STT-RAM Based Energy-Efficient Hybrid Cache Architecture for 3D Chip Multiprocessors

Fen Ge, Lei Wang, Hao Lu, Ning Wu, Fang Zhou, and Ying Zhang

Abstract—**With increasing the number of cores on a chip in Chip-Multiprocessors (CMPs), more cache resources are needed, and as a result, the leakage power consumption of the cache accounts for a larger proportion of the total chip power consumption. The emerging non-volatile memory (NVM) is expected to replace traditional memory devices due to its high density, near zero leakage power, and nonvolatility. In this paper, we use STT-RAM, a most promising candidate of NVM, to construct a energy-efficient hybrid cache architecture for 3D CMP. For the hybrid cache architecture design, we proposed a spherical placement approach to determine the optimal placement of STT-RAM and SRAM cache banks. This paper further proposes an optimized hybrid cache dynamic migration scheme, to reduce the data migration jitter and solve the problem of data migration failure in the hybrid cache architecture. The experimental results show that our proposed hybrid cache architecture with spherical placement and optimized data migration scheme can achieve 34.94% energy saving on average with only 1.49% performance degradation, compared with the architecture which uses pure SRAM as the cache in the same capacity.**

Index Terms—**Chip multiprocessors, data migration, hybrid cache, non-volatile memory**

I. Introduction

MODERN computer architecture has shifted from single-core chip to multi-core chip designs, which are also referred to as chip multiprocessors (CMPs). As Moore's law continues, the number of cores in CMP is increasing rapidly every generation. With increasing the number of cores, more cache resources are needed to feed all the cores. The three-dimensional integrated circuits (3D ICs) technology, where multiple cache layers are stacked vertically, has been proved to be a promising solution to increase the cache resources on a chip [1-4]. This is because the 3D ICs design can mitigate the large off-chip memory access latency and bandwidth constraints. The CMP that uses 3D ICs design is called 3D CMP. However, cache architecture is one of the most power-hungry parts of the 3D CMP, because the leakage power in cache systems has become an import contributor in the overall chip power consumption. Therefore, it is very necessary to reduce the power consumption generated by the cache architecture.

The emerging non-volatile memories (NVM) have potential application in the memory architecture to reduce the leakage power consumption due to their benefits such as high storage density and near zero leakage power consumption [5]. Spin-transfer torque random access memory (STT-RAM) is the most promising candidate of NVM, because it combines the access speed of SRAM, high density and low leakage power of DRAM. Although STT-RAM has so many advantages, it has large write latency and write power consumption, which restricts it to directly replace traditional memory devices. In order to overcome these disadvantages of STT-RAM, a hybrid cache architecture is adopted. In the hybrid cache architecture, the ratio of SRAM to STT-RAM and their placement affect the power consumption in cache architecture. Besides, the data migration scheme between SRAM and STT-RAM has significant effects in the dynamic power consumption. Therefore, in the hybrid cache architecture, the optimal placement of STT-RAM and SRAM and the migration scheme are worth researching problems.

In this paper, we focus on energy-efficient hybrid cache architecture design for 3D CMPs. Specifically, we propose a spherical placement approach and an optimized dynamic migration scheme for the hybrid cache integrating SRAM with STT-RAM. Parts of our work have been presented in [6]. This paper expands the previous work with a further analysis of the hybrid cache construction and the impact of placement approach and migration scheme on performance of 3D CMP.

The rest of the paper is organized as follows: section II summarizes related work; section III describes the proposed hybrid cache architecture for 3D CMP; section IV presents the hybrid cache placement approach. Section V describes the hybrid cache dynamic migration scheme; Experimental results are demonstrated in section VI and we finally conclude our work in section VII.

II. Related Work

In recent years, hybrid cache architecture has attracted considerable attention. A number of recent studies have proposed to integrate NVMs with SRAM to construct hybrid caches in CMPs. Sun et al. [7] first proposed using STT-RAM as the L2 cache stacked atop CMPs to reduce

Manuscript received January 7, 2019.This work was supported in part by the Natural Science Foundation of China under Grant 61774086 and 61701228, the Natural Science Foundation of Jiangsu Province under Grant BK20160806, and the Fundamental Research Funds for the Central Universities under grant NS2016041 and NS2017023.

