A Novel Hybrid Last Level Cache Based on Multi-retention STT-RAM Cells

Hongguang Zhang[1(✉)], Minxuan Zhang[1,2], Zhenyu Zhao[1], and Shuo Tian[1]

[1] College of Computer, National University of Defense Technology, Changsha 410073, People's Republic of China
{zhanghongguangl4, mxzhang, zyzhao, tianshuol4}@nudt.edu.cn

[2] National Key Laboratory of Parallel and Distributed Processing, National University of Defense Technology, Changsha 410073, People's Republic of China

Abstract. Spin-transfer torque random access memory (STT-RAM) is one of the most promising substitutes for universal main memory and cache due to its excellent scalability, high storage density and low leakage power. A much larger cache capacity in the same die footprint can be implemented with STT-RAM because its area is only 1/9 to 1/3 that of SRAM. However, the non-volatile STT-RAM also has some drawbacks, such as long write latency and high write energy, which limit its application in cache design. To solve the two problems, we relax the retention time of STT-RAM to optimize its write performance and energy, and propose a novel multi-retention STT-RAM hybrid last level cache (LLC) architecture, which is realized with three different kinds of cells. In addition, we design the data migration scheme to manage its block allocation, thus improving overall system performance further. The experiment results show that our multi-retention hybrid LLC reduces the total power consumption by as much as 96.6 % compared with SRAM LLC, while having almost the same (at 99.4 %) instruction per cycle (IPC).

Keywords: STT-RAM · Last level cache · Multi-retention · Data migration

1 Introduction

Power has been the dominator of the increasing of CPU's frequency since one decade ago. This has generated a considerable volume of research in multi-core processor to provide sustainable performance enhancement of computer system. However, the gap of access speed between main memory and processor is becoming larger and has been the bottleneck of overall system performance. Cache is developed to alleviate this mismatch problem.

SRAM has been the mainstream of caches for many years because it owns high access speed, low dynamic power and other good characters. However, with more and more cores are embedded on chip, caches need larger size. However, increasing capacity of SRAM caches lead to high leakage power, which takes up the dominator of

J. Wu and L. Li (Eds.): ACA 2016, CCIS 626, pp. 28–39, 2016.
DOI: 10.1007/978-981-10-2209-8_3

the microprocessor's overall power consumption, therefore, researchers are focusing on alternative substitutes for SRAM.

Spin-transfer torque random access memory (STT-RAM) is regarded as the most promising replacement for SRAM because it has almost all desired characters of the universal memory and cache, such as high storage density, fast read access speed and non-volatility. However, we are faced with two drawbacks of STT-RAM, namely, long write latency and high write energy, which result in the reduction of system performance and the enhancement of dynamic power consumption.

Hybrid cache scheme is proposed to address the write access speed and energy of STT-RAM. For example, the SRAM/STT-RAM hybrid cache in [7, 8] moves write intensive data blocks into SRAM region to reduce the average write latency. However, even a small SRAM partition can bring in very high leakage power. Researchers discover that relaxing the data retention time could significantly optimize its write performance, which can even exceed that of SRAM. That makes the multi-retention hybrid cache architecture possible. In [2], a new cache hierarchy is proposed to improve the overall system performance with multi-retention STT-RAM cell, and the outcome is good.

In this paper, we relax the retention time of STT-RAM and propose a novel multi-retention STT-RAM hybrid last level cache with three kinds of STT-RAM cells, which is different with the design in [2, 11], to obtain an improvement of overall performance. We simulate the proposed cache design on architecture simulator, and collect the test results of benchmarks to analysis the overall system performance and power consumption.

2 STT-RAM Features

2.1 MTJ Features

The magnetic tunnel junction (MTJ) shown in Fig. 1 is the basic storage device of STT-RAM. The MTJ has two layers, namely, free layer and reference layer. The magnetic direction of reference layer is fixed, however, that of free layer can be switched by current. If the magnetic directions of two magnetic layers are parallel, the MTJ is in low-resistance state; otherwise it is in high-resistance state.

