An Architecture-level Cache Simulation Framework Supporting Advanced PMA STT-MRAM

Bi Wu*, Yuanqing Cheng†, Ying Wang‡, Aida Todri-Sanial¶, Guangyu Sun§, Lionel Torres¶ and Weisheng Zhao†
*School of Software Engineering
†School of Electrical and Information Engineering
Beihang University, Beijing, China, 100191
‡Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China, 100190
¶LIRMM laboratory, CNRS, Montpellier, France, 34095
§CECA, EECS, Peking University, Beijing, China, 100871

Email: yuanqing@ieee.org

***Abstract*—With integration density on-chip rocketing up, leakage power dominates the whole power budget of contemporary CMOS technology based memory, especially for SRAM based on-chip cache. To overcome the aggravating "power wall" issue, some emerging memory technologies such as STT-MRAM (Spin transfer torque magnetic RAM), PCRAM (Phase change RAM), and ReRAM(Resistive RAM) are proposed as promising candidates for next generation cache design. Although there are several existing simulation tools available for cache design, such as NVSim and CACTI, they either cannot support the most advanced PMA (Perpendicular magnetic anisotropy) STT-MRAM model or lack the ability for multi-banked large capacity cache simulation. In this paper, we propose an architecture level design framework for cache design from device level up to array structure level, which can support the most advanced PMA STT-MRAM technology. The simulation results are analyzed and compared with those produced by NVSim, which prove the correctness of our framework. The potential benefits of PMA STT-MRAM used as multi-banked L2 and L3 cache are also investigated in the paper. We believe that our framework will be helpful for computer architecture researchers to adopt the PMA STT-MRAM in on-chip cache design.**

I. Introduction

As the technology node continuously shrinks, CMOS based memory suffers from severe static power consumption challenge due to the aggravating transistor sub-threshold leakage. The standby power of cache memory occupies a large fraction of the total power budget [1]. To attack the elevated "power wall" problem, several emerging non-volatile memory (NVM) technologies, such as PCRAM [2], ReRAM [3], and Spin-Torque Transfer memory (STT-MRAM) [4], are proposed as promising candidates for the future memory architecture design because of their high density, good scalability, and ultra low leakage. Among these technologies, STT-MRAM has some unique properties like fast read/write speed and high endurance. Therefore, it is a competitive candidate of on-chip cache to replace SRAM.

MTJ is the data storage device of STT-MRAM. It is a sandwich-like structure consisting of two ferromagnetic layers and one barrier in between as shown in Fig. 1a. One ferromagnetic layer, called pinned layer, has a fixed magnetic direction. Another ferromagnetic layer is called free layer. Its magnetization direction depends on the spin polarized current direction flowing through it. When the magnetization of free layer is parallel to that of pinned layer, its resistance is low. Otherwise, it has high resistance. Therefore, depending on MTJ resistance, a '0' or '1' can be stored.

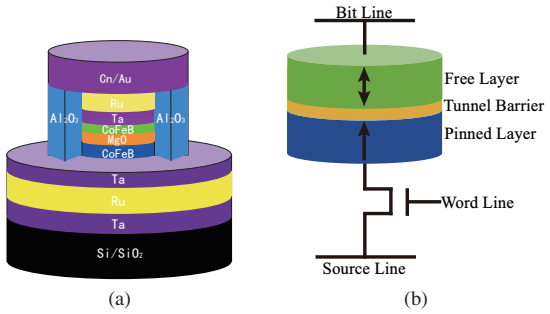

Fig. 1: (a) Perpendicular anisotropy magnetic tunneling junction structure (b) The commonly used 1T1MTJ cell structure

The most widely used STT-MRAM cell structure is 1T1MTJ (i.e., one transistor and one MTJ as shown in Fig. 1b. The wordline transistor control access to data stored in MTJ. In a write operation, a write voltage is imposed between bitline and source line to make the switching current flow through MTJ. Depending on the voltage polarity, a '0' or '1' can be written. For instance, if the current flows from bitline to source line, the MTJ magnetization will be in parallel with that of pinned-layer. Thus, '0' is written into the cell. In order to read data from memory cell, a small sensing voltage is imposed between bitline and source line, the current flowing through MTJ is compared with that flowing through the reference cell. The reference cell consists of a MTJ in parallel state and one in anti-parallel state. The sensing current difference can tell what data is stored in the cell. Since the current difference is usually very small, a sense amplifier is necessary to amplify the difference and feed the amplified signal to the next stage circuit. In the paper, we adopt the commonly used 1T1MTJ as

978-1-4673-7849-9/15/$31.00 ©2015 IEEE

our cell structure.