Fen Ge, the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics(NUAA), Nanjing 210016, China (email: gefen@nuaa.edu.cn)

Lei Wang, the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics(NUAA), Nanjing 210016, China (email: 18326952570@163.com)

Hao Lu, the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics(NUAA), Nanjing 210016, China (email: paper_lew@163.com)

Ning Wu, Fang Zhou, and Ying Zhang, are the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics(NUAA), Nanjing 210016, China (email: wunee@nuaa.edu.cn, zfnuaa@nuaa.edu.cn, tracy403@nuaa.edu.cn).

power consumption of CMPs significantly. They also proposed a hybrid cache architecture composed of the SRAM and STT-RAM technologies to improve the performance reduction problem due to long STT-RAM write latency. Wu et al. [8-9] proposed to use STT-RAM and PRAM to implement the lower-level cache. Two types of hybrid cache architectures (HCAs) are evaluated – inter-level and intra-level, in which NVMs are utilized either as the entire L3 cache or the slow-accessed region in L2 cache. Their experiments show that these HCAs can improve the energy efficiency and performance. Li et al. [10] integrated SRAM with STT-RAM to construct a novel hybrid cache architecture for CMPs, and they also proposed energy-aware read and write mechanisms for the hybrid cache to improve performance for workloads with different write patterns. Lee et al. [11] considered the L1 SRAM cache and external memory to analyze hybrid cache architecture, and models for average memory access time, power consumption, and area of cache memory were proposed to compare various cases adopting different memory types and benchmark programs. A hybrid non-uniform cache architecture (NUCA) by combining SRAMs and STT-RAMs with different operating voltage/pulse width settings was proposed in [12]. However, these above work do not consider the optimal placement issue of NVMs and SRAM banks in the hybrid cache architecture of 3D CMP.

There have also been related work investigating cache access issues in hybrid cache architectures. Lin et al. [13] proposed to use hybrid cache partitioning to improve write operations in NUCA architectures. In the proposed design, each partition of the L2 cache contains four regions, one for SRAM Bank and three for STT-RAM Bank, based on the number of write operations and the way each processor core accesses for SRAM and STT-RAM. Two access-aware policies were proposed to reduce the write pressure and balance the write distribution on STT-RAM regions. Chen et al. [14] proposed a dynamically reconfigurable hybrid cache architecture for the last-level cache of processors. In this architecture, hit counters are added in the cache structure to dynamically adjust the ratio of NVM and SRAM. However, the proposed dynamically configurable hybrid cache architecture may result in data migration jitter. Ahn et al. [15] introduced the concept of write-intensity prediction to optimize the hybrid cache architecture. The work shows that the multiprocessor has a high probability of frequent write operations on the cache when running. Following these instructions, it is possible to predict which of the cache banks have greater write intensity in the next operation. In their study, the decision threshold for the cache bank that is most likely to perform data migration is difficult to determine. Low thresholds increase the chance of unnecessary migration, and high thresholds reduce the significance of migration. Li et al. [16] proposed two compilation-based approaches to improve the energy efficiency and performance of STT-RAM-based hybrid cache by reducing the migration overheads. This work uses compilation-based techniques, while our paper focuses on the architectural level technologies to improve the efficiency of hybrid caches.

Based on the problems presented above, this paper studies hybrid cache architecture for 3D CMP based on STT-RAM and proposes a hybrid cache spherical placement and an optimized hybrid cache dynamic migration scheme to minimize the power consumption generated by the hybrid cache architecture and data migration.

III. The Hybrid Cache Architecture for 3D CMP

In a 3D CMP architecture, the processor layer is generally placed near the heat sink, on the bottom of the chip [17-18]. The processor core has private L1 cache, which consists of instruction cache and data cache. The cores are connected via network-on-chip (NoC). Above the processor layer, multiple level shared caches are stacked. As shown in Fig. 1, two layers of level 2 cache are stacked. Each cache layer has 16 cache banks, which are connected via NoC. Different layers are connected by through silicon vias (TSVs).

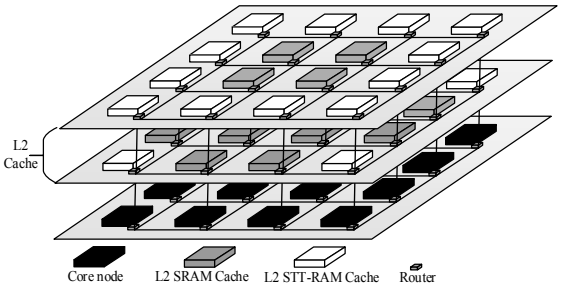

Fig.1. Hybrid cache architecture for 3D CMP

Without considering the different memory technologies of each bank, the local bank that is closer to the processor has shorter access latency due to shorter wire delay. This architecture is called NUCA [19]. The NUCA is also commonly adopted in 3D CMP. In NUCA-based cache, for a given core, the access to closer cache banks is much faster compared to the farther banks. One common way to improve system performance is to dynamically move frequently accessed data during runtime to closer banks. This technique is called dynamic NUCA (D-NUCA) [19]. Using D-NUCA in CMP, the access to the cache banks is not uniform. Most of the access operations occur on a small portion of the bank in the central of the cache layer [20], in order to reduce access time.