The most widely used STT-RAM storage cell is one transistor one MTJ (1T1J) at present. In memory array, the STT-RAM cell is connected to word line (WL), bit line (BL) and source line (SL). The WL is used to select the specific row, and the voltage gap between SL and BL is used to complete write and read operation. When executing a read operation, we add a negative voltage between SL and BL and use a sense amplifier to get the current flowing throw the MTJ, thus knowing the current resistance of MTJ. When writing "0" into STT-RAM cell, there is a positive voltage between SL and BL. However, when writing "1" into STT-RAM, a negative voltage is applied. The current used to switch the MTJ's state is called switching current, and its value is mainly determined by the write pulse width, which is represented by T_w in this paper.

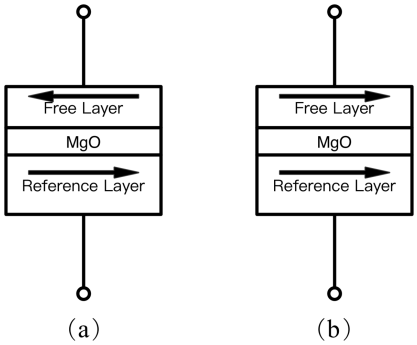

Fig. 1. The MTJ design (1T1J). (a) MTJ in high-resistance state. (b) MTJ in low-resistance state.

2.2 MTJ Non-volatility

The MTJ's non-volatility can be analyzed quantitatively with the retention time of MTJ. We use τ to represent its retention time. τ is related to the thermal stability factor Δ and can be calculated with Eq. (1) [1].

$$\tau \approx \tau_0 \, exp(\Delta) \tag{1}$$

τ_0: The attempt time and set as 1 ns.
Δ is derived from Eq. (2).

$$\Delta = \frac{E_F}{k_B T} = \frac{M_s V H_K}{2 k_B T} \tag{2}$$

M_s: The saturation magnetization.
H_k: The effective anisotropy field.
T: The working temperature.
K_B: The Boltzmann constant.
V: The volume for the STT-RAM write current.

From Eqs. (1) and (2), we can know that the data retention time of a MTJ decreases exponentially when its working temperature T increases.

According to the different T_w, MTJ has three regions, namely, the thermal activation, dynamic reverse and processional switching. Their distribution is shown as Fig. 2.

The switching current in each working region can be calculated approximately by Eqs. (3)–(5) [2].

$$J_C^{Therm} T_w = J_{C0}\left(1 - \frac{1}{\Delta} ln\left(\frac{T_w}{\tau_0}\right)\right) \quad (T_w > 20 \text{ ns}) \tag{3}$$

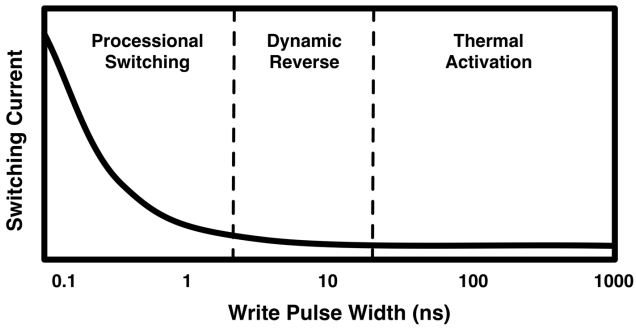

Fig. 2. The three working region of MTJ

$$J_c^{Dyn}T_w = \frac{J_c^{Therm}T_w + J_c^{Prec}T_w e^{(-A(T_w - T_{PIV}))}}{1 + e^{(-A(T_w - T_{PIV}))}} \quad (20\text{ ns} > T_w > 3\text{ ns}) \tag{4}$$

$$J_C^{Prec}T_w = J_{C0} + \frac{C\ ln\left(\frac{\pi}{2\theta}\right)}{T_w} \quad (T_w < 3\text{ ns}) \tag{5}$$

$$J_{C0} = \left(\frac{2e}{\hbar}\right)\left(\frac{\alpha}{\eta}\right)(t_F M_S)(H_k \pm H_{ext} + 2\pi M_S) \tag{6}$$

Where $J_C T_w$ is required switching current, A, C, and T_{PIV} are fitting parameters, J_{CO} is the threshold of switching current density, e is the electron charge, $\hbar$ is the reduced Planck constant, α is the damping constant, H_{ext} is the external field, η is the spin transfer efficiency, t_F is the free layer thickness.

