Area, Power, and Latency Considerations of STT-MRAM to Substitute for Main Memory

Youngbin Jin, Mustafa Shihab and Myoungsoo Jung
Computer Architecture and Memory Systems Laboratory
Department of Electrical Engineering, The University of Texas at Dallas
{youngbin.jin, mustafa.shihab, jung}@utdallas.edu

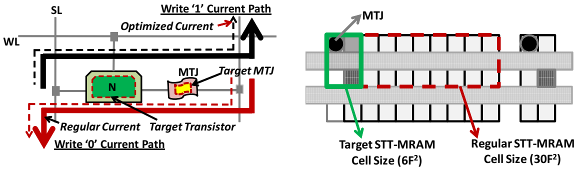
Abstract—STT-MRAM is one of the most promising non-volatile memory technologies with the potential of becoming a universal memory. However, because of its area, power and latency limitations, STT-MRAM is facing critical bottlenecks in substituting DRAM for main memory. Compared to modern DRAM technology, STT-MRAMs cell area and write power consumption are about four times larger and higher, respectively. In this paper, we study diverse device-level parameters of STT-MRAM to make the storage capacity of STT-MRAM comparable to DRAM with better performance as well as power consumption behavior. We then present analytic models to finely tune the thermal stability factor, which is related to STT-MRAM's magnetic tunnel junction (MTJ) and the corresponding transistor, and address the challenges that storage-class STT-MRAM faces in replacing DRAM as a working memory. Our preliminary evaluation results show that, our early-stage optimized STT-MRAM can offer shorter latency and lower power consumption than a baseline DRAM by on average 18.4% and 66.2%, respectively.

◆

1 INTRODUCTION

Dynamic Random Access Memory (DRAM) is being used as the main memory in all forms of modern computing devices, thanks to its successful development and implementation. However, DRAMs face two main challenges in modern computer architecture: i) memory scaling and ii) power consumption. Specifically, computing is growing around two times faster than DRAM storage capacity or bandwidth. In addition, power becomes the primary design constraint in diverse computing domains ranging from embedded system to graphics processing unit (GPU) to high performance computing (HPC), while DRAM requires high operating power and frequent refresh cycles to preserve the data in their volatile storage medium. Even though many studies have been performed to achieve low power and energy efficient DRAM technologies, there is dearth of non-volatile memory (NVM) technology optimizations to directly replace them as a main memory, which can address both of DRAM's scaling-down and power consumption problems.

Spin-Transfer Torque Magnetoresistive RAM (STT-MRAM) is a promising memory technology with the potential of becoming a universal memory. In the recent years, it has received serious attention from the research community as an attractive candidate for replacing both SRAM as a cache, and DRAM as the main memory. Prevailing studies have looked at using STT-RAM with the intention of exploiting its scalability, zero leakage power, high endurance advantages. Specifically, [17] explores the possibility of using STT-MRAM technology to completely replace DRAM in main memory by using partial write and row buffer write bypass. [18] indicates STT-MRAM as a DRAM replacement, in an attempt to reduce system-level power consumption and remove DRAM refresh time in HPC domain. While these prior studies provide detailed insights and observations to replace DRAM with STT-MRAM, their efforts unfortunately have missed out on the vital issue of the memory-level area/density mismatch between DRAM and the STT-MRAM. Even though STT-MRAM has one-transistor one-register (1T1R) structure similar to DRAMs one-transistor one-capacitor (1T1C), its transistor size in practice is much bigger than DRAMs one. The STT-MRAM cell ($32F^2$) is around 4x larger than the DRAM cell ($6 \sim 8F^2$), which renders STT-MRAM difficult to attain DRAM-like density. Further, the write power of STT-MRAM is 3.5x higher than usual DRAM specifications, which might not be acceptable in many computing domains.



(a) (b)

Fig. 1: STT-MRAM optimization: (a) circuit diagram (b) cell layout.