Therefore, we propose to integrate SRAM with STT-RAM to construct a shared hybrid L2 cache, according to the cache access characteristics with D-NUCA architecture. The main problem in the hybrid cache construction is to determine the placement of cache banks with the different memory technologies STT-RAM and SRAM. A placement example is given in Figure 1, the banks with white color represent the placements of STT-RAM cache bank, and the gray banks represent the placements of SRAM cache bank. The proposed placement approach for the hybrid cache architecture design will present in next section.

IV. HYBRID CACHE SPHERICAL PLACEMENT APPROACH

This section first analyzes the average access distance for cache banks in 3D CMP with D-NUCA architecture and then describes the proposed hybrid cache spherical placement approach.

A. Analysis of the cache bank average access distance

According to the cache access characteristics of D-NUCA mentioned above, the access to the cache banks is not uniform in a 3D CMP. As shown in Fig. 2, we take two 4×4 cache layers as an example to calculate the average access distance of each cache bank in a 3D CMP architecture. The average access distance for a cache bank is calculated by the average route hops from each processor core to this cache bank. It is found that cache banks with short access distances are distributed in the central area of this cache layer, the cache banks with large access distance are distributed in the edge of this cache layer. And in different cache layers, the average access distance is different. The average access distance in the bottom layer tends to be smaller than that in the top layer, because the bottom layer is near the processor layer. Therefore, the spatial distribution of cache banks with different access layers tends to be hemispherical.

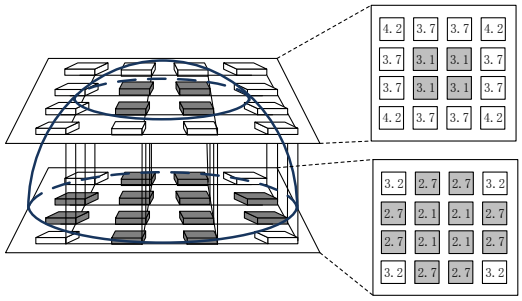

Fig.2. Cache bank average access distance in 3D CMP

B. Hybrid Cache Spherical Placement Approach

Based on the relation between average access distance and placement, we propose a hybrid cache spherical placement approach. The spherical placement approach is described as follows.

In a cache layer, the cache banks with short average access distance preferably use SRAM, and that with large average access distance use STT-RAM. The cache banks with short average access distance are distributed in the central of the cache layer. Therefore, the SRAM cache banks in a cache layer are distributed like a circle, and the radius of the circle is related to the distance between the cache layer and the processor layer. For the cache layer near the processor layer, the average access distance of the cache banks is smaller, so the radius of the SRAM distributed circle is large. As for the cache layer far away from the processor layer, the radius of the circle is small. In general, the spatial distribution of SRAM cache banks tends to be one hemisphere. The 3D CMP architecture with hybrid cache spherical placement design is shown in Fig. 3.

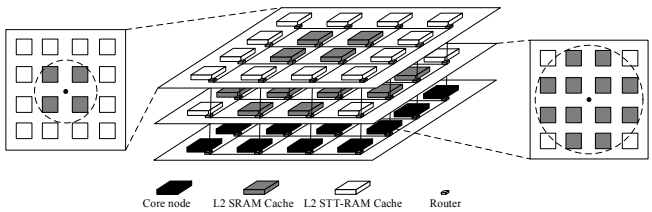

Fig.3. 3D CMP architecture with hybrid cache spherical placement design

The above hybrid cache spherical placement design takes into account the processor power consumption, NVM characteristics and the cache access characteristics of D-NUCA architecture.

