. Chatterjee, M. Rasquinha, S. Yalamanchili, and S. Mukhopadhyay. A scalable design methodology for energy minimization of STTRAM: a circuit and architecture perspective. *IEEE Transactions on Very Large Scale Integration*, 2010. 9

[2] Y. Chen, X. Wang, H. Li, L. H., and D. V. Dimitrov. Design Margin Exploration of Spin-Torque Transfer RAM (SPRAM). In *Quality Electronic Design, 2008. ISQED 2008. 9th International Symposium on*, pages 684–690, 2008. 10

[3] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L.-C. Wang, and Y. 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. 7

[4] X. Dong, N. P. Jouppi, and Y. Xie. PCRAMsim: System-level performance, energy, and area modeling for phase-change RAM. In *Proceedings of the International Conference on Computer-Aided Design*, pages 269–275, 2009. 9

[5] X. Dong, X. Wu, G. Sun, Y. Xie, H. Li, et al. Circuit and Microarchitecture Evaluation of 3D Stacking Magnetic RAM (MRAM) as a Universal Memory Replacement. In *Proceedings of the Design Automation Conference*, pages 554–559, 2008. 10

[6] F. Fishburn, B. Busch, J. Dale, D. Hwang, et al. A 78nm $6F^2$ DRAM technology for multigigabit densities. In *Proceedings of the Symposium on VLSI Technology*, pages 28–29, 2004. 9

[7] C. Lin, S. Kang, Y. Wang, K. Lee, X. Zhu, W. Chen, X. Li, W. Hsu, Y. Kao, M. Liu, Y. Lin, M. Nowak, N. Yu, and L. Tran. 45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell. In *Proceedings of International Electron Devices Meeting*, pages 57 –59, 2009. 8

[8] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. De. Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances. In *Proceedings of International Electron Devices Meeting*, pages 707–710, 2009. 8

[9] W. Xu, H. Sun, X. Wang, Y. Chen, and T. Zhang. Design of last-level on-chip cache using spin-torque transfer RAM (STT RAM). *IEEE Transactions on Very Large Scale Integration*, 19(3):483 –493, 2011. 9

[10] W. Zhao and Y. Cao. New generation of predictive technology model for sub-45 nm early design exploration. *IEEE Transactions on Electron Devices*, 53(11):2816 –2823, nov. 2006. 8