Although these cell area and power concerns are a hitch for STT-MRAM, it is far from a dead-end for STT-MRAM's future as main memory. We believe, it opens up a whole new set of opportunities for exploration. One of the promising solutions to address the problem is to tweak STT-MRAMs' thermal stability factor, which is related to the cell area and the thickness of the Magnetic Tunnel Junction (MTJ) used for storing data. As shown in Figure 1, the area of magnetic layers and the corresponding transistor can be reduced by optimizing the STT-MRAM's thermal stability factor, which leads to high storage capacity comparable to DRAM technologies. However, the thermal stability factor optimization is not only related to cell area/MTJ thickness, but also diverse device parameters including data retention time (non-volatility), power consumption, and reliability. As a consequence, finding the optimal values of thermal stability factor and corresponding device parameters is key to making STT-MRAM substitute main memory.

In this study, the main goal is to introduce an optimized STT-MRAM that can offer - i) higher (DRAM-like) density, ii) lower power consumption, and iii) shorter latency. To achieve this goal, we classify critical device parameters, including thermal stability factor, that make STT-MRAM highly comparable to modern DRAM technology. We then present the analytic models to optimize MTJ, leading to low operating power, better performance, yet reliable memory system. Based on our analytic model, we provide Pareto optimal points and detailed design parameter through a memory-level optimization study.

2 BACKGROUND

Magnetic Tunnel Junction (MTJ). MRAM is a magnetic storage element that uses MTJ which consists of two ferromagnets separated by a thin insulator. Typically a few nanometers apart, the insulating layer is thin enough for the electrons to tunnel from one ferro-magnet into the other, when current is applied.

One of the two plates in the MTJ, called *fixed layer*, is a permanent magnet set to a particular polarity; the other plate's, named *free layer*, field can be changed to match that of an external field to store memory. If free layer and reference layer are in parallel, MTJ has low resistance, and if they are in anti-parallel state, MTJ has high resistance. These two different levels of resistance denote value (0 or 1) of the stored data [10].

This cell structure is one of the critical factors behind the large area issue of STT-MRAM. The STT-MRAM cell area is dominated by the transistor, and the MTJ is much smaller in size. However, it is the physical characteristic of MTJ that requires the transistor to be of very large size (i.e. high W/L) to be able to drive the high switching current for writing data. The cell area would reduce significantly, if the transistor can be sized down.

Read Operation. In STT-MRAM, reading is carried out by measuring the electrical resistance of the cell. A particular cell is selected by powering an associated transistor that switches current from a supply line through the cell to ground. Due to the magnetic tunnel effect, the electrical resistance of the cell changes due to the orientation of the fields in the two plates. By measuring the resulting current, the resistance inside any particular cell can be determined, and from this the polarity of the writable plate. If the two plates have the same polarity, it is typically considered meaning a '0', while if the two plates are of opposite polarity denotes a '1'. Since the read operation is non-destructive, it involves the sensing of the cell and no write-back. Therefore, it only takes a couple of nanoseconds, and makes STT-MRAM a promising working memory.

Write Operation. STT-MRAM writes data to the cells using spin-aligned (polarized) electrons to directly torque the magnetic state. To rotate the direction of the free layer and write data, certain level of current needs to be applied to the STT-MRAM cell. If write current is higher than this critical current, free layer will rotate, and write operation is completed. In practice, the STT-MRAM write operation requires 2.6x more operating time than its read operation, which is one of the critical points to be addressed while replacing DRAM.

This writing process is what makes the current STT-MRAM cell so large for most part. If transistor in the STT-MRAM cell cannot drive current higher than the critical current (I_C), then write failure will occur [3]. Due to this reason, transistors with large W/L ratio are required for controlling the write current, and this in turn makes the cell size very large. If the critical current required to ensure a successful write can be reduced, it will allow us to use transistors with smaller W/L ratio, and will make the STT-MRAM cell much smaller.

However, the latency of the writing operation also depends on the current value. If we decrease the critical current for attaining smaller cell area, it may increase the write latency. But, we can counter this situation by manipulating the thermal stability factor.