Firstly, the hybrid cache design integrating STT-RAM with SRAM not only reduces the large static power consumption of SRAM, but also guarantees the system performance. In the 3D CMP with D-NUCA architecture, a small part of the cache banks in the middle of the cache layer carries nearly 50% of the processor's access requests, and the data migration of D-NUCA also generates more write operations in the middle of the cache layer. The write power consumption of STT-RAM is seven to eight times that of SRAM, and the write delay is more than six times that of SRAM. Therefore, if STT-RAM with the same capacity is used in the middle of the cache layer instead of SRAM, the performance of the system may be greatly degraded. Because of the high write power consumption of STT-RAM, the advantage of almost zero static power consumption may also be weakened in the environment of frequent write operations. Therefore, SRAM cache banks are located in the central area of the cache layer and STT-RAM cache bank are located in the edge of the cache layer using the proposed hybrid cache spherical placement strategy.

Secondly, the design of hybrid cache spherical placement also considers the characteristics of data migration in the 3D CMP with D-NUCA architecture. In 3D CMP, the processor layer is often located at the bottom of the chip due to heat dissipation issues. Therefore, for data migration in the vertical direction, only one direction is migrated towards the processor layer. The vertical latency of 3D CMP is relatively small, so if cache migrates from SRAM Bank far from the processor layer to STT-RAM Bank near the processor layer, the data access latency is even longer than before, such data migration is invalid. Therefore, the proposed spherical placement of hybrid cache avoids the case that the number of SRAM banks at the top is larger than the number of SRAM banks at the bottom, which further ensures the effectiveness of data migration in the hybrid cache architecture.

V. HYBRID CACHE DYNAMIC MIGRATION SCHEME OF 3D CMP

In the 3D CMP with our proposed spherical placement of hybrid cache, the traditional NUCA may cause data migration problems presented as follows.

A. Problems with data migration

1) Data migration jitter

In 3D CMP, different cores accessing the same cache line may generate two requests in two opposite direction for data migration. Fig. 4 shows the jitter problem with 3D CMP data migration. As shown in the figure, core 3 and core 15 share a cache line in bank 7. At time T1, the access frequency of core 15 to the shared cache line reaches the threshold of data migration, and according to the traditional data migration scheme, the shared cache line will be migrated from bank 7 to bank 11. At time T2, the access frequency of core 3 to the shared cache line also meets the data migration condition, and the shared cache line is relocated from bank 11 to bank 7. If both core 3 and core 15 have frequent access to the shared

cache line for a certain amount of time, the shared cache lines will migrate between bank 11 and bank 7 frequently. This frequent migration phenomenon of a cache line in the two opposite directions between the banks, called the data migration jitter.

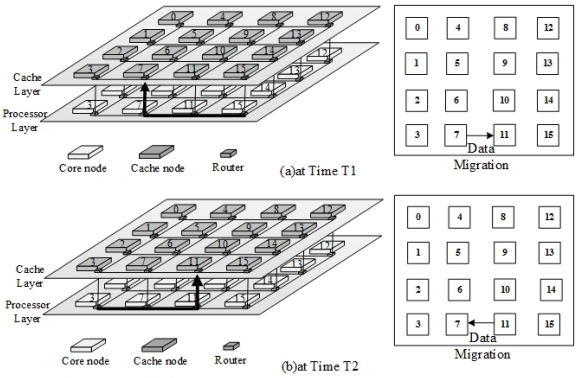

Fig.4. CMP data migration jitter

For the entire CMP, when data migration jitter occurs, the data migration will not reduce the average data access latency, and itself will produce additional power consumption overhead. Frequent data migration increases the amount of data on the on-chip network, so data migration jitter problems have a negative impact on the power and performance of CMP.

The problem of data migration jitter is mainly due to frequent data migration in two opposite directions. Therefore, increasing the threshold of data migration may reduce the frequency of data migration. However, simply increasing the threshold of data migration will weaken the effect of data migration on system performance. Therefore, how to reduce the frequency of data migration in two opposite directions while ensuring normal data migration is important in solving the data migration jitter problem.

2) Hybrid cache data migration failure

In hybrid cache architectures, the latency of data access depends on the physical distance between the processor core and the cache bank holding the data, and the type of cache bank holding the data. For example, in our proposed architecture, if data migrates from STT-RAM to SRAM, the average access distance and write latency will both be reduced. However, when migrating data originally stored in the SRAM to the STT-RAM, the write latency will increase even though the average access distance is reduced.

In the 2GHz system, the latency caused by the routing of the on-chip network is about 2 clock cycles. The read latency of SRAM and STT-RAM is about 6 clock cycles. The write latency of STT-RAM is more about 36 clock cycles. When data migrates from SRAM to STT-RAM, the average access distance is shorted, but the write latency of 30 clock cycles is caused. Therefore, in this case the data migration of the hybrid cache architecture loses its significance.