When MTJ feature size is above 90nm, its magnetization is within its surface plane. This type of MTJ is called in-plane MTJ [5]. As the fabrication process shrinks, high switching current density requirement from in-plane MTJ can not keep compatible with CMOS technology node. To overcome this challenge, PMA STT-MRAM is proposed. The magnetization of PMA MTJ is perpendicular to the surface and has much lower switching current requirement. Subsequently, PMA technology based STT-MRAM is more suitable for on chip cache when the feature size shrinks below 90nm. To explore the full potential of PMA STT-MRAM for cache design, it is highly desirable to develop an architecture level simulation tool for cache design, which can accurately capture PMA STT-MRAM behaviors and extensively explore the design space to obtain the optimal solution.

NVSim [6] is a well known simulator supporting emerging nonvolatile memory design. It can optimize cache design with some specific optimization targets, e.g., dynamic power consumption, access time, area, leakage, etc. However, to accommodate other available nonvolatile memory technologies such as PCRAM and ReRAM, it has to abstract the common properties of all emerging nonvolatile technologies and make some simplifications on STT-MRAM behaviors. In addition, NVSim lacks the support of sub-40nm most advanced PMA STT-MRAM technology. Moreover, it can only simulate a single bank cache. With the development of many core processor, it expects that large capacity cache will be integrated on-chip, which is usually organized as multi-bank structure. As a result, the capability of simulating multi-banked large capacity cache is highly desirable. Another cache simulation tool CACTI [7] supports multi-banked cache design and includes the network on chip model to estimate communication congestions between different banks. Unfortunately, it can not support emerging nonvolatile memory technology, and can not be used for STT-MRAM cache simulation.

In this paper, we propose an architecture level simulation framework which can facilitate PMA STT-MRAM design. Our main contributions are as follows,

- Supporting most advanced PMA MTJ models and multi-banked cache structure in our cache simulation framework,
- Incooperating corresponding read/write circuits and reference cell structure for cache array modeling,
- Exploring the potential benefits of using PMA STT-MRAM as large capacity cache including L2 and LLC based on our established simulation framework.

The rest of the paper is organized as follows. Section II presents related work and motivates our research work through analyzing drawbacks of several mainstream simulators. Section III presents our PMA STT-MRAM simulation framework from device model up to array level modeling. Experimental results compared with those of NVSim and analysis are given in Section IV. Section V concludes this paper.

II. Related Work & Motivation

With technology node continuously shrinking, leakage power occupies a large fraction of chip power consumption. This problem is much more severe for large capacity on chip cache. To reduce standby power, some non-volatile emerging memory technologies are proposed to replace SRAM or DRAM based memory, such as PCRAM, ReRAM and STT-MRAM. Among them, STT-MRAM is compatible with CMOS process technology and comparable access speed with SRAM. Therefore, it attracts much attention from both academia and industry.

At device level, Hosomi et al. proposed to use spin-RAM as a novel nonvolatile memory [8]. After that, Oh et al. explored novel MTJ structure that can approach Gb integration density [9]. Kim et al. evaluated the scalability of PMA STT-MRAM towards sub-20nm regime [10]. There are also many research efforts on circuit and architecture level design. Wang et al. proposed a new STT-MRAM cell structure to improve cell sensing reliability [11]. Many researchers investigate the possibility for replacing on chip cache by MRAM and proposed many effective techniques to exploit MRAM's full potential [12] [13].