3 DESIGN CONSIDERATIONS

To construct STT-MRAM that can replace DRAM as main memory, there are three different key considerations that need to be addressed: i)Pareto optimal point regarding cell area and MTJ thickness, ii) Reliability of write operation, and iii) Retention time, which is also related to power and latency. In this section, we first discuss the critical parameters associated to all these key considerations. We then study the analytic models regarding these parameters, and explain the scopes for optimization these models can provide.

3.1 Critical Parameters.

Thermal Stability Factor. For reliable data storage in STT-MRAM, the MTJ is designed to have two discrete, stable orientations - the parallel and the anti-parallel configuration. Thermal stability factor refers to the stability of these two states. Each of the stable orientations has an associated energy level. The energy levels of both orientations are equivalent, but there is an energy barrier to overcome when switching from one orientation to the other. The stability of a given magnetic state over time is intrinsically linked to the magnitude of the energy barrier between the two possible stable orientations of magnetization. It is critical to observe that the ability to write to a new state is also intrinsically linked to the height of this energy barrier. During the write process, energy must be supplied to change the magnetization of the free layer in the desired direction. The greater the energy barrier between the two orientations, the greater the magnitude of the energy required for the writing process will be.

This increase in writing energy being tied to the increase in stability originates the fundamental tension that exists between the factors controlling data retention time and the scalability of memory cell to smaller lithographic processes.

Critical Current (I_C). Critical current is the minimum amount current required to perform a write operation in a STT-MRAM cell. In terms of physics, the current density at which the STT-MRAM overcomes the damping and therefore magnetization motion is excited, is called the critical current density J_{c0} [1]. If the critical current can be reduced, we can achieve smaller cell area for STT-MRAM.

Retention Time. The retention time of a MTJ is a characterization of the expected time until a random bit-flip occurs and is determined by the thermal stability of the MTJ. High stability indicates the cell is unlikely to suffer from random bit-flips but makes it more difficult to write, requiring either higher currents or more time. The stability is estimated by the thermal stability factor (Δ).

By manipulating the Thermal Stability Factor and the thickness of the free layer of the MTJ, we can decrease/relax the retention time of the STT-MRAM to optimize it in terms of area, power and latency.

3.2 Analytic Model

In this section, we discuss the analytic models for optimizing STT-MRAM to be used as main memory. For proper optimization we need to integrate the factors such as critical current, thermal stability factor, and retention time in our model.

Critical Current. Critical Current I_{c0} is given as:

$$I_{c0} = (\frac{4ek_BT}{h}) \cdot \frac{\alpha}{\eta} \cdot \Delta \cdot (1 + \frac{4\pi M_eff}{2H_K}) \qquad (1)$$

Where, e = magnitude of the electron charge, k_B = Boltzmann constant, T = Temperature (Kelvin), h = Plank's Constant, α = LLGE damping constant, η = $STT - MRAM$ efficiency parameter, Δ = Thermal Stability Factor, 4 π M_{eff} = Effective demagnetization field, and H_K = Anisotropy Field Term.

To achieve DRAM-like density, we need to decrease STT-MRAM cell area, and for that we need to decrease the critical current. One can observe from Equation 1 that, other than thermal stability factor, all the components of critical current are constants or fixed value parameter. Thus we can decrease critical current by decreasing thermal stability factor.

Thermal Stability Factor. Now, the equation model for thermal stability factor is as follows [1] -

$$\Delta = \frac{E_b}{k_B T} = \frac{H_K M_S V}{2 k_B T} \tag{2}$$

where, E_b = Energy Barrier, k_B= constant, T = Temperature, H_K = Anisotropy Field Term, M_S = Saturation Magnetization, k_B= constant and V = Volume of the MTJ = $Area of the MTJ \cdot Thickness of the MTJ = A \cdot t_h$.

Here, the critical parameters are the area and the thickness of the MTJ. We can observe that, while keeping everything else constant, we can reduce thermal stability factor by reducing MTJ area. Also, by changing the thickness of the MTJ (i.e. thickness of the free layer), we can maintain same thermal stability factor for different area of the MTJ.

Data Retention Time. Analytically, data retantion time is defined as [4][6]:

$$\tau = \tau_0 \cdot exp(\frac{E_b}{k_B T}) \tag{3}$$

where, τ_0 = operating frequency.

Equation 2 and Equation 3 show that, retention time and area are related exponentially. Thus, by manipulating retention time of the STT-MRAM cell we can optimize its area.

Finally, we know that writing in the STT-MRAM cell is basically switching the free layer of the MTJ in the desired direction. [3][13] provide a equation model for the switching probability. It considers the switching probability as a function of write pulse signal time, area, thickness, temperature, and write current magnitude. The switching probability model is [13]-

$$P_{sw}(t_{pulse}) = 1 - exp(-\frac{t_{pulse}}{Duration}) \tag{4}$$

where,

$$\frac{1}{Duration} = (\frac{2}{C + \ln(\frac{\pi^2 \Delta}{4})})(\frac{\mu_B P}{em(1+P^2)})(I_{WR} - I_{c0}) \tag{5}$$

where, m = the free layer magnetic moment, μ_B = Bohr magneton constant, C = Eulers constant, P = tunneling spin polarization of ferromagnetic layers, I_{WR} = Write current, and e = magnitude of the electron charge.

Based on these analytic models, we can tweak the thermal stability factor, critical current, and retention time to reduce the area of the STT-MRAM cell by 18.4%; which leads to DRAM comparable storage capacity. We can also use this model to optimize STT-MRAM's power and latency values.

4 MEMORY LEVEL OPTIMIZATION

DRAM Baseline. For verifying the results of optimization based on the analytic models, we need to set a DRAM baseline or reference point. In this study, we set the standard 45nm process technology as our DRAM baseline. While there are some smaller process technology available for DRAM, the 45nm process is good enough to serve the purpose of this study, which is to examine the possibility of optimizing STT-MRAM to a level where it may replace DRAM as a main memory. The DRAM baseline setting [15] is based on a $6F^2$ cell structure. DRAM and STT-MRAM read and write operations are different than each other, so it is not practical to compare their read/write power directly. The write energy data was extracted from [15], and was converted to power value considering operating setup. From this procedure, we got the actual size of DRAM as $12150nm^2$ and write power as 88μW.

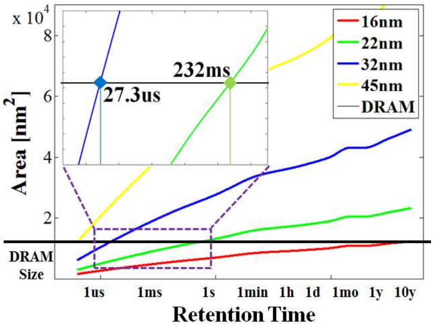


Fig. 2: Area optimization.

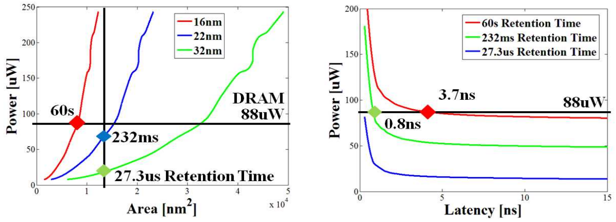
Area Optimization. The goal of area optimization is to reduce STT-MRAM cell size to set it up for designing high density main memory. First, from Equation 2 thermal stability factor was decreased by reducing MTJ's area and thickness. Then, by applying that reduced thermal stability factor into Equation 1, lowered critical current was attained. Finally, the transistor was sized-down based on this new critical current. From Equation 4, it is revealed that the trade-off factor for this process is reduction/relaxation of STT-MRAM's retention time. This whole process of area optimization was carried out for multiple retention times and CMOS process technologies (Figure 2).

From this figure we can observe that STT-MRAM cell area decreases sharply with smaller retention time. This trend verifies our analytic model, because we know that decreasing retention time reduces thermal stability factor, which in turn decreases area by reducing critical current. As the reference point, the area of a DRAM (for 45nm process) is also shown in the figure. In order to replace DRAM a main memory, STT-MRAM's cell area has to be equal to or smaller than that of DRAM. We can observe the proposed STT-MRAM successfully fulfills that criteria for the 16nm, 22nm and 32nm process for retention times of 60s , 232ms and 27.3 μs, respectively.