Due to the data migration from SRAM to STT-RAM, the traditional data migration scheme in hybrid cache architecture can not result in the expected improvement in system performance.

B. Hybrid Cache Data Migration Scheme

Based on the above mentioned problems, we propose an optimized scheme for data migration in 3D CMP hybrid cache architecture.

The scheme proposed in this paper assumed that the priority of data migration in different directions is: X direction> Y direction> Z direction. If the data migration priority in the Z direction is set to high, all the data that needs to be migrated will be migrated to the cache layer at the bottom layer near the processor layer. Thus, the data transmission in the Z direction will become crowded. Besides, the data in the bottom cache layer is often the most frequently accessed data, and replacing the data in the bottom layer will make the cache performance decreased.

In addition, the migration of data from SRAM to STT-RAM is forbidden. Based on the hybrid cache spherical placement proposed in this paper, the STT-RAMs are located at the edge of the cache layer and are spatially distributed outside the SRAM. If data migrates from SRAM to STT-RAM, it is likely to increase the average access distance of the cache bank.

In order to realize the proposed data migration scheme, we modify the traditional structure of the cache line. As shown in Fig. 5, each cache line added the X and Y two bits to preserve the previous migration information. The X flag bit is used to record the data migration in the west and east directions, and the Y flag bit is used to record the data migration in the north and south directions. If X is a ‘0’ , it means that data has not been migrated in the west and east directions or has been migrated to the west direction in the existing cache bank. If X is a ‘1’, it means that data has been migrated to the existing cache bank in east direction. If Y is a ‘0’, it indicates that data has not been migrated to north and south migration or data has been migrated to the north direction in the existing cache bank. And if Y is a ‘1’, it indicates that the data has been migrated in the south direction to the existing cache bank.

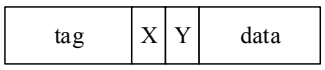

Fig.5. Modified cache line structure

In the proposed hybrid cache data migration scheme, when the cache line receives the migration request, it judges the state of the last data migration. If the migration request is in the same direction as the last migration, then the data migrates. If the direction of the migration request is different from the direction of the last data migration, the corresponding flag bit is modified but the current data migration is refused. We take the data migration between west and east direction request as an example. If the cache line receives a data migration request from the west direction, it will judge the state of the flag bit X at this moment. If the value of X is a ‘0’ at this moment, the data migration is allowed directly. If the value of X is a ‘1’ at this time, the value of X is rewritten to a ‘0’, but no data migration is performed.

VI. EXPERIMENTAL RESULTS

In this section, we present the evaluation of the proposed hybrid cache architecture with the spherical placement and data migration scheme. First, the evaluation methodology is

illustrated. Second, we present the experimental results and the analysis.

A. Experiment Setup

In this paper, a gem5-based simulation platform [21] is used. Experiments are conducted using the Spec2006 benchmark suit [22] to obtain the simulation results of the power consumption and performance of the 3D CMP. We simulate the novel 3D CMP architecture with the hybrid cache placement and data migration scheme proposed in this paper, compared with the traditional 3D CMP architecture. Table I describes the main simulation parameters applied in our experiments.

TABLE I
SIMULATION PARAMETERS

Parameter	Vaule
Core	16 processor cores, 2GHz, ALPHA ,4 × 4 Mesh
L1 Caches	Private, instruction and data cache, 32KB per core, 2-way set associative, 64 B block size, LRU replacement, 2-cycle latency
L2 Cache	Shared, 16-way set associative, 64 B block size, LRU replacement, 6-cycle latency for SRAM,36-cycle latency for STT-RAM
Cache Coherence Protocol	MESI
Main Memory	4G, 300-cycle latency

We model a 16-core 3D CMP system with two-level on-chip cache hierarchy similar to Fig.1. Each core has private L1 instruction and data caches. The capacity of L1 cache is 16×32KB SRAM.

We compare the proposed hybrid cache architecture with the spherical placement and data migration scheme to three experimental groups as shown in Table II. The experimental group 1 and the experimental group 4 both use SRAM as the L2 cache, but the capacity of the cache is different. In the experimental group 1, one layer of 16MB cache layer is stacked on the processor layer, and the experimental group 4 is stacked two 16MB cache layer. The experimental group1 and 4 are compared to evaluate the impact of different cache capacity on the system power consumption and performance. L2 cache in experimental group 2 is a hybrid structure of SRAM and STT-RAM, in which SRAM is located on one layer and STT-RAM is located on the other layer. Experimental group 3 uses the proposed hybrid L2 cache architecture with the spherical placement and the optimized data migration scheme. Experimental group 2 and 3 are compared to evaluate the impact of different hybrid cache architectures on the system power consumption and performance.