As mentioned above, although STT-MRAM has promising prospects for building cache due to its high density, fast access speed, nonvolatile property, it still suffers from high write power and long write latency. All these obstacles lead to many difficulties when using STT-MRAM in the practical use for computer architecture [14] [15]. To deal with these challenges, both academia and industry are working on optimize STT-MRAM from device fabrication to architecture design, such that it can replace the current SRAM based cache successfully. At circuit level, there are some mature simulators like Cadence Spectre, which can support MTJ model simulation. However, it can not be used for cache design space exploration due to time consuming simulation procedure. At system level, there are also some well known simulators like NVSim and CACTI. But as we mentioned before, both of them have their own shortcomings. NVSim includes several emerging nonvolatile memory technologies and does not optimize specifically for STT-MRAM especially for the most advanced PMA STT-MRAM. For instance, in NVSim, all nonvolatile memory technology based cache share the same sense amplifier architecture and write circuits. However, the read/write circuit of PMA STT-MRAM is different from those for PCRAM and ReRAM due to its unique access requirements. Another shortage is NVSim can only simulate the single bank structure, no matter how large the cache capacity is. This also does not fit contemporary large capacity cache design, which usually splits cache into multiple banks to reduce access latency.

CACTI is another widely used cache simulator, which can support both SRAM and eDRAM multi-bank simulation. The latest version, CACTI6.5 has improved in many aspects. First, it has changed from simple linear scaling based on original 0.8 micron process to using technology parameters based on the ITRS roadmap. Second, it adds the support for DRAM memory. Third, it includes more wire models for inter-bank network communication modeling, more reasonable repeater sizing and spacing for delay optimization, and low-swing differential wires signal considering low power design. Moreover, it can support both NUCA (Non-Uniform Cache Architecture) and UCA (Uniform Cache Architecture). Finally, its empirical network contention model can estimate the impact of network congestion on bank access cycle time and power

consumption. The CACTI simulation results are verified by real chip cache parameters. Unfortunately, it can not support the emerging technologies such as PCRAM, STT-MRAM, etc. In this paper, we establish our STT-MRAM simulation framework based on CACTI to ensure the simulation accuracy.

III. Our Proposed Cache Simulation Framework

A. Device level modeling

Our work is based on a compact model characterizing 40nm MTJ electromagnetic behaviors [16]. In the model, MTJ switching current can be calculated by the following formula,

$$I_{c0} = \alpha \frac{\gamma^e}{\mu_B g}(\mu_0 M_s) H_k V \tag{1}$$

The thermal stability of MTJ can be derived by

$$E = \frac{\mu_0 M_s \times V \times H_k}{2} \tag{2}$$

See Table I for the symbol denotations used in above two formulas. Comparing with experimental results, the compact model can capture PMA MTJ electrical behavior accurately according to [16].

TABLE I: Parameters used in PMA MTJ compact model [16]

Symbol	Definition	Value
ϕ	Potential barrier height of MgO	0.4
H_k	Anisotropy field	$113.0 \times 10^3 A/m$
M_s	Saturation magnetization	$456.0 \times 10^3 A/m$
t_{ox}	Oxide barrier height	0.85nm
μ_0	Permeability in free space	$1.2566 \times 10^{-6} H/m$
μ_B	Bohr magneton	$9.274 \times 10^{-24} J/T$
γ	Gyromagnetic ratio	$1.76 \times 107 rad/(s \cdot T)$
g	Spin polarization efficiency factor	Depending on technology node
e	Electron charge	$1.6 \times 10^{-19} C$
α	Magnetic damping constant	0.027
F	Material dependent constant(for R_p)	664
RA	MTJ resistance area product	$5\Omega \cdot \mu m^2$

The above MTJ model is used to construct our 1T1MTJ cell structure.

TABLE II: Parameters of 1T1MTJ cell

configure	Cell Area	Cell Aspect Ratio	TMR	Access CMOS Width
Index	$21F^2$	2.3	200%	3F

In the real 1T1MTJ cell production process, MTJ is usually placed over the access transistor. And the area of MTJ is smaller than that of access transistor. So in our framework we use access transistor area to represent total area of 1T1MTJ cell. To calculate the area of access transistor, we use the area model in [12]. Fig. 2 is sketch map of that model.