Based on the above analysis, we can adjust the value of J_C and Δ by changing several related parameters, such as M_S, t_F, H_k and V.

We get three $I_C - T_w$ curves shown in Fig. 3. for STT-RAM cells whose retention time are 2.5 years ($\Delta = 38.9$), 3 s ($\Delta = 21.8$) and 30 μs ($\Delta = 10.3$). In this paper they are called HRS, MRS and LRS respectively.

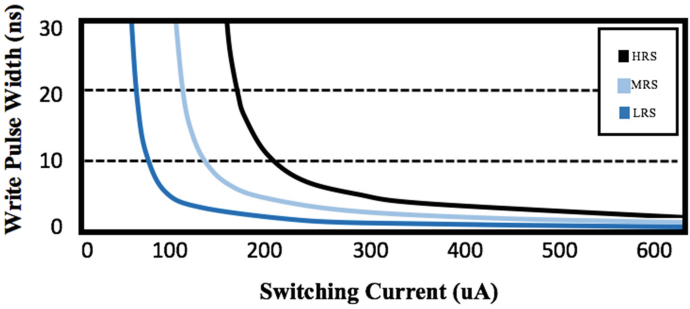

Fig. 3. The $I_c - T_w$ for HRS, MRS and LRS MTJ cells

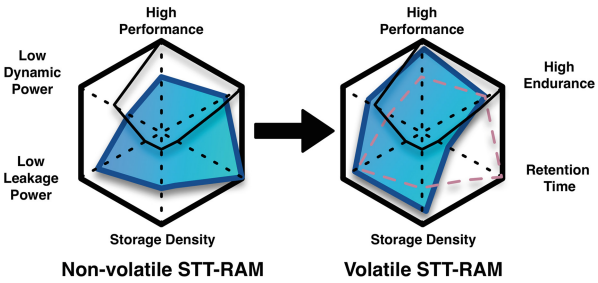

Fig. 4. The difference between non-volatile and volatile STT-RAM

It is clear that the higher the retention time, the lower the $I_c - T_w$ curve. With the same switching current, LRS's write pulse width is the shortest one and HR is the longest. If their write pulse widths are the same, the switching current required by LRS is the lowest. The performance difference between non-volatile and volatile STT-RAM is shown in Fig. 4 [3]. The dotted border is optimal and black line is SRAM. The blue region is STT-RAM.

3 STT-RAM LLC Design

3.1 Cache Parameters

Although the long retention time of STT-RAM can offer low leakage power consumption, it leads to long write latency and high write energy. To reduce the write latency and energy, we relax the retention time of STT-RAM to improve its write performance.

In Sect. 2, we find that the STT-RAM cells whose retention time are relaxed to μs and ms level can satisfy the access speed of all level caches. So we simulate the proposed HRS, MRS and LRS cells on NVSim [6] to get their parameters in 1 MB last level cache design. The results are shown in Table 1 .

Table 1. The parameters for multi-retention STT-RAM cells.

Parameters	SRAM	LRS	MRS	HRS
Area/F^2	125	21	22	23
Switching Time/ns	/	2.0	5.0	10.0
Retention Time	/	30 μs	3.0 s	2.5 years
Read Latency/ns	2.735	2.085	2.097	2.210
Read Latency/Cycles	6	5	5	5
Read Energy/nJ	0.181	0.083	0.087	0.099
Write Latency/ns	2.301	2.431	5.427	10.936
Write Latency/Cycles	5	5	11	22
Write Energy/nJ	0.112	0.479	1.016	1.978
Leakage Power/mW	1261.7	26.9	31.1	36.2

From Table 1, it can be seen that the performance varies with different retention time. LRS's access speed is even better than SRAM, while HRS's write latency is longer than 10 ns.

3.2 Hybrid LLC Architecture

In previous section, we get their overall performance of LRS, MRS and HRS cells. We find that LRS owns the fastest access speed, so if we adopt LRS to design LLC, the LLC's performance can be enhanced significantly. However, it should be noticed that the data stored in LRS or MRS blocks will be invalid after its short retention time, so we must use refresh scheme to improve the reliability. For LLC with large capacity (1 MB or larger), it can be foreseen that the refresh energy and the hardware overhead are unbearable in this situation. Typically, the hardware overhead is 0.80 %. So it is not suitable to design LLC with LRS or MRS purely. Considering the existed SRAM/STT-RAM hybrid cache architecture [13], which fully utilize both the fast write speed of SRAM and the excellent features of STT-RAM, and other designs in [4, 5], the hybrid LLC based on volatile STT-RAM is possible. A multi-retention hybrid cache design is proposed in [2], however, the large capacity of LLC offers more choices, so we propose to design an optimized novel multi-retention hybrid cache architecture.

We find that if we add a MRS-Region in LRS/HRS hybrid LLC, its performance can be promoted further and power consumption can be reduced although the hardware overhead is a bit higher than the original design. The reason why we do not expand LRS-Region is that the block-refresh and counter-reset happen frequently in LLC in case that the size of LRS-Region is too large, thus leading to a very high power consumption. In addition, the large amount of counter requires larger on-chip area and hardware overhead. These factors make it unsuitable to expand LRS-Region further. The LRS/MRS hybrid LLC is also one choice, however, the retention time of MRS can not make sure all data are reliable though the retention time of MRS is longer than LRS. We still need the refresh scheme, thus contributing to serious refresh power consumption problem.

Based on the above analysis, the LRS/MRS/HRS multi-retention hybrid LLC is one of the best choices that we can find at present. we separate the 1 MB LLC into 16 ways, way0 is LRS-Region and realized by LRS cells, way1–3 is MRS-Region and made by MRS cells, way4–15 is HRS-Region and consist of HRS cells only.

To improve the reliability of LRS-Region and MRS-Region, we add a refresh-counter and an access-counter for every LRS or MRS block. The refresh-counter is used to monitor the duration that the data has been stored in that block while the access-counter is utilized to record its read access number during the retention time. The refresh counters are controlled by a global clock whose period is T_{gc}. The value of refresh-counter is N_{ref}, and that of access-counter is N_{ac}. At the end of each T_{gc}, all refresh-counters will be increased by 1. If there is a read access to one block, its access-counter is increased by 1. However, if there is a write access to the block, both its refresh-counter and access-counter are initialized to 0. The maximum value of refresh-counter N_{max} depends on their different retention time. When a LRS or MRS block's N_{ref} reaches N_{max}, we do not conduct a refresh operation but check its

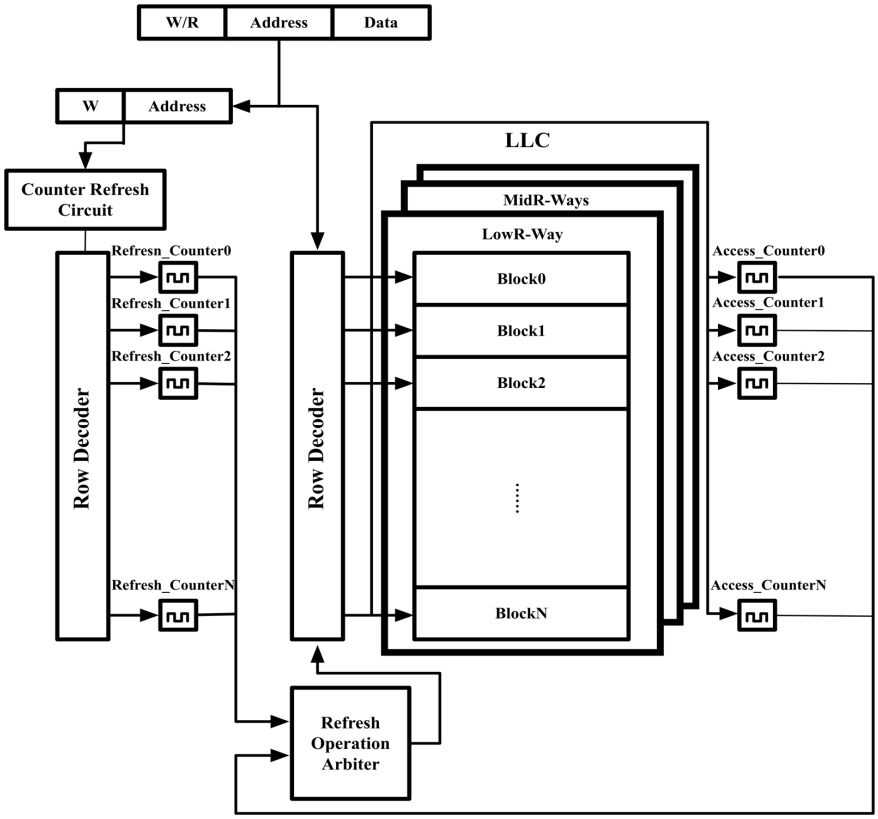

Fig. 5. The counter-based writeback refresh scheme

N_{ac}. We write it back to HR-Region in case of $N_{ac} > 5$, otherwise write it back to main memory. The whole scheme shown in Fig. 5 is called Counter-based Writeback Refresh Scheme (CWRS).

The design of counter is shown as Fig. 6. The hardware overhead of CWRS is (4 bits × 2 × 4)/(64 bytes × 16) = 0.39 %, the overall area needed is $(4 \times 125F^2 \times 2 \times 4)/(64 \times 8 \times 40F^2 \times 16) = 1.22$ %. Based on simulation results, these counters' power consumption takes up only less than 1 % of the total power consumption, which has little influence on the overall performance.

To improve overall system performance, we create a write intensive block prediction table (WIBPT) to predict and monitor write intensive blocks. WIBPT has 64 entries, and each entry consists of an address and a counter. We divide all write intensive blocks (WIB) into three levels, namely, WIB1, WIB2 and WIB3, to support the migration scheme in our hybrid LLC.

When a request comes to LLC, firstly we detect what kind of operation it is and if it is a hit. If it is a miss, we allocate a LRS block for it. If the request is a write hit, we detect if its address is already in WIBPT. If so, its access counter is increased by 1, otherwise we add its address to WIBPT and reset the counter to 0. If WIBPT is full, we

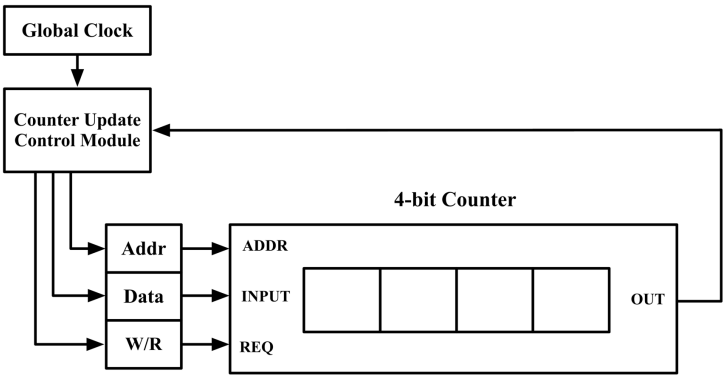

Fig. 6. The counter design

kick the LRU entry and add this new address. Then we detect the value of its counter, if the counter is less than 4, we define it as WIB1 and do nothing; if it is larger than 4 and less than 8, we define it as WIB2 and swap it with blocks in MRS-Region; if the counter is larger than 8 [12], we name it as WIB3 and migrate it to LRS-Region. A migration operation needs read the data from original cache block firstly, and then write it to the target. It consumes two read and write operations. This dynamic power is added into the final results.

The proposed data migration policy is demonstrated by Fig. 7. In this way, we obtain a better tradeoff between performance and power consumption. To illustrate, the

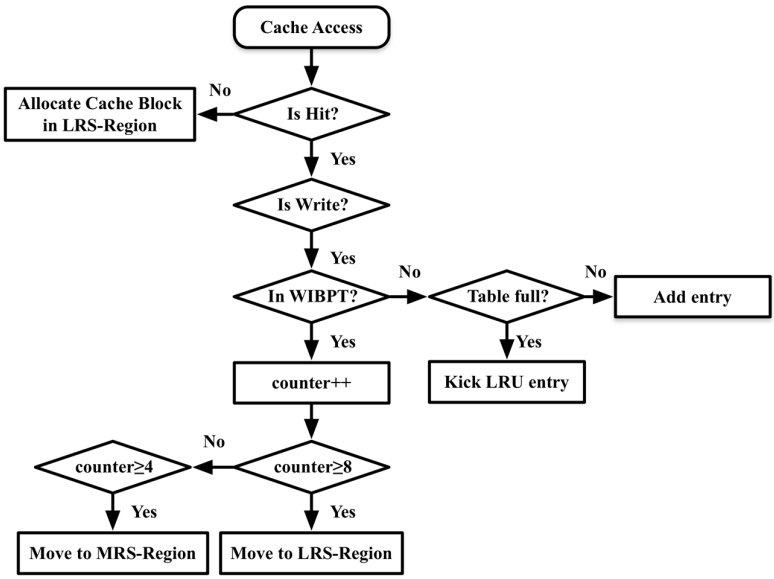

Fig. 7. The migration scheme

overall system performance can be improved significantly, while the total power consumption is much lower than SRAM LLC.

Compared with SRAM/STT-RAM Hybrid LLC, our design can have better overall performance and leakage power with the same migration scheme. The extra power consumption that MRS-Region brings in can be ignored because the number of refresh and reset operations in MRS-Region is limited. However, the refresh circuits of MRS-Region lead to extra hardware overhead.

4 Simulation

4.1 Experimental Setup

We evaluate proposed multi-retention hybrid LLC on GEM5 [9, 10]. GEM5 is an universal architecture simulator. It has a highly configurable simulation framework, including support for various universal ISAs and multiple cache coherence protocols (MESI, MOESI, etc.).

The configuration for GEM5 is shown as Table 2. The private L1 cache is 32 KB, the private L2 cache is 256 KB, and the shared L3 cache is 1 MB. The ISA we use is X86 instruction set.

Table 2. GEM5 configuration

Computer system	Configuration
CPU	X86, O3, 2 GHz
L1 Icache	Private, 32 KB, 2-way
L1 Dcache	Private, 32 KB, 2-way
L2 Cache	Private, 256 KB, 8-way
L3 Cache	Shared, 1 MB, 16-way
Main Memory	1024 MB, 1-channel

4.2 Architectural Simulation

We simulate SPEC CPU2006, including 401.bzip2, 403.gcc, 429.mcf, 445.gobmk, 456.hmmer, 458.sjeng and 462.libquantum, on proposed multi-retention hybrid LLC, and compare its performance [instruction per cycle, (IPC)] as well as power consumption with SRAM LLC. We also simulate high-retention STT-RAM LLC and SRAM/STT-RAM hybrid LLC (1 SRAM-way and 15 STT-RAM-ways) as samples. The LRS/HRS hybrid design shares almost the same with SRAM/STT-RAM hybrid LLC, so we do not simulate it again here. All outcomes are normalized to the results of SRAM LLC.

The final IPC results are shown as Fig. 8. It can be seen that the overall performance of our proposed LLC design is the best one among the three STT-RAM cache architecture, which is 0.6 % lower than SRAM LLC. The performance of SRAM/STT-RAM Hybrid LLC is 2.8 %lower than SRAM LLC. The performance of HRS LLC is the lowest one, at 94.8 %.

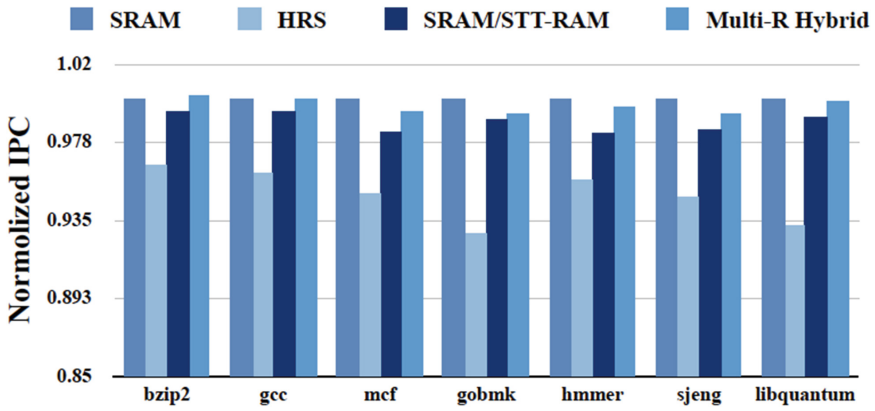

Fig. 8. The normalized IPC results

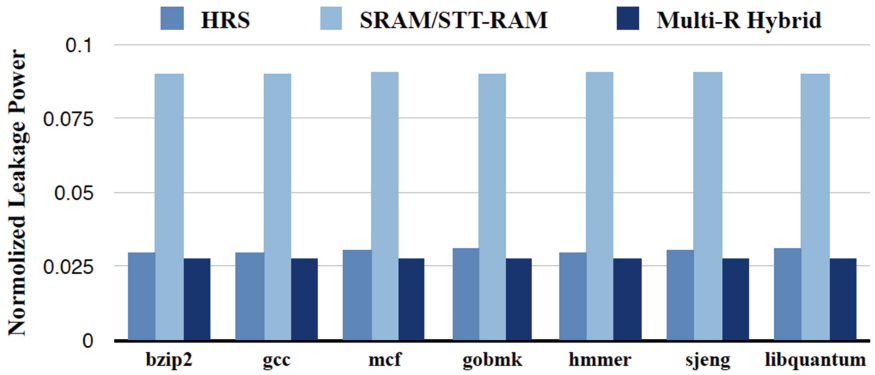

Fig. 9. The normalized leakage power consumption

The leakage power consumption results are shown as Fig. 9. The SRAM/STT-RAM Hybrid LLC has the highest leakage power consumption, at 9.0 %, while that of Multi-R Hybrid LLC is only 2.7 %.

The dynamic power consumption results are shown in Fig. 10. We can find that the average dynamic power consumptions of the three STT-RAM-based LLC designs are all much higher than SRAM. The HRS LLC shares the highest one, at 582 %. The Multi-R Hybrid LLC (at 401 %) is a bit higher than SRAM/STT-RAM Hybrid LLC (at 382 %).

The overall power consumption shown in Fig. 11 is the sum of leakage and dynamic power consumption. The overall power consumption of Multi-R Hybrid LLC is only 3.2 % that of SRAM, which is 52.3 % lower than SRAM/STT-RAM Hybrid LLC (at 9.1 %).

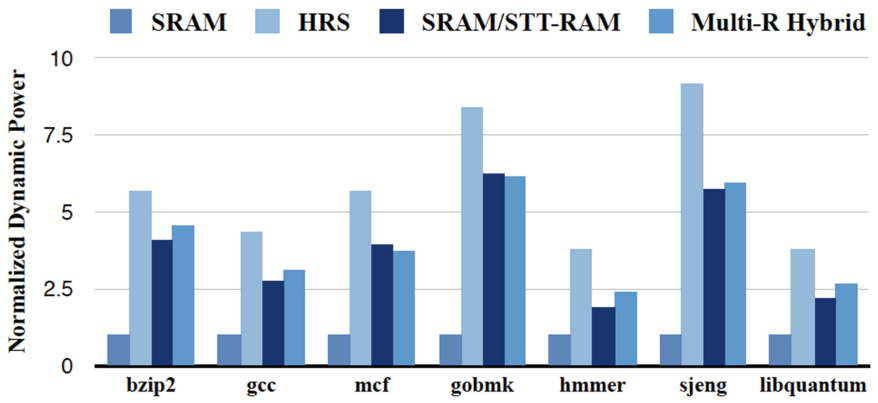

Fig. 10. The normalized dynamic power consumption

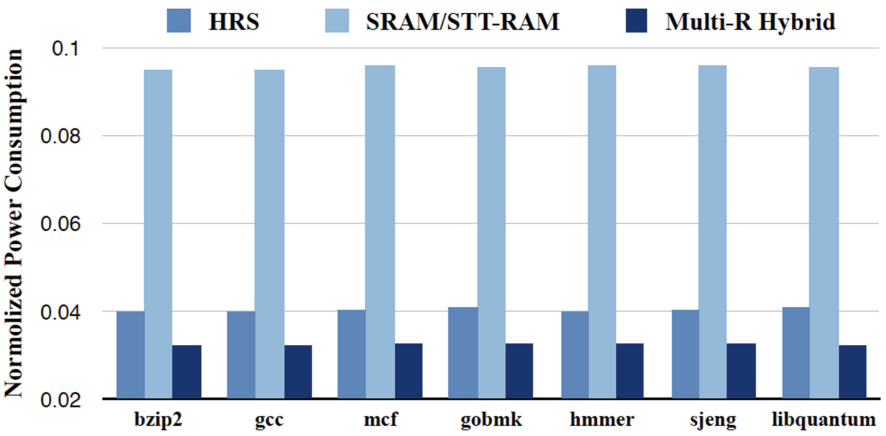

Fig. 11. The normalized overall power consumption

5 Conclusion

In this paper, we propose a novel hybrid last level cache architecture based on three different kinds of STT-RAM cells. Each kind of cells has totally different write performance.

Our simulation results show that the proposed Multi-R Hybrid design has almost the same overall performance with SRAM LLC (at 99.4 %), while having only 3.2 % power consumption. In addition, the total on-chip area of Multi-R Hybrid LLC can be saved by 81.6 % ideally. Compared with SRAM/STT-RAM Hybrid LLC, the Multi-R Hybrid LLC's IPC is increased by 2.2 % while its power consumption is reduced by 70 %.

Acknowledgements. The project is sponsored by National Science and Technology Major Project, "The Processor Design for Super Computer" (2015ZX01028) in China and the Excellent Postgraduate Student Innovation Program (4345133214) of National University of Defense Technology.

References

1. Jog, A., Mishra, A.K., et al.: Cache revive: architecting volatile STT-RAM caches for enhanced performance in CMPs. In: IEEE Design Automation Conference, pp. 243–253 (2012)
2. Sun, Z., Bi, X., et al.: STT-RAM cache hierarchy with multiretention MTJ design. IEEE Trans. Very Large Scale Integr. Syst. **22**(6), 1281–1294 (2014)
3. Smullen, C., Mohan, V., et al.: Relaxing non-volatility for fast and energy-efficient STT-RAM caches. In: IEEE Symposium on High-Performance Computer Architecture, pp. 50–61 (2011)
4. Li, J., Shi, L., et al.: Low-energy volatile STT-RAM cache design using cache-coherence-enabled adaptive refresh. ACM Trans. Des. Autom. Electron. Syst. **19** (1), 1–23 (2013)
5. Zhao, J., Xie, Y.: Optimizing band width and power of graphics memory with hybrid memory technologies and adaptive data migration. In: Proceedings of the International Conference Computer-Aided Design, pp. 81–87 (2012)
6. NVSim. http://www.rioshering.com/nvsimwiki/index.php
7. Li, Q., Li, J., et al.: Compiler-assisted STT-RAM-based hybrid cache for energy efficient embedded systems. IEEE Trans. Very Large Scale Integr. Syst. **22**(8), 1829–1840 (2014)
8. Raychowdhury, A., et al.: Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances. In: IEEE International Electron Devices Meeting, pp. 1–4 (2009)
9. Binkert, N., Beckmann, B., et al.: The gem5 simulator. ACM SIGARCH Comput. Archit. News **39**(2), 1–7 (2011)
10. Gem5. http://gem5.org
11. Sun, Z., Bi, X., Li, H.: Multi retention level STT-RAM cache designs with a dynamic refresh scheme. In: 44th Annual IEEE/ACM International Symposium on Microarchitecture, pp. 329–338 (2011)
12. Ahn, J., Yoo, S., et al.: Write intensity prediction for energy-efficient non-volatile caches. In: IEEE International Symposium on Low Power Electronics and Design, pp. 223–228 (2013)
13. Wang, Z., Jimenez, D., et al.: Adaptive placement and migration policy for an STT-RAM-based hybrid cache. In: 20th IEEE International Symposium on High Performance Computer Architecture, pp. 13–24 (2014)