We use a set of SPEC2006 benchmarks for multi-programmed workloads, and the characteristics of the benchmark applications are shown in Table III.

B. Result Analysis

We first evaluate the impact of the proposed data migration scheme in 3D CMP with D-NUCA architecture

TABLE II
EXPERIMENTAL GROUP SIMULATION PARAMETERS CONFIGURATION

Parameter	SRAM	STT-RAM	Cache Coherence Protocol	Cache Hierarchy
Group 1	16 MB, 16 Bank	-	D-NUCA	4×4 Mesh
Group 2	16 MB, 16 Bank	16 MB, 16 Bank	D-NUCA	4×4×2 Mesh
Group 3	16 MB, 16 Bank	16 MB, 16 Bank	Optimized D-NUCA	4×4×2 Mesh, Spherical placement
Group 4	32MB, 16Bank×2	-	D-NUCA	4×4×2 Mesh

TABLE III
THE CHARACTERISTICS OF THE BENCHMARK APPLICATIONS

Application	Read operation ratio	Write operation ratio	Type of Application
bzip2	86.2%	13.8%	compute-intensive
libquantum	100%	0%	cache-intensive
hmmer	63.6%	36.4%	compute-intensive
lbm	15.7%	84.3%	cache-intensive
mcf	94.5%	5.5%	compute-cache -balance

(referred to as optimized D-NUCA the figure), compared with the traditional data migration scheme (referred to as D-NUCA in the figure). The system configuration used in the simulation is shown in Table I. Fig. 6 shows a comparison of the normalized system performance. The performance is measured by Instruction Per Clock (IPC). The IPC of the 3D CMP with the proposed optimized D-NUCA architecture is improved by 1.87% on average, and its system performance is improved by 4.43% for lbm benchmark. The main reason is that the optimized D-NUCA solves the problem of data migration failure in hybrid cache architecture. Especially for benchmarks with high proportion of write operations, which account for 84.3% of the total number of accesses, the benefit of storing more data in SRAM Bank is much greater than that of reducing data access distance.

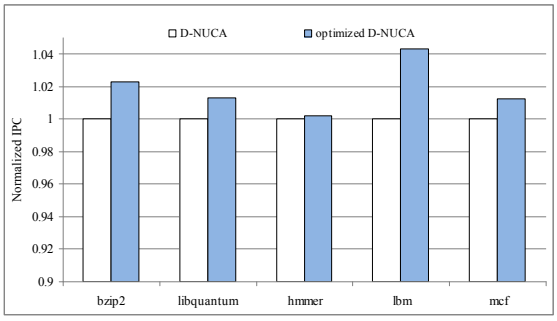

Fig.6. Performance comparison for different data migration scheme

Fig. 7 shows the normalized system power consumption comparison. The system power consumption of the 3D CMP with the proposed optimized D-NUCA architecture decreases by 2.07% on average. The main reason is that the optimized D-NUCA eliminates the unnecessary data migration in the traditional migration scheme. Eliminating unnecessary data migration not only reduces the power consumption generated by cache read-write operation, but also reduces the power consumption of data transmission through NoC routers and links.

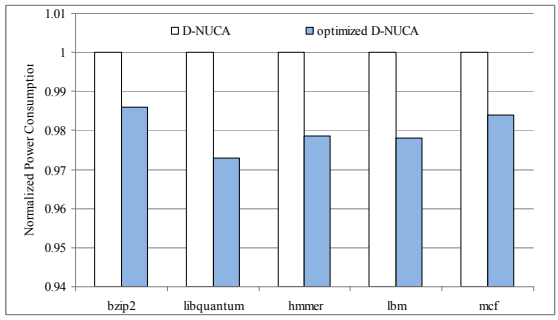

Fig.7. Power consumption comparison for different data migration scheme

The detailed evaluation results of our proposed hybrid cache architecture are shown in Fig. 8, Fig. 9 and Fig. 10. Fig. 8 shows a comparison of the normalized system performance with IPC. The experimental group 3 on average achieves 11.12% performance improvements than that of experimental group 1, 2.38% improvements than that of experimental group 2 and 1.49% degradation than that of experimental group 4. For compute-intensive applications such as bzip2 and hmmer, the performance of the system is mainly due to the speed of the processor, so increasing the capacity of the cache does not significantly improve the system performance. While the libquantum and lbm two programs are cache-intensive applications, the program will produce a large number of read and write operations, and the increase in cache capacity will reduce the data access latency. Therefore, with the benchmark libquantum and lbm, the system performance is greatly improved.