We can get the cell area with the equation below:

$$Area_{access-transistor} = Area_{Diffusion} + Area_{Gate} \tag{3}$$

$$AspectRatio = \frac{L_{Diffusion} + L_{Gate}}{W_{Gate}} \tag{4}$$

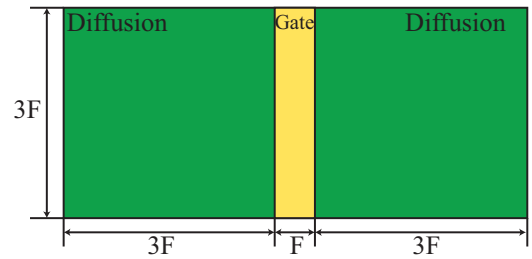

Fig. 2: Area model of 1T1MTJ cell

B. Peripheral circuit modeling

1) Read circuits modeling: To read data in SRAM cell, it requires to precharge the bitline voltage to a middle value between zero and V_{dd}. Then, depending on current magnitude difference on the two bitlines, data can be read out by the sense amplifier. Instead of using differential signal sensing scheme as in SRAM, STT-MRAM compares the sensing current with that flowing through reference cell. Therefore, the sensing margin is smaller compared to SRAM due to the small resistance between parallel and antiparallel state. Therefore, it requires a more accurate sense amplifier to sense data out reliably.

The sensing circuit used in our framework is shown in Fig. 3. It consists of two parts: reference generator circuit [17] and precharge sensing amplifier circuit(PCSA) [18]. The former part is to convert current difference to voltage difference. The latter part amplifies this voltage difference to full swing signal that can be recognized by next stage circuitry.

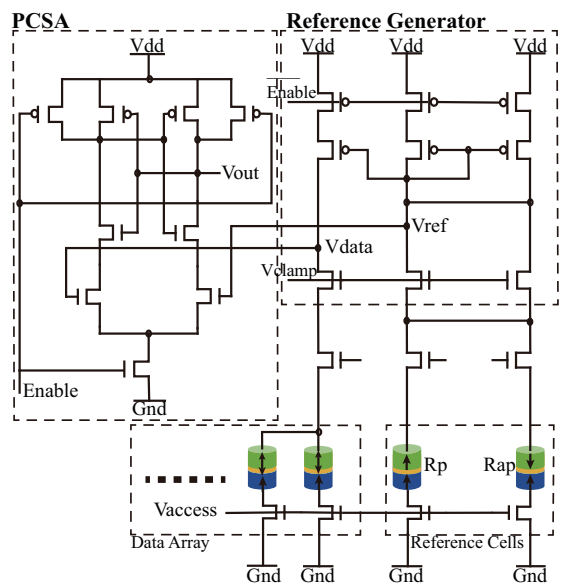

Fig. 3: Reference generator and sense amplifier circuitry [17] [18]

As in CACTI, one sensing circuitry can be shared among multiple cell columns depending on the multiplexing share

degree. To account for the area overhead of reference cells, we assume that two columns of reference cells are inserted per subarray. The sense amplifier area overhead can be calculated as follow:

$$AREA_{sa} = AREA_{sense-circuit} + 2 \times AREA_{ref-column} \quad (5)$$

In order to calculate read power, latency and leakage of a single cell, we use Cadence Spectre to simulate the hybrid CMOS/MTJ sensing circuit. Since the read latency and power of read ‘ 1’ are larger than those when read ‘0’, we use the data of reading ‘1’ to cover the two cases.

2) Write circuit modeling: CACTI has no specific write circuit because of the structure of SRAM cell. Read operation shares the same circuitry with write. However, the sensing voltage is much smaller than writing voltage. Therefore, it requires a specific write circuit to produce the switching current in order to change the magnetization of free layer.