Power Optimization. Even though we can introduce DRAM comparable storage capacity, for replacing DRAM, STT-MRAM power consumption should also be lower than DRAM. While the read power is minute and better than DRAM, the high write power of STT-MRAM needs to be optimized for making an attempt of replacing DRAM.

By definition, power P = I^2R. In this case, P = ${I_C}^2 R_{MTJ}$. Resistance level is fixed for MTJ and we decrease power by reducing critical current. Now, critical current is proportional to cell area. Figure 3a portrays power consumed by STT-MRAM for varying cell area. It also shows DRAM area and power from [15]. Using the DRAM values as reference, we can easily locate the operating region for STT-MRAM. On a side note, for smaller process technology transistors should be able to drive same current (i.e. same W/L ratio) with smaller physical area. 16nm process transistor should drive same current as 32nm process transistor, using smaller area. Figure 3 confirms this trend as well.

Latency Optimization. Finally, the model is examined for latency optimization. In this regard, there is a key point to notice. In the earlier part of this optimization, based on Equations 1 and 3, thermal stability factor and critical current was lowered in order to reduce STT-MRAM's cell area and its power consumption for write operation. However, this lowering of the critical current has a adverse effect on the write latency of the STT-MRAM. From Equation 5 and 6, it can be seen that, using higher current shortens the required switching pulse width. This in turn leads to



(a) (b)

Fig. 3: STT-MRAM optimization: (a) Power optimization. (b) Latency optimization.

TABLE 1: Proposed STT-MRAM Implementations.

Retention Time	Latency (ns)	Power (μW)	Thermal Stability Factor
60s	4.20	86.81	27.26
232ms	2.30	59.39	19.26
27.3μs	1.00	28.59	10.21

smaller write latency. Figure 3b demonstrates the latency optimization for STT-MRAM to replace the baseline DRAM as main memory, through plotting latency vs. power for various retention times. Once again, the baseline DRAM power value is shown for reference purposes. The region where STT-MRAM latency meets/supersedes DRAM in terms of latency, is pointed out.

Potential Implementations. Comparison of our optimized STT-MRAM and reference DRAM is shown in Figure 4. Optimized STT-MRAM reduced power by 98.64%, 67.48%, and 32.48%, when MTJ retention time is 60s, 232ms, and 27.3μs respectively. Optimized STT-MRAM's latency values, normalized to that of the baseline DRAM, are 0.31, 0.17, and 0.074, for MTJ retention times of 60s, 232ms, and 27.3μs respectively.

5 RELATED WORK

The retention time of STT-MRAM can be varied by regulating different physical parameters, which can result in performance improvement in terms of latency and power consumption. Of late, researchers are exploring this concept with genuine interest. There exist many efforts to address long write latency and poor power consumption behaviors by relaxing the retention time and non-volatility. [16] proposed both L1 cache and lower level cache designs using multi-retention level STT-MRAM cache, which can significantly reduce the total energy, while improving write for both level 2 and level 3 caches. Similarly, [12] claimed that retention-relaxed STT-MRAM can replace SRAM in processor caches by reducing the high dynamic energy and slow write latencies.

Unfortunately, all these works overlook the area drawback of STT-RAM, the most critical barrier on the path of this technology to be implemented as a main memory. Unlike previous work, we looked into the problem by looking at it from a device/material level perspective and offer the Pereto optimal point to address poor power and long write latency issues behind storage-class STT-MRAM.

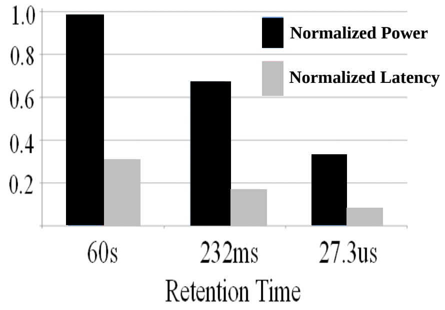


Fig. 4: Power and latency comparison - normalized to the baseline DRAM.

6 CONCLUSION AND FUTURE WORK

In this paper, we have presented a probable course of action for replacing DRAM with an optimized STT-MRAM. We have suggested methods for tuning the regular STT-MRAM, for DRAM-like density and low latency and power consumption. For this purpose, we studied the critical physical parameters of STT-MRAM, such as critical current, thermal stability factor and retention time. Based on that study we developed the analytic models for modifying the STT-MRAM for desired performance. Finally, we carried out the area, power and latency optimization for STT-MRAM, and presented the evaluation.

We have carried out our comparative evaluation between STT-MRAM and DRAM with 45nm DRAM technology as the baseline or reference point. In future, we plan to explore similar evaluation for smaller process technologies (i.e. 32nm, 22nm etc.).

REFERENCES

[1] Khvalkovskiy, A. V., et al. *Basic principles of STT-MRAM cell operation in memory arrays.* Journal of Physics D: Applied Physics 46.7 (2013): 74001-74020.

[2] Zhao, W. S., et al. *Failure and reliability analysis of STT-MRAM.* Microelectronics Reliability 52.9 (2012): 1848-1852.

[3] Li, Jing, et al. *Design paradigm for robust spin-torque transfer magnetic RAM (STT MRAM) from circuit/architecture perspective.* Very Large Scale Integration (VLSI) Systems, IEEE Transactions on 18.12 (2010): 1710-1723.

[4] Sun, Jonathan Z., et al. *Spin angular momentum transfer in a current-perpendicular spin-valve nanomagnet.* Integrated Optoelectronic Devices 2004. International Society for Optics and Photonics, 2004.

[5] Dorrance, Richard William. *Modeling and Design of STT-MRAMs.* PhD diss (2011).

[6] Rizzo, N. D., et al. *Thermally activated magnetization reversal in submicron magnetic tunnel junctions for magnetoresistive random access memory.* Applied physics letters 80.13 (2002): 2335-2337.

[7] Sun, J. Z. *Spin-current interaction with a monodomain magnetic body: A model study.* Physical Review B 62.1 (2000): 570.

[8] Raychowdhury, Arijit, et al. *Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances.* Electron Devices Meeting (IEDM), 2009 IEEE International. IEEE, 2009.

[9] Huai, Yiming. *Spin-transfer torque MRAM (STT-MRAM): Challenges and prospects.* AAPPS Bulletin 18.6 (2008): 33-40.

[10] Li, Hai, and Yiran Chen. *Nonvolatile Memory Design: Magnetic, Resistive, and Phase Change.* CRC Press, 2011.

[11] Lee, Benjamin C., et al. *Architecting phase change memory as a scalable dram alternative.* ACM SIGARCH Computer Architecture News 37.3 (2009): 2-13.

[12] Smullen, Clinton W., et al. *Relaxing non-volatility for fast and energy-efficient STT-RAM caches.* High Performance Computer Architecture (HPCA), 2011 IEEE 17th International Symposium on. IEEE, 2011.

[13] Fong, Xuanyao, et al. *Failure Mitigation Techniques for 1T-1MTJ Spin-Transfer Torque MRAM Bit-cells.* 1-1.

[14] Diao, Zhitao, et al. *Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory.* Journal of Physics: Condensed Matter 19.16 (2007): 165209.

[15] Lee, Benjamin C., et al. "Architecting phase change memory as a scalable dram alternative." ACM SIGARCH Computer Architecture News 37.3 (2009): 2-13.

[16] Sun, Zhenyu, et al. "Multi retention level STT-RAM cache designs with a dynamic refresh scheme." Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture. ACM, 2011.

[17] Kultursay, Emre, et al. "Evaluating STT-RAM as an energy-efficient main memory alternative." Performance Analysis of Systems and Software (ISPASS), 2013 IEEE International Symposium on. IEEE, 2013.

[18] Bergman, Keren, et al. "Exascale computing study: Technology challenges in achieving exascale systems." Defense Advanced Research Projects Agency Information Processing Techniques Office (DARPA IPTO), Tech. Rep 15 (2008).