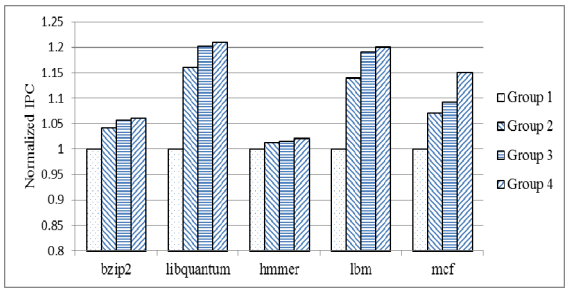

Fig.8. Normalized system performance comparison

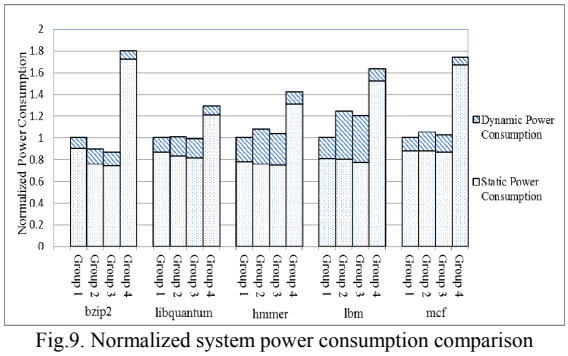

Fig.9. Normalized system power consumption comparison

Fig. 9 shows the normalized system power consumption comparison. The leakage power consumption of experimental group 2 and the experimental group 3 is relatively low, because these two experimental groups all use STT-RAM as cache. When running two cache-intensive programs of hmmer and lbm frequently, the dynamic power consumption of the overall power consumption is increasing due to the high write power consumption of STT-RAM. The experimental group 3 on average increases 2.53% power consumption than that of experimental group 1, but achieves 2.77% energy saving than that of experimental group 2, and 34.94% energy saving than that of experimental group 4.

Fig. 10 shows the system performance-power consumption ratio of different experimental groups. By the analysis of system performance-power ratio, we can evaluate the performance gains with increasing the power consumption of the different cache architectures. The experimental results show that the system using our proposed hybrid cache architecture has shown a better performance-power consumption ratio in most cases.

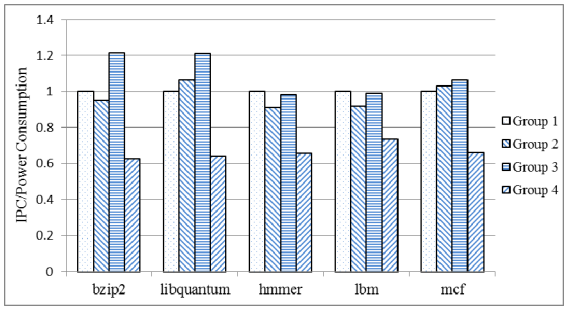

Fig.10. System performance- power consumption ratio comparison

VII. CONCLUSION

In this paper, we proposed a novel hybrid cache architecture composed of SRAM and STT-RAM for 3D CMP to reduce system power consumption. For the hybrid cache architecture design, we proposed a spherical placement approach to determine the optimal placement of STT-RAM and SRAM cache banks. And an optimized hybrid cache dynamic migration scheme is also proposed to solve the problem caused by traditional data migration scheme in 3D CMP. The experiments carried out with SPEC2006 benchmarks show that the proposed hybrid cache architecture with spherical placement and optimized data migration scheme are able to reduce power consumption by 34.94% on average with only 1.49% performance degradation, compared with the architecture which uses pure SRAM as the cache in the same capacity.

REFERENCES

[1] W. Haensch, "Why should we do 3D integration?" in *Proc. 45th ACM/IEEE Design Automation Conference*, Anaheim, 2008, pp. 674-675.

[2] Y. Xie, G. H. Loh, B. Black, and K. Bernstein, "Design space exploration for 3D architectures," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 2, no. 2, pp. 65-103, Apr. 2006.

[3] G. H. Loh and Y. Xie, "3D Stacked Microprocessor: Are We There Yet?" *IEEE Micro*, vol.30, no. 3, pp. 60-64, 2010.

[4] K. N. Chen and C. S. Tan, "Integration schemes and enabling technologies for three-dimensional integrated circuits," *IET Computers & Digital Techniques*, vol. 5, no.3, pp. 160-168, 2011.

[5] Y. Xie, "Modeling, architecture, and applications for emerging memory technologies," *IEEE Computer Design and Test*, vol.28, no.1, pp. 44-51, 2011.

[6] L. Wang, F. Ge, H. Lu, N. Wu, Y. Zhang, and F. Zhou, "A Spherical Placement and Migration Scheme for a STT-RAM Based Hybrid Cache in 3D chip Multi-processors," *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2018*, 4-6 July, 2018, London, U.K., pp.236-240.

[7] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen, "A novel architecture of the 3D stacked MRAM L2 cache for CMPs," in *Proc. IEEE 15th Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2009, pp. 239-249.

[8] X. Wu, J. Li, L. Zhang, E. Speight, R. Rajamony, and Y. Xie, "Hybrid Cache Architecture with Disparate Memory Technologies," in *Proc. ISCA*, 2009, pp. 34-45.

[9] X. Wu, J. Li, L. Zhang, E. Speight, and Y. Xie, "Power and performance of read-write aware hybrid caches with non-volatile memories," in *Proc. Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Apr. 2009, pp. 737-742.

[10] J. Li, C. J. Xue, and Y. Xu, "STT-RAM based energy-efficiency hybrid cache for CMPs," in *Proc. IEEE/IFIP 19th Int. Conf. VLSI Syst.-Chip (VLSI-SoC)*, Oct. 2011, pp. 31-36.

[11] S. Lee, J. Jung, and C.-M. Kyung, "Hybrid cache architecture replacing SRAM cache with future memory technology," in *Proc. IEEE Int. Symp.Circuits Syst. (ISCAS)*, May 2012, pp. 2481-2484.

[12] R. J. Behrouz and H. Homayoun, "NVP: Non-uniform Voltage and Pulse width Settings for Power Efficient Hybrid STT-RAM", in *Proc. IEEE international Green Computing Conference*, 2014, pp. 1-6.

[13] C. Lin, J.-N. Chiou, "High-Endurance Hybrid Cache Design in CMP Architecture with Cache Partitioning and Access-Aware Policies," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, pp. 2149-2161, 2015.

[14] Y. Chen, J. Cong, H. Huang, B. Liu, C. Liu, M. Potkonjak, and G. Reinman, "Dynamically reconfigurable hybrid cache: An energy-efficient last level cache design," in *Proc. Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Piscataway, 2012, pp.12-16.

[15] J. Ahn, S. Yoo, K. Choi, "Write intensity prediction for energy-efficient non-volatile caches," in P*roc. IEEE Int Symp on Low Power Electronics and Design (ISLPED)*, Piscataway, 2013, pp.223-228

[16] Q. Li, J. Li, L. Shi, M. Zhao, C. J. Xue, Y. He, "Compiler-assisted STTRAM-based hybrid cache for energy efficient embedded systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol.22, pp. 829-1840, 2014.

[17] X. Zhou, Y. Xu, Y. Du, Y. Zhang, and J. Yang, "Thermal Management for 3D Processors via Task Scheduling," in *Proc. 37th International Conference on Parallel Processing*, Portland, Sept. 2008, pp.115-122.

[18] H. Wang, Y. Fu , T. Liu, and J. Wang, "Thermal management via task scheduling for 3D NoC based multi-processor," in *Proc. International SoC Design Conference*, Seoul , Nov. 2010, pp. 440-444.

[19] C. Kim, D. Burger, and S. W. Keckler, "An adaptive, non-uniform cache structure for wire-delay dominated on-chip caches," in *Proc. 10th Int. Conf. Archit. Support Program. Lang. Oper. Syst. (ASPLOS)*, Oct. 2002, pp. 211-222.

[20] B. M. Beckmann, D. A. Wood, "Managing Wire Delay in Large Chip-Multiprocessor Caches," in *Proc.International Symposium on Microarchitecture*, 2004, pp.319-330.

[21] N. Binkert et al., "The gem5 simulator," *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1-7, May 2011.

[22] J. L. Henning, "SPEC CPU2006 benchmark descriptions," *ACM SIGARCH Comput.* Archit. *News*, vol. 34, no. 4, pp. 1-17, 2006.