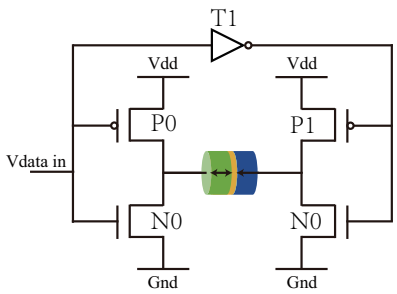

Fig. 4: PMA STT-MRAM write circuitry

Fig. 4 shows the write circuitry used in our framework. The inverter T1 translates V_{data_in} into two opposite states to control transistors establishing corresponding write path. For example, if V_{data_in} is ‘1’, then N0, P1 will turn on. A parallel state can be written into MTJ. The write behavior simulated by Cadence Spectre is listed in Table III. Because write ‘1’ state can cover the worst case of write operation, so we use write ‘1’ state parameters as the write operation paramters.

TABLE III: Simulation results of read and write circuits in PMA STT-MRAM by Cadence Spectre

Parameter	T_{sense0}(ns)	T_{sense1}(ns)	T_{write0}(ns)	T_{write1}(ns)
Result	0.088	0.053	2.7	5.8
Parameter	P_{sense0}(pJ)	P_{sense1}(pJ)	P_{write0}(pJ)	P_{write1}(pJ)
Result	0.0054	0.0035	0.196	0.348

T_{sense0} (or T_{sense1}) is the delay to read data ‘0’ (or ‘1’) out from memory cell, T_{write0} (or T_{write1}) is the delay to write data ‘0’ (or ‘1’) into memory cell, P_{sense0} (or p_{sense1}) is the power of reading data ‘0’ (or ‘1’) out of memory cell, P_{write0} (or p_{write1}) is the power of writing data ‘0’ (or ‘1’) into memory cell.

C. Array level modeling

Considering the above mentioned differences between S-RAM and PMA STT-MRAM cache modeling, our framework can be presented as in Fig. 5.

As shown in the figure, the input MTJ parameters are obtained from hybrid CMOS/MTJ circuit simulation. The transistor parameters are derived from ITRS technology report. Then, those device and circuit level data are used for array level parameter calculations.

At cell level calculation, we reduce the bitline number to one for STT-MRAM. SRAM based cell structure using differential sensing scheme requires two bitlines in each memory cell. STT-MRAM has only one bitline, which affects the bitline parasitics calculations. We calculate the power and delay of STT-MRAM mat based on single bitline cell structure.

At mat and subarray level, we add the delay and power of sense and write circuits obtained from circuit simulation. The sense amplifier area overhead is calculated based on the new write and sensing circuitry. Moreover, two columns of reference cells are added in each subarray. Another difference of STT-MRAM subarray structure from SRAM structure is that the former one does not need bitline precharge and bitline restore circuits due to its unique sensing mechanism. Thus, we assign the bitline precharge values(power, delay) and bitline restore values(delay, power) to 0. Additionally, we separate read and write path in STT-MRAM modeling and consider its impact on power, latency and area calculations. In contrast, CACTI does not distinguish the differences between cell read and cell write operation.

IV. Experimental Results

A. Verification of our proposed framework

To verify the framework, we simulate 2MB cache with 40nm feature size. The cache configuration is tabulated in Table IV.

TABLE IV: Cache configuration parameters for our framework validation

configure	Capacity	Block size	Associativity	Bank
Index	2M	64bytes	1	1
configure	Access mode	Interconnect	ECC	Design objective
Index	Sequential	Conservative	None	Delay

Using the configuration above, the simulation results obtained from NVSim and our simulator is shown in Table V. The optimization objective is delay.

TABLE V: The comparative simulation result between NVSim and This work

Parameter	NVSim	This work	Percentage(%)
Read Power(nJ)	0.782	0.452	66.3
Read latency(ns)	1.272	1.380	-8.0
Write Power(nJ)	0.856	0.954	-11.4
Write latency(ns)	10.345	6.888	50.1
Area(mm^2)	2.585	3.77895	-46.1

As shown in the table, there are some differences in the results produced by NVSim and our simulator. In the next section, we will analyze the difference to prove the correctness of our simulator.

First, we can observe that the read power obtained by NVSim is 66.3% higher than that of our simulator. The difference is caused by the calculation method of read operation. In

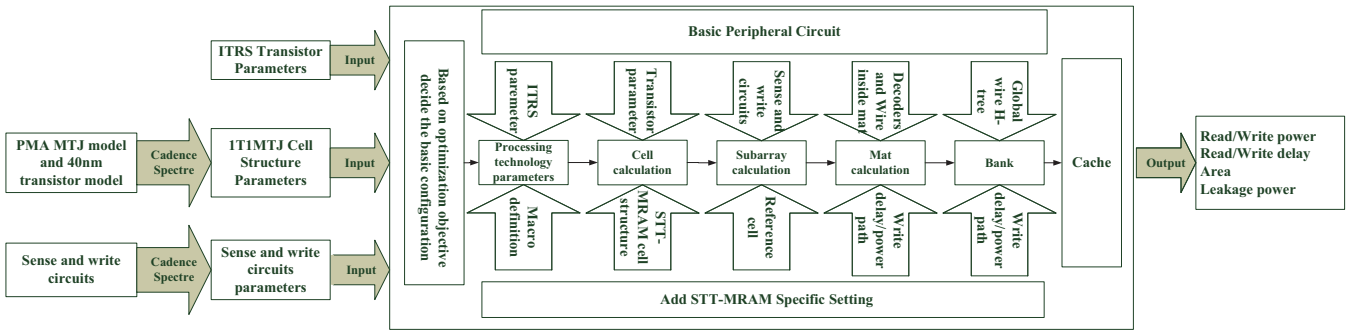

Fig. 5: The schematic of our proposed PMA STT-MRAM cache design simulation framework

NVSim, it separates the read process in two parts. The first part is the cell read power, and the second part is sensing power. However, when it calculates the total read power, it adds the sensing power twice leading to the read power much higher than the results of our simulator. As for the writing power, we can observe that there is only small difference between the two (about 10%), meaning that our modeling method coincides with NVSim. Additionally, refering to read/write power values produced by NVSim, the difference is only about 9%, which is not reasonable according to our circuit level simulation results as shown in Table V.

Next, taking read/write latency into account, the write latency estimated by NVSim is 50% higher than our simulation results. In our work, we use Cadence Spectre to simulate the read and write latencies with advanced read/write circuit shown in Fig. 3 and Fig. 4 respectively. Then, the simulation data are used for higher mat and sub-array level modeling. However, we found that under 40nm feature size the sense amplifier and write circuit delays are fixed to 0.805ns and 10ns respectively. The relative larger delay values of NVSim indicates that the MTJ model is in-plane STT-MRAM model, which has distinct characteristics compared with PMA STT-MRAM.

At last, considering area calculation, 2MB cache area estimation made by NVSim is 46.1% less than that of our simulation framework. This because we add two column reference cells in each subarray which incurs a non-negligible area overhead. Although NVSim also takes reference cell into account, we find that they only consider reference cells during power sense amplifier energy calculation. The reference area overhead is not included in cache area estimation. Another source of our extra area overhead comes from the sense amplifier. We use reference generator circuit(shown in Fig. 3) to convert currents to voltages. This part requires large PMOS transistors to ensure sensing reliability, which also introduces extra area occupation. From above analysis, our proposed simulation framework can obtain more accurate simulation results for PMA STT-MRAM cache.

TABLE VI: Different bank amount under same condition simulation

Parameter	Read latency(ns)	Write latency(ns)	Leakage power(mW)
Single-bank	3.02	10.64	5925
16-bank	2.57	7.48	4208

As multi-core or many core architecture becomes more common for high performance computing, the capacity of on-chip cache is also expected to increase more rapidly. To reduce access time, large capacity cache is usually organized in multiple banks and interconnected by network on chip. In the experiments, we first use 32MB 16-way UCA cache under 40nm technology node for simulation. In order to highlight the importance of supporting multi-bank cache simulation, which can not be supported by NVSim, we first make the cache be only a single bank, and get the simulation results from NVSim simulation. Then, we split 32MB cache into 16 banks, and obtain simulation results by our simulator. The comparisons are shown in Table VI. We can observe that results of multi bank organization vary largely from those of single bank organization. The leakage power of the latter case is much lower than the former case, which indicates the benefits of multi-bank in term of power consumption. Additionally, note the multi-bank access time is lower than that of single-bank. And each bank can be accessed in parallel. Therefore, the throughput of 16-bank organized cache is much higher than that of single bank case.

B. Exploration of the potential to use PMA STT-MRAM as large capacity on-chip cache

In the following experiments, we will use our proposed simulation framework to evaluate the benefits brought by PMA STT-MRAM, especially for constructing large capacity L2 or LLC cache. The simulation results are compared with those of SRAM cache under the same configuration. The configuration parameters are listed in the Table VII.

TABLE VII: Configuration parameters of L2 and L3 cache for simulations

Parameter	capacity	associa	block size	bank
L2 cache	1MB	8 way	64 bytes	1
L3 cache	8MB	8 way	64 bytes	4

The memory cell parameters used in the cache design simulation are obtained by Cadence Spectre with 40nm PMA STT-MRAM model. Read latency for 1T1MTJ cell is 0.088ns. The write latency is 5.8ns. The read power is 3.5pJ. The write power is 348pJ. With the above parameters, we can get the simulation results when using PMA STT-MRAM as L2 and L3 cache as shown in Table VIII and Table IX.

TABLE VIII: L2 cache simulation result

Result	Read latency(ns)	Write latency(ns)	Area(mm^2)
STT-MRAM	1.37	6.81	2.34
SRAM	1.56	1.56	4.72
Result	Read power(nJ)	Write power(nJ)	Leakage power(mW)
STT-MRAM	0.192	0.953	98.3
SRAM	0.3	0.3	1223

TABLE IX: L3 cache simulation result

Result	Read latency(ns)	Write latency(ns)	Area(mm^2)
STT-MRAM	1.96	7.00	21.56
SRAM	2.33	2.33	38.04
Result	Read power(nJ)	Write power(nJ)	Leakage power(mW)
STT-MRAM	0.486	1.794	880
SRAM	0.733	0.733	10168

In L2 cache we just use single bank configuration because of relatively low capacity of L2 cache. As shown from the results, STT-MRAM is better than SRAM in terms of read latency/power, area and leakage. But the write power and latency of PMA STT-MRAM are larger than those of SRAM. This is because MTJ needs larger current and longer time to change the magnetization of free layer.

For the area, traditional SRAM is composed of 6 transistors and STT-RAM 1T1MTJ cell just occupies about one transistor area. This leads to significant area benefits of STT-MRAM. Although reference generator circuit used in our STT-MRAM introduces some area overheads, STT-MRAM still has large area benefits after considering this factor.

Another promising part when replacing SRAM with PMA STT-MRAM is the leakage power. PMA STT-MRAM is nonvolatile, which is different from SRAM implicating that STT-MRAM doesn't need extra standby power to keep the data. This benefit paves the way for ultra low leakage power cache design.

From above simulation results, it indicates that PMA STT-MRAM can achieve better read performance, small area, very low leakage power except for high write latency and energy. Consequently, STT-MRAM is more suitable for L3 cache which usually has largest capacity in the whole cache hierarchy. Since it is further from the processor compared to L1, L2 cache and usually has much lower write activity, the relatively slow write operation can be tolerated.

V. Conclusion

As cache capacity increases rapidly in morden multi-core and many-core processors, power consumption becomes a bottleneck for large capacity cache design. As a promsing candiate, PMA STT-MRAM has negligible leakage power and comparable access speed with SRAM. Therefore, it is highly desirable to build an architecture level cache simulation tool supporting PMA STT-MRAM. In this paper, we propose an architecture level cache simulation framework incooperating most advanced 40nm PMA STT-MRAM technology. We buid the framework from device model to array level considering the unique requirements on read/write circuitry, sub-array design issues of PMA STT-MRAM. The simulation results are firstly compared with those obtained by NVSim to show the effectiveness of our simulation framework. Then, we use it to evaluate the potential of PMA STT-MRAM as on chip cache. By comparing with SRAM, we conclude that PMA STT-MRAM is much more suitable to work as L3, which requires large capacity and can tolerate longer write latency.

Acknowledgment

The work is sponsored by China NSFC fund No. 61401008 and Beijing NSF fund No. 4154076.

References

[1] J. Wang et al., "OAP: An obstruction-aware cache management policy for STT-RAM last-level caches," in *Proc. of Design, Automation Test in Europe Conference Exhibition (DATE)*, 2013, pp. 847–852.

[2] S. Raoux et al., "Phase-change random access memory: A scalable technology," *IBM Journal of Research and Development*, vol. 52, no. 4.5, pp. 465–479, 2008.

[3] S. Gaba et al., "Memristive devices for stochastic computing," in *Proc. of IEEE International Symposium onCircuits and Systems (ISCAS)*, 2014, pp. 2592–2595.

[4] C. Chappert et al., "The emergence of spin electronics in data storage," *Nature Materials*, vol. 6, no. 11, pp. 813–823, 2007.

[5] S. H. Kang et al., "Emerging materials and devices in spintronic integrated circuits for energy-smart mobile computing and connectivity," *Acta Materialia*, vol. 61, no. 3, pp. 952–973, 2013.

[6] X. Dong et al., "NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 7, pp. 994–1007, 2012.

[7] N. Muralimanohar et al., "CACTI 6.0: A tool to understand large caches," *University of Utah and Hewlett Packard Laboratories, Tech. Rep*, 2009.

[8] M. Hosomi et al., "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM," in *Proc. of IEEE International Electron Devices Meeting (IEDM)*, 2005, pp. 459–462.

[9] S. C. Oh et al., "On-axis scheme and novel MTJ structure for sub-30nm Gb density STT-MRAM," in *Proc. of IEEE International Electron Devices Meeting (IEDM)*, 2010, pp. 12.6.1–12.6.4.

[10] W. Kim et al., "Extended scalability of perpendicular STT-MRAM towards sub-20nm MTJ node," in *Proc. of IEEE International Electron Devices Meeting (IEDM)*, 2011, pp. 24.1.1 – 24.1.4.

[11] W. Kang et al., "DFSTT-MRAM: Dual Functional STT-MRAM Cell Structure for Reliability Enhancement and 3-D MLC Functionality," *IEEE Transactions on Magnetics*, vol. 50, no. 6, article no.3400207, 2014.

[12] G. Sun et al., "A novel architecture of the 3D stacked MRAM L2 cache for CMPs," in *Proc. of IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2009, pp. 239–249.

[13] X. Zhang et al., "Exploring potentials of perpendicular magnetic anisotropy STT-MRAM for cache design," in *Proc. of IEEE International Solid-State and Integrated Circuit Technology (ICSICT)*, 2014.

[14] M. Chang et al., "Technology Comparison for Large Last-Level Caches (L^3Cs): Low-Leakage SRAM, Low Write-Energy STT-RAM, and Refresh-Optimized eDRAM," in *Proc. of IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2013, pp. 143–154.

[15] J. Ahn et al., "Lower-bits cache for low power STT-RAM caches," in *Proc. of IEEE International Symposium on Circuits and Systems (ISCAS)*, 2012, pp. 480–483.

[16] Y. Wang et al., "Compact model of magnetic tunnel junction with stochastic spin transfer torque switching for reliability analyses," *Microelectronics Reliability*, vol. 54, no. 9-10, pp. 1774–1778, 2014.

[17] J. Kim et al., "A novel sensing circuit for deep submicron spin transfer torque MRAM (STT-MRAM)," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 1, pp. 181–186, 2012.

[18] W. Zhao et al., "High speed, high stability and low power sensing amplifier for MTJ/CMOS hybrid logic circuits," *IEEE Transactions on Magnetics*, vol. 45, no. 10, pp. 3784–3787, 2009.