Thermosiphon: A Thermal Aware NUCA Architecture for Write Energy Reduction of the STT-MRAM based LLCs

Bi Wu[1,2], Yuanqing Cheng[1,2], Pengcheng Dai[1,2], Jianlei Yang[1,3], Youguang Zhang[1,2]
Dijun Liu[4], Ying Wang[5] and Weisheng Zhao[1,2]
Fert Beijing Institute, BDBC, Beihang University, Beijing, China[1]
School of Electronic and Information Engineering, Beihang University, Beijing, China[2]
School of Computer Science and Engineering, Beihang University, Beijing, China[3]
China Academy of Telecommunication Technology(CATT), Beijing, China[4]
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China[5]
{bi.wu, yuanqing, weisheng.zhao}@buaa.edu.cn, wangying2009@ict.ac.cn

ABSTRACT

As the speed gap of the modern processor and the off-chip main memory enlarges, on-chip cache capacity increases to sustain the performance scaling. As a result, the cache power occupies a large portion of the total power budget. STT-MRAM (Spin Transfer Torque Magnetic Memory) is proposed as a promising solution for the low power cache design due to its high integration density and ultra-low leakage. Nevertheless, the high write power and latency of STT-MRAM become new barriers for the commercialization of this emerging technology. In this paper, we investigate the thermal effect on the access performance of STT-MRAM and observe that the temperature can affect the write delay and energy significantly. Then, we explore the NUCA (Non-Uniform Cache Access) design of the CMPs (Chip-Multi-Processors)with STT-MRAM based LLC (Last Level Cache). A thermal aware data migration policy, called "Thermosiphon", which takes advantage of the thermal property of STT-MRAM, is proposed to reduce the LLC write energy. This policy splits the LLC into different regions based on the thermal distribution and adaptively migrate write intensive data considering the temperature gradient among different thermal regions. Compared to the conventional NUCA design, our proposed design can save 22.5% write energy with negligible hardware overhead.

1. INTRODUCTION

Memory bandwidth instead of CPU processing speed has become a severe bottleneck of modern computing systems for further performance scaling, especially when entering into the many core era [12]. As a result, a large shared last level cache is beneficial and thought indispensable to overcome the "memory wall" issue. For example, the Intel Xeon-Phi deploys as large as 30MB on-chip L2 cache [2]. It is expected that more cache will be integrated on-chip as core count and working sets of applications continuously increase. Therefore, on-chip cache organization and interconnection are vital to sustain high memory bandwidth and reasonable access latency.

However, the growing array size increases the worst-case load of bitlines/wordlines and deteriorates the cache performance in the uniform access designs. Organizing on-chip cache, especially LLC cache, into many banks and connecting them with NoCs (Network-on-Chips) is an effective way to improve performance of multi-core and many-core processors [15]. Different from conventional UCA (Uniform Cache Access) architecture, LLCs of CMPs are commonly designed with NUCA architecture, which can be classified as S-NUCA (Static NUCA) and D-NUCA(Dynamic NUCA) [5].

In addition, the growing cache capacity makes conventional SRAM based cache suffer from the severe leakage power in the deep sub-micron regime [24]. To deal with this problem, several emerging non-volatile memory (NVM) technologies, such as PCRAM [18], ReRAM [17], and STT-MRAM [4], are proposed as alternatives to conventional SRAM for the future memory design. Among them, STT-MRAM has fast access speed, high endurance and process compatibility with CMOS technology, and has become a competitive candidate for LLC design [11]. Since STT-MRAM write operation is a time and energy consuming procedure, many techniques are proposed to reduce the write energy [16].

Besides all these well-known advantages of STT-MRAM, we also found that their thermal characteristics are worth studying and exploitable for energy-efficient cache operation, which seldomly be addressed in prior literatures. As a common concern, a surge in the power consumption makes the thermal issue a challenging problem for multi-core or many-core processor design. The induced high temperature and severe thermal gradient threaten the chip reliability and aggravates leakage power significantly. Unfortunately, the thermal issue is either completely ignored in the traditional SRAM-based NUCA design or taken as an undesirable characteristic that sometimes necessitates the power-gating of active cache banks for data integrity gurantee. However, compared to its SRAM counterpart that rejects high temperature, STT-MRAM has a unique thermal property: with the temperature rising, the write latency and energy decrease significantly. According to our analysis, the on-chip temperature non-uniformity in common multi-core processors are able to cause sufficiently large operation performance/power

978-1-5386-3093-8/17/$31.00 ©2017 IEEE

gradient among the STT-MRAM cells in different regions of the chip. Noting that the write operation of STT-MRAM is a energy hungry and time-consuming process, the thermal property of STT-MRAM and the on-chip temperature non-uniformity provide us an opportunity to improve STT-MRAM cache energy efficiency.

In this paper, we propose a novel thermal-aware NUCA design for the STT-MRAM based LLC, which is called "Thermosiphon"[1] While traditional NUCA exploits the non-uniformity in the access latency of distributed cache banks for performance gains, the proposed "Thermosiphon" architecture benefits from the spatial variation of STT-MRAM performance/power caused by the increasing temperature gradient. Due to the thermal awareness feature, it is expected to bring about two direct benefits. First of all, depending on the thermal distribution on-chip, the multi-bank Thermosiphon LLC are dynamically partitioned into the hot region and the cool region, and the cache banks in different thermal regions can self-adjust their write pulses appropriately instead of conforming to the worst-case write pulse setting, so that the average write performance and efficiency can be improved. In addition, with a thermal-aware data migration policy, Thermosiphon can actively herd the cache blocks that favor the high temperature to the locations where they can have better write performance/efficiency. As a result, we could improve the write latency and energy of STT-MRAM LLC without degrading the performance significantly. Our main contributions are as follows,

- We classify STT-MRAM LLC into different regions based on the thermal distribution on-chip, and each region has different write latency and energy in order to improve the write energy efficiency. To the best of our knowledge, it is the first work to exploit the on-chip thermal gradient to optimize STT-MRAM based LLC write energy efficiency.
- To effectively take advantage of the thermal property of STT-MRAM LLC, we propose a novel thermal aware NUCA design. Different data migration policies are adopted in different thermal regions so that most of the write operations can benefit from the performance improvement and energy reduction brought by high temperature without hurting the spatial locality in NUCA cache.
- The experimental results show that "Thermosiphon" can reduce the write energy by 22.5% on average compared to the conventional NUCA architecture design. Additionally, the performance can also be improved slightly due to the reduction of data swappings induced by data migration.

The rest of the paper is organized as follows. Section 2 presents the preliminaries of STT-MRAM and NUCA design. Section 3 describes the motivation of our work by examining the unbalanced thermal distribution on-chip and the thermal behavior of STT-MRAM access operation. Section 4 details the proposed thermal aware NUCA design and investigates the design tradeoffs. The experimental results are given in Section 5. Section 6 introduces the related work, and Section 7 concludes the paper.

[1]The design is named after a natural phenomenon "Thermosiphon" since our proposed NUCA design tries to promote the write intensive data from one region to another region, much like the fluid flowing with the thermosiphon effect

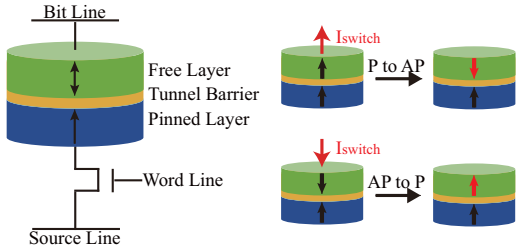

Figure 1: 1T1MTJ STT-MRAM cell structure.

2. PRELIMINARIES OF NUCA AND STT-MRAM

2.1 NUCA architecture design

The UCA structure, until very recently, is commonly adopted for on-chip cache design. It assumes that the delay of cache access is uniform which is determined by the delay to access the furthest bank. However, as the cache capacity increases, UCA induced cache access latency degrades cache performance dramatically and the NUCA (Non-Uniform Cache Architecture) is proposed to improve the cache performance [5]. In S-NUCA design, the data block position is fixed using some static address mapping scheme. In D-NUCA design, the data block can migrate from one bank to another bank within the same bankset. Therefore, D-NUCA can make sure that cores have quick accesses to the frequently used blocks and be more adaptive to memory behavior fluctuation compared to S-NUCA. Therefore, we target the D-NUCA based LLC design in the paper.

2.2 Introduction to STT-MRAM

STT-MRAM is one of the most promising candidates for the next generation memory technology because of its unique properties like fast access speed, extremely low standby power, high integration density, etc. [10]. Fