Evaluation of Spin-Hall-assisted STT-MRAM for Cache Replacement

Liang Chang, Zhaohao Wang*, Yuqian Gao, Wang Kang, Youguang Zhang, and Weisheng Zhao*
Department of Electronic and Information Engineering, Beihang University
Fert Beijing Research Institute, Beihang University, Beijing, China 100191
*Email: zhaohao.wang@buaa.edu.cn and weisheng.zhao@buaa.edu.cn

***Abstract*—Emerging spin orbit torque (SOT) promises to achieve high-speed write operation for magnetoresistive random access memory (MRAM) since it can eliminate the incubation delay of the conventional spin transfer torque (STT). Such a speed improvement allows the MRAM to be used as low-level cache in the computer architecture. Among various SOT technologies, spin-Hall-assisted STT is a potential candidate thanks to its magnetic-field-free benefit. In this work, we evaluate the potential of the spin-Hall-assisted STT-MRAM in the cache replacement. Firstly, the bit-cell parameters are obtained from the circuit-level simulation. Then, the cache evaluation and system-level simulation are performed with NVSim and Gem5 simulators. Simulation results validate the advantage of the spin-Hall-assisted STT in the write speed and energy. Moreover, for the large capacity (about >1 MB), the spin-Hall-assisted STT-MRAM is a competitive candidate for replacing the conventional SRAM-based cache.**

Keywords—spin orbit torque (SOT); spin transfer torque (STT); magnetoresistive random access memory (MRAM); cache replacement; NVSim; GEM5

I. INTRODUCTION

With the continuous scaling of the conventional CMOS technology, the mainstream memory architecture of the computing system suffers from two major challenges: on the one hand, the increasing leakage current of CMOS transistors leads to soaring static power consumption [1], especially in static and dynamic random access memories (SRAM and DRAM); on the other hand, the interconnection between the processor and memory causes a relative large access latency. As one of the solutions, emerging non-volatile memories (NVMs) are considered promising candidates for constructing the next-generation RAM, since they can avoid the static power by adopting the sleep mode and decrease the interconnection latency by integrating them above the CMOS circuits. Among various NVMs, spin transfer torque-based magnetoresistive RAM (STT-MRAM) attracts lots of research interest thanks to its high speed, low power, infinite endurance, and CMOS compatibility [2]–[4]. Up to now, abundant STT-MRAM chips have been demonstrated by industry and academia, including the test chip and commercial products [5]–[6].

Although the STT-MRAM shows the great application potential in building the high-performance memory architecture, its write speed is limited by an intrinsic incubation delay of the STT. Such a speed bottleneck makes it difficult to use the STT-MRAM as low-level cache. Recently, spin-orbit torque (SOT) has been proposed to provide high-speed and low-energy write operation for the MRAM [7]–[11]. Nevertheless, for the SOT-MRAM using perpendicular magnetic anisotropy (PMA), an additional magnetic field is required to achieve the deterministic switching. Although the magnetic field can be generated by the antiferromagnetic layer integrated into the device, it lowers the retention stability. In this context, we proposed an alternative write technology called spin-Hall-assisted STT for the MRAM [12]. With this technology, the incubation delay can be eliminated by the SHE-induced SOT. Therefore, the fast, reliable and magnetic-field-free write operation can be obtained.

In this paper, to evaluate the application potential of the spin-Hall-assisted STT-MRAM in the computer architecture, we present a device-to-system study of the spin-Hall-assisted STT-MRAM aiming to the cache replacement. The rest of this paper is organized as follows: in Section II we introduce the preliminary knowledge about the MRAM, STT and SOT. In section III, the single bit-cell performance of the spin-Hall-assisted STT-MRAM is evaluated based on a compact device model and STMicroelectronics CMOS 28 nm design kit. Then, the experimental methodology for the cache evaluation is presented in section IV. Afterwards, the cache parameters and the system-level performance are obtained with the NVSim and Gem5 simulator in section V. Finally, we summarize this work with a conclusion.

II. BACKGROUND AND BASICS

A. Magnetic tunnel junction and spin transfer torque

The basic cell of the MRAM is magnetic tunnel junction (MTJ) [13], whose core part is composed of an oxide barrier sandwiched between two ferromagnetic (FM) layers, as shown in Fig. 1(a). One FM layer is called reference layer (RL) or pinned layer (PL) which has a stable magnetization. The other

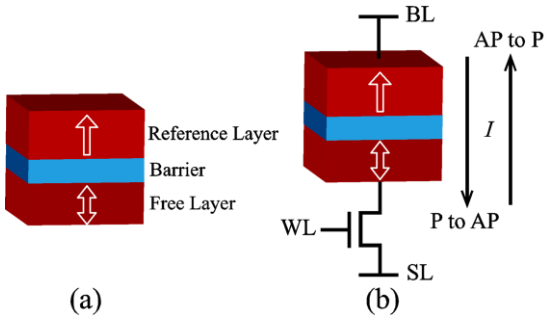

Fig. 1. (a) Core structure of the MTJ. (b) one-transistor-one-MTJ structure switched by STT. Only a bidirectional current is required to achieve the write operation.

FM layer is free layer (FL), whose magnetization direction can be switched to be parallel or antiparallel (P or AP) to that of the RL. The P or AP state of the relative magnetization orientation corresponds to the low or high tunnel resistance of the MTJ, which is tunnel magnetoresistance (TMR) effect. This effect enables the MTJ to storage 1-bit non-volatile data and to constitute the MRAM.

The write technology of the MTJ is a research focus as it directly influences the essential performance of the MRAM. Currently, the STT is widely used as the mainstream write approach for the MRAM [14]–[15]. The STT is induced by a large enough current flowing through the MTJ. Depending on the current direction, the magnetization of the FL can be switched between P and AP states, as shown in Fig. 1(b). Compared to the early field-induced magnetization switching (FIMS), the STT provide lower write current (1~10 MA/cm^2) and better scalability. However, its switching speed is limited by an intrinsic incubation delay, as explained below.

As Eq. (1), the strength of the STT is proportional to the cross product of the magnetizations of the FL and RL. During the initial stage of the switching process, these two magnetizations are nearly collinear. It is the thermal fluctuation that causes a small angle between them. Therefore, the initial STT is relative weak, leading to an incubation delay. Increasing the write current can lower the incubation delay, but add the risk of the barrier breakdown. Incubation delay is a vital factor prohibiting the application of the STT-MRAM in low-level cache.

$$\boldsymbol{\tau}_{STT} = J\xi\mathbf{m}\times\left(\mathbf{m}_r\times\mathbf{m}\right) \tag{1}$$

where J is the current density, ξ is the device-dependent parameter, $\boldsymbol{m}$ and $\boldsymbol{m}_r$ are the unit vectors along the magnetization of the FL and RL, respectively.

B. Spin orbit torque

Recently the SOT has been proposed to overcome the drawbacks of the STT-MRAM [7]–[10]. The structure of the SOT-based MTJ is shown in Fig. 2(a), where the FL is contacted to a heavy metal stripe. Due to the strong spin orbit coupling, an in-plane charge current passing the heavy metal stripe can induce Rashba effect or SHE. The Rashba effect produces an efficient in-plane magnetic field, and the SHE generates the pure spin current along the perpendicular direction. Hence both of them can provide torques (i.e. SOT) used for the magnetization switching of the FL.

To obtain higher thermal stability and better scalability, the PMA-based MTJ (p-MTJ) is preferred to the in-plane-

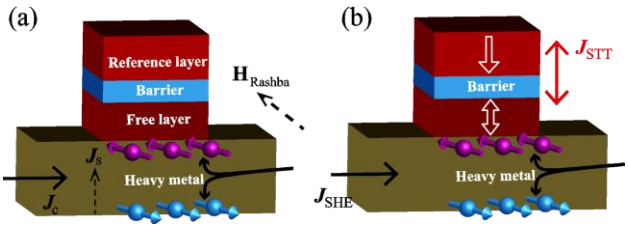

Fig. 2. (a) Three-terminal MTJ switched by spin orbit torque, which originates from SHE or Rashba effect. (b) Schematic of Spin-Hall-assisted STT.

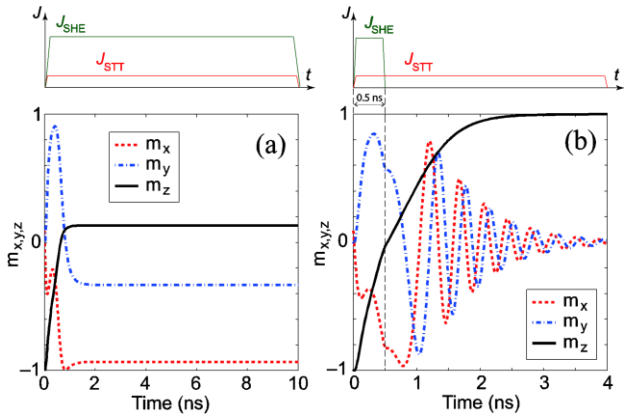

Fig. 3. Typical simulation results of spin-Hall-assisted STT: (a) Under the action of a large enough SHE write current and an STT write current, the incubation delay of the magnetization dynamics is eliminated, but the final magnetization is not switched to the perpendicular axis. (b) If the SHE write current is removed at 0.5 ns, the STT can continue to achieve the complete switching.

anisotropy MTJ. For the p-MTJ, both the efficient magnetic field induced by Rashba effect and the polarization direction of the pure spin current induced by SHE are nearly vertical to the easy-axis of the MTJ, thus the initial SOT is stronger than the conventional STT. The incubation delay can be eliminated, and high switching speed can be obtained. However, to achieve the deterministic switching of the p-MTJ, the SOT has to be used in combination with an additional in-plane magnetic field. In order to avoid the use of the magnetic field, we propose the spin-Hall-assisted STT [12].

C. Spin-Hall-assisted STT

The schematic of the spin-Hall-assisted STT is shown in Fig. 2(b), where two currents are required to achieve the write operation of the MTJ. The current passing the heavy metal is responsible for inducing the SHE, and the other current flowing through the MTJ is used for generating the STT. To produce the significant SHE, the SHE write current density needs to be large enough. The magnetization dynamics of the spin-Hall-assisted STT can be described by the Landau-Lifshitz-Gilbert (LLG) equation, as Eq. (2). Typical simulation results are shown in Fig. 3. As can be seen, thanks to the assistance of the SHE, the incubation delay is eliminated, as explained above. But the combined effect of the SHE and STT makes the final magnetization stabilized at a direction between the in-plane and perpendicular axis (see Fig. 3(a)). For switching the magnetization to the desired perpendicular axis, the SHE write current must be removed at an appropriate time (~ 0.5 ns), then the STT continues to achieve the remaining switching process, as Fig. 3(b).

$$\frac{\partial\mathbf{m}}{\partial t} = -\gamma\mu_0\mathbf{m}\times\mathbf{H}_{eff} + \alpha\mathbf{m}\times\frac{\partial\mathbf{m}}{\partial t} - \xi PJ_{STT}\mathbf{m}\times\left(\mathbf{m}\times\mathbf{m}_r\right) - \xi\eta_{SH}J_{SHE}\mathbf{m}\times\left(\mathbf{m}\times\boldsymbol{\sigma}_{SHE}\right) \tag{2}$$

where $\mathbf{H}_{eff}$ is the effective field, $\boldsymbol{\sigma}_{SHE}$ is the unit vector along the polarization direction of the SHE-induced spin current. γ is gyromagnetic ratio, μ_0 is the vacuum permeability, P is the spin polarization of the tunnel current, η_{SH} is the spin Hall

angle, J_{STT} and J_{SHE} are the STT and SHE write current densities, respectively.

III. Device Model and Memory Design

In this section, the spin-Hall-assisted STT is used for designing the MRAM. The circuit simulation is performed with a compact model of the spin-Hall-assisted STT-MTJ and STMicroelectronics CMOS 28 nm design kit. The parameters of the bit-cell are obtained based on the simulation results.

A. Compact model of the spin-Hall-assisted STT-MTJ

In order to quantitatively describe the electrical behaviors of the spin-Hall-assisted STT-MTJ, we developed a compact model which can run on standard circuit simulator (e.g. Spectre). This model is proposed based on the related physical theories, including two main modules: the magnetization dynamics is described by Eq. (2); the tunnel resistance can be calculated by Brinkman model and Slonczewski model [12].

Using the developed model and STMicroelectronics CMOS 28 nm design kit, we designed and simulated the spin-Hall-assisted STT-MRAM equipped with read/write peripheral circuits. The parameters of a single bit-cell are obtained from the simulation results.

B. Circuit-level analysis of the spin-Hall-assisted STT-MRAM

The overall architecture of the spin-Hall-assisted STT-MRAM is shown in Fig. 4(a). As can be seen, each three-terminal MTJ is connected with two access transistors, forming a 2T1R bit-cell [16]. The write circuit includes four driving transistors, which are controlled by logic signals in order to produce the bidirectional STT write current and unidirectional SHE write current, as Fig. 4(b). The read circuit is implemented with a pre-charge sensing amplifier (PCSA) [17]. Note that the direction of the SHE write current is unique as it has no deterministic influence on the magnetization switching.

For the comparison, we also designed a conventional STT-MRAM. The bit-cell is composed of one transistor and one MTJ (1T1R), as Fig. 1(b). The write circuit is shown in Fig. 4(c), where the four driving transistors are used for generating the bidirectional STT write current. The sensing amplifier is the same as that of the spin-Hall-assisted STT-MRAM.

The read/write operation of the above two MRAMs is validated by the transient simulation. The simulation results of a bit-cell in the spin-Hall-assisted STT-MRAM are shown in Fig. 5. The main parameters are configured as Table I. As can

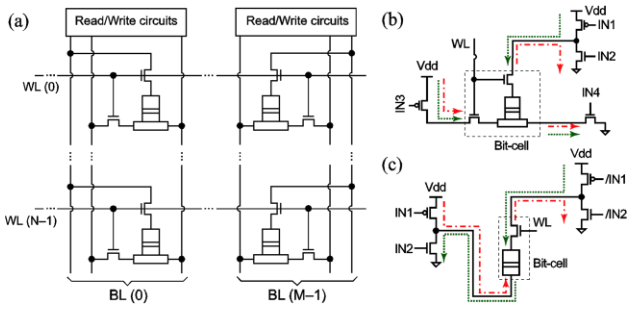

Fig. 4. (a) The overall structure of the spin-Hall-assisted STT-MRAM, (b) Write circuits for the spin-Hall-assisted STT-MRAM, (c) Write circuits for the conventional STT-MRAM.

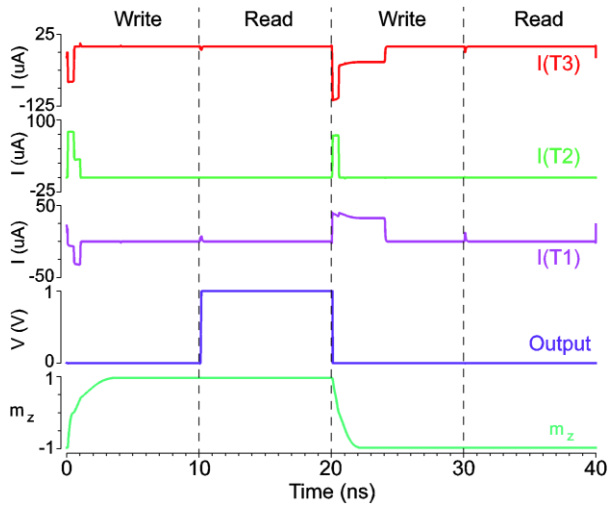

Fig. 5. Transient simulation of a bit-cell in the spin-Hall-assisted STT-MRAM.

be seen, the incubation delay is eliminated (see m_z), thus high-speed operation can be achieved.

TABLE I. Parameters for Circuit-level Simulation

Parameters	Values
Gilbert damping constant	0.03
TMR ratio	120%
Spin Hall angle	0.3
Thermal stability factor	31.56
R.A. for the MTJ	$10\ \Omega \cdot \mu m^2$
Free layer volume of the MTJ	50 nm × 50 nm × 0.8 nm
Dimensions of the heavy metal	60 nm × 50 nm × 3 nm
Width of the access transistors	150 nm × 2 ($10\ F^2$)

IV. Experimental Methodology

Spin-Hall-assisted STT-MRAM promises to serve as high-speed cache since it eliminates the incubation delay of the conventional STT-MRAM. In this section, we present the experimental methodology for evaluating the performance of the spin-Hall-assisted STT-MRAM for cache replacement.

A. Experiment platform

The experimental platform is shown in Fig. 6. Firstly, the bit-cell parameters are extracted from the aforementioned circuit-level simulation, as Table II. Then these parameters are fed into NVSim [18], which is an NVM evaluation simulator including area calculation, latency estimation, energy analysis, etc. Afterwards, based on the evaluation results from NVSim, system-level simulation is implemented with a cycle-accurate simulator Gem5 (Syscall emulation mode) [19]. This simulator could take the user-defined configuration to evaluate the instruction per cycle (IPC), execution time, etc. In addition, we modified Gem5 to take into account the asymmetry between read and write latencies [20]–[23]. We modified McPAT energy and power evaluation tool to evaluate cache hierarchy energy consumption [24].

B. Experimental configuration

With our experimental platform, various experiment workloads are used to evaluate the advantages and drawbacks

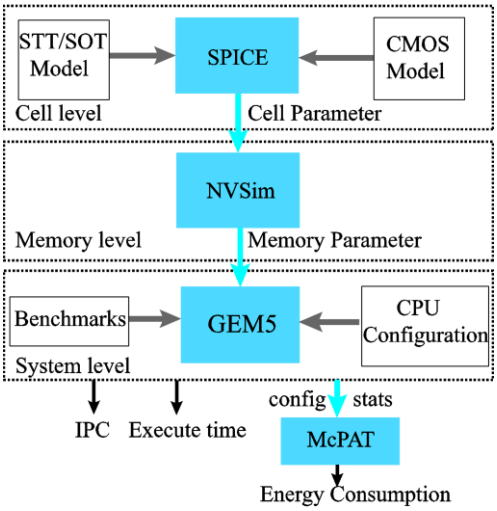

Fig. 6. Experimental platform for the cache evaluation and system-level simulation.

of three caches: spin-Hall-assisted STT-MRAM, the conventional STT-MRAM, and 6T-SRAM. The baseline is configured as Table III. A single ARMv7 core with a 32 KB instruction/data L1 cache and a 1 MB L2 cache is selected to run those workloads. We choose 14 data-intensive benchmarks from SPEC CPU 2006 as the test cases for the proposed cache technologies. Two criterions are used for estimating the performance: the one is speed measurement, reflecting the running speed of a single task; the other is throughout measurement, indicating the number of the instructions executed within a fixed duration. We run two billion instructions for each benchmark. The detailed evaluation results will be presented and discussed in the next section.

TABLE II. PARAMETERS FOR A SINGLE-CELL MRAM

	Spin-Hall-assisted STT-MRAM	**Conventional STT-MRAM**
Read latency	167 ps	243 ps
Read power	21 μW	16 μW
Write current (AP to P/P to AP)	STT: ~6 μA/~37 μA SHE: ~73 μA/~73 μA	113~124 μA/55~64 μA
Write latency[a] (AP to P/P to AP)	492 ps/562 ps	984.7 ps/1995 ps
Write energy	48.7 fJ	113.2 fJ

[a] the required delay for $m_z = 0$.

TABLE III. PARAMETERS FOR SYSTEM-LEVEL SIMULATION

CPU	Single-ARMv7(detailed) cores, 2Ghz, out-of-order
L1	Inst./Data 32KByte/32KByte, 64Byte line 4-way, 1 bank, Write-back SRAM : Lat(Read/Write) 0.6ns/0.6ns STT : Lat(Read/Write) 0.9ns/8.6ns SOT : Lat(Read/Write) 0.9ns/1.0ns
L2	1M, 64Byte line 8-way, 1 bank, Write-back SRAM : Lat(Read/Write), 2.0ns/2.0ns STT : (Read/Write), 2.1ns/10.2ns SOT : (Read/Write), 2.1ns/2.0ns
Execution Unit	2x ALU, 2x CALU, 2x FPU
Main Memory	8GB DDR3 1600MHz, 120cycle, 12.8GB/s.
Workload	Lbm, Mcf, Soplex, Libquantum, Leslie3d, Milc, Bzip2, Hmmer, Astar, Gromacs, Namd, Perlbench, Povray, Sjeng

V. ARCHITECTURE AND SYSTEM-LEVEL EVALUATION

In this section, we obtain area, latency, energy consumption and leakage power of three caches with NVSim. The performance comparison among them is also demonstrated. The capacity ranging from 32 KB to 16 MB are measured to find a good trade-off for the spin-Hall-assisted STT-cache. Finally, the IPC and execution time of three caches are estimated by Gem5 in order to shown their benefits and shortcomings.

A. Cache evaluation

Scaling Capacity: The scaling behaviors are shown in Fig. 7. The corresponding results are normalized to a 6T-SRAM cache for various capacities of the spin-Hall-assisted STT-cache and the conventional STT-cache. As can be seen, the performance of both the spin-Hall-assisted STT-cache and the conventional STT-cache is improved as the capacity increases. In particular, in larger capacity (e.g. > 1MB) these two caches show better performance than SRAM-based cache. Moreover, the spin-Hall-assisted STT-cache is advantageous to the conventional STT-cache in the aspect of read energy, write latency and write energy. The detailed analysis is shown as follows.

Area: It is anticipated that there is an optimal capacity for the cache to reduce miss rate. As the capacity is large enough, the cache area is dominated by the cell size instead of the peripheral circuits. According to Fig. 7(a), the SRAM area increases significantly with the capacity since the bit-cell includes 6 transistors. However, the MRAM cell is about 10 F^2 at the 28 nm technology node, which is smaller than the SRAM cell. Therefore both the spin-Hall-assisted STT-cache and the conventional STT-cache show higher storage density than the SRAM cache. For instance, an 8 MB MRAM cache occupies the same area as a 1 MB SRAM cache.

Read performance: Fig. 7(b) and (c) show the results of the read latency and read energy, respectively. When the capacity is relative small, the read performance of the MRAM cache is inferior to the SRAM cache. With the capacity increasing, the read latency and read energy of the MRAM cache are improved. Once the capacity is larger than 2 MB, the MRAM cache shows better read performance than the SRAM cache. The performance gap between the spin-Hall-assisted STT-cache and the conventional STT-cache is negligible, as they use the same sensing amplifier.

Write performance: The write latency and energy consumption are shown in Fig. 7(d) and (e), respectively. Among these three caches, the write latency of the SRAM cache is the smallest. The difference of the write latency between SRAM cache and MRAM cache decreases with the capacity increasing. Moreover, the spin-Hall-assisted STT-cache shows smaller write latency than the conventional STT-cache, which can be attributed to the assistance of the SHE.

The results of the write energy show the similar trend to the write latency. Overall, the write energy of the MRAM cache drops with the capacity. Especially above 512 KB, the MRAM cache consumes smaller write energy than SRAM

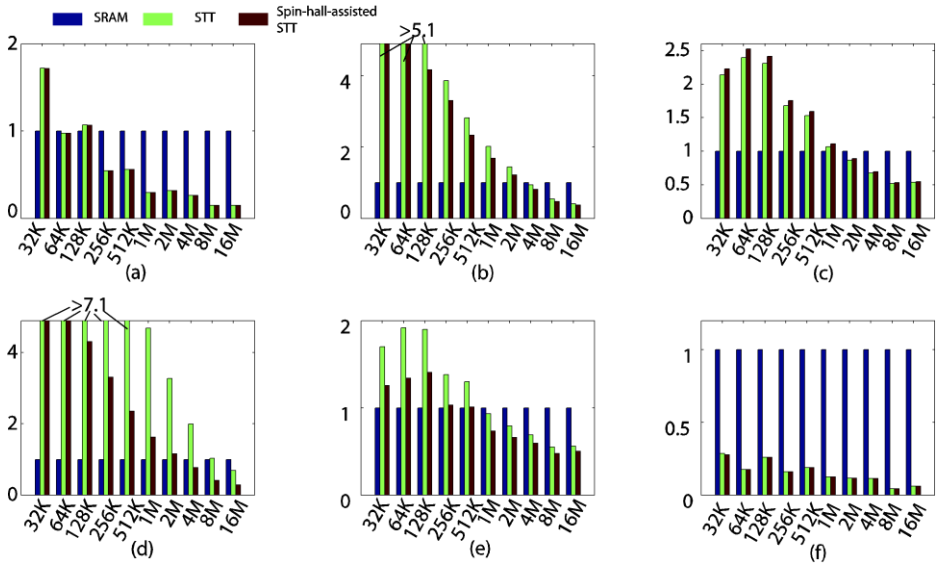

Fig. 7. Scaling behavior of the spin-Hall-assisted STT-cache and the conventional STT-cache normalized to SRAM-based cache.

cache. Moreover, compared to the conventional STT-cache, the spin-Hall-assisted STT-cache needs smaller write energy thanks to the decrease in the write current (see Table II).

Leakage Power: For the MRAM cache, the most of leakage power originates from the CMOS circuits since the storage device (MTJ) is non-volatile. Thus it can be inferred that the MRAM cache has ultralow leakage power compared to the SRAM cache, in agreement with the results of Fig. 7 (f). For example, the leakage power reaches higher than 10^4 mW when the capacity is larger than 8 MB, however, it is only 550 mW for the spin-Hall-assisted STT-cache. Even if the capacity is as large as 16 MB, its leakage power is only 1106 mW. Therefore, the spin-Hall-assisted STT-cache is suitable for the power-efficient embedded processor.

Summary: To sum up, the spin-Hall-assisted STT-MRAM shows promising potential in replacing the SRAM multi-level hierarchy cache, especially in the large capacity. However, it is difficult for the conventional STT-MRAM to act as cache due to the larger write latency.

B. System-level evaluation

To fully show the benefit of the spin-Hall-assisted STT-cache in the modern computer system, we also compare the SRAM-based multi-level caches with MRAM-based cache by system-level simulation.

Performance: With the same capacity (1MB), we evaluate the performance of on-chip cache under various benchmarks. The overall results are shown in Fig. 8. Thanks to the advantage of the spin-Hall-assisted STT-cache in read and write operation, some benchmarks achieve significant performance improvement (e.g. lbm, namd). Nevertheless, a few benchmarks suffer from little improvement due to infrequent data communication with off-chip memory. The maximum performance improvement can be up to 12%, and the average improvement is 9%.

Execution time: The results of execution time are shown in Fig. 9. Generally, the spin-Hall-assisted STT-cache reduces the access latency while used as L2 cache level. This is attributed to its high speed advantage at the large capacity. Most of benchmarks achieve ultra-fast speed because they need to use the cache frequently. The average reduction of execution time is nearly 4% compared to SRAM-based cache and 10% to STT-cache.

Energy Consumption: To present the advantage of our

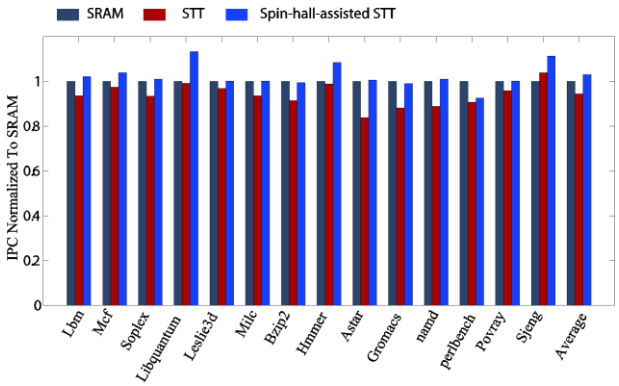

Fig. 8 IPC evaluation normalized to SRAM under various benchmarks.

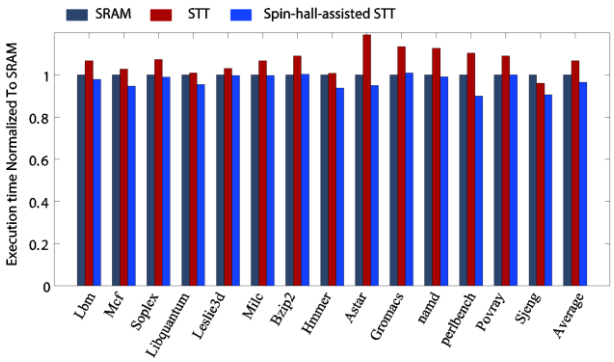

Fig. 9 Execution time evaluation normalized to SRAM under various benchmarks.

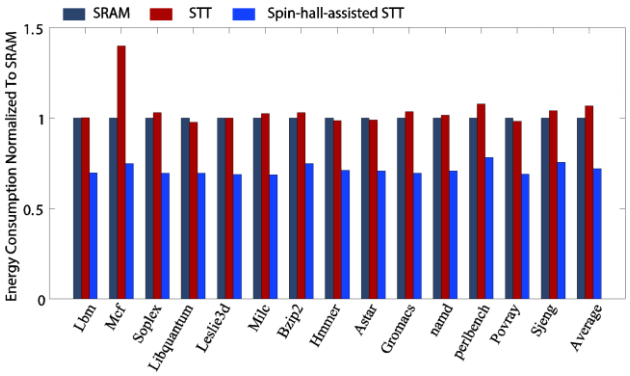

Fig. 10 Energy evaluation normalized to SRAM under various benchmarks.

proposed spin-Hall-assisted STT-cache, Fig. 10 shows our comparison of cache energy consumption based on various benchmarks. The results indicate that the spin-Hall-assisted STT cache consume the smallest energy. The average energy reduction nearly 27.98% compare to SRAM-based cache and 32.58% compare to STT-MRAM based cache replacement.

VI. Conclusion

In this paper, we have discussed the application potential of the spin-Hall-assisted STT-MRAM in the cache replacement by the device-to-system simulation. Compared to the conventional STT, the spin-Hall-assisted STT can eliminate the incubation delay and achieve high-speed and low-energy write operation for the MRAM. According to simulation results, the performance of the MRAM-based cache is improved as the capacity increases. Generally, the improvement of the spin-Hall-assisted STT-cache is more significant than the conventional STT-cache. For 1 MB capacity, the average performance improvement of the spin-Hall-assisted STT-cache is 9% compared to SRAM baseline. The average execution time reduction is 4% to SRAM baseline and 10% to the conventional STT-cache. In addition, spin-Hall-assisted STT-cache reduces 27.98% energy consumption compared to SRAM. As a result, the spin-Hall-assisted STT-cache promises to provide a comparable latency and energy consumption to SRAM-based cache, which makes it a viable candidate for on-chip memory and even low-level cache.

References

[1] N. S. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J. S. Hu, M. J. Irwin, M. Kandemir, and V. Narayanan, “Leakage current: Moore's law meets static power,” Computer, vol. 36, no. 12, pp. 68–75, 2003.

[2] S. Yuasa, A. Fukushima, K. Yakushiji, T. Nozaki, M. Konoto, H. Maehara, H. Kubota, T. Taniguchi, H. Arai, H. Imamura, K. Ando, Y. Shiota, F. Bonell, Y. Suzuki, N. Shimomura, E. Kitagawa, J. Ito, S. Fujita, K. Abe, K. Nomura, H. Noguchi, and H. Yoda, “Future Prospects of MRAM Technologies,” in IEEE-IEDM, pp. 3.1.1–3.1.4, 2013.

[3] W. Kang, et al, Spintronics: Emerging ultra-low power circuits and systems beyond MOS technology. ACM J. Emerg. Technol. Comput. Syst., vol. 12, no. 2, 2015.

[4] W. Kang, et al, Reconfigurable Codesign of STT-MRAM Under Process Variations in Deeply Scaled Technology. IEEE Trans. Electron Devices, vol. 62, no. 6, pp. 1769–1777, 2015.

[5] S Ikeda, et al. “Recent Progress of Perpendicular Anisotropy Magnetic Tunnel Junctions for Nonvolatile VLSI, ” SPIE, vol. 2, no. 3, pp. 1240003, 2012.

[6] J. M. Slaughter et al., "High density ST-MRAM technology (Invited)," in IEDM, San Francisco, CA, 2012

[7] I. M. Miron, K. Garello, G. Gaudin, P.-J. Zermatten, M. V. Costache, S. Auffret, S. Bandiera, B. Rodmacq, A. Schuhl, and P. Gambardella, “Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection,” Nature, vol. 476, pp. 189–193, 2011.

[8] L. Liu, C.-F. Pai, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman, “Spin-Torque Switching with the Giant Spin Hall Effect of Tantalum,” Science, vol. 336, no. 6081, pp. 555–558, 2012.

[9] S. Fukami, C. Zhang, S. DuttaGupta, et al. “Magnetization switching by spin-orbit torque in an antiferromagnet-ferromagnet bilayer system,” Nat Mater., 2016.

[10] A. van den Brink, G. Vermijs, A. Solignac, et al. “Field-free magnetization reversal by spin-Hall effect and exchange bias,” Nat Commun, pp. 7: 10854, 2016.

[11] W. Kang, Z. Wang, Y. Zhang, et al. Spintronic logic design methodology based on spin Hall effect–driven magnetic tunnel junctions. Journal of Physics D: Applied Physics, 2016, 49(6): 065008.

[12] Z. Wang, W. Zhao, E. Deng, J. O. Klein, and C. Chappert, “Perpendicular-anisotropy magnetic tunnel junction switched by spin-Hall-assisted spin-transfer torque,” J. Phys. D: Appl. Phys., vol. 48, no. 6, pp. 065001, 2015.

[13] C. Chappert, A. Fert, and F. Nguyen Van Dau, “The emergence of spin electronics in data storage,” Nature Mater., vol. 6, pp. 813–823, 2007.

[14] J. C. Slonczewski, “Current-driven excitation of magnetic multilayers,” J. Magn. Magn. Mater., vol. 159, no. 1–2, pp. L1–L7, 1996.

[15] L. Berger, “Emission of spin waves by a magnetic multilayer traversed by a current,” Phys. Rev. B, vol. 54, no. 13, pp. 9353–9358, 1996.

[16] Z. Wang, W. Zhao, E. Deng, Y. Zhang, and J. O. Klein, “Magnetic non-volatile flip-flop with spin-Hall assistance,” Phys. Status Solidi RRL., vol. 9, no. 6, pp. 375–378, 2015.

[17] W. Zhao, C. Chappert, V. Javerliac, and J.-P. Nozière, “High Speed, High Stability and Low Power Sensing Amplifier for MTJ/CMOS Hybrid Logic Circuits,” IEEE Trans. Magn., vol. 45, no. 10, pp. 3784–3787, 2009.

[18] X. Dong, C. Xu, Y. Xie and N. P. Jouppi, "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 31, no. 7, pp. 994-1007, July 2012.

[19] N. Binkert, B. Beckmann, G. Black, S.K. Reinhardt, A.Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Swell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood. 2011. “The gem5 simulator”. SIGARCH Comput. Archit. News, August 2011.

[20] G. Prenat *et al*., "Ultra-Fast and High-Reliability SOT-MRAM: From Cache Replacement to Normally-Off Computing," in IEEE Transactions on Multi-Scale Computing Systems, vol. 2, no. 1, pp. 49-60, Jan.-March 1 2016.

[21] F. Oboril, R. Bishnoi, M. Ebrahimi and M. B. Tahoori, "Evaluation of Hybrid Memory Technologies Using SOT-MRAM for On-Chip Cache Hierarchy," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 34, no. 3, pp. 367-380, March 2015.

[22] Y. Zhang, C. Zhang, J. Nan, Z. Zhang, X. Zhang, J. O. Klein, D. Ravelosona, G. Sun, and W. Zhao, "Perspectives of Racetrack Memory for Large-Capacity On-Chip Memory: From Device to System," in IEEE Transactions on Circuits and Systems I: Regular Papers , vol.PP, no.99, pp.1-10, 2016.

[23] G. Sun, C. Zhang, P. Li, T. Wang, and Y. Chen, "Statistical Cache Bypassing for Non-Volatile Memory," in IEEE Transactions on Computers , vol.PP, no.99, pp.1-1, 2016.

[24] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen and N. P. Jouppi, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," *in MICRO*, New York, NY, 2009, pp. 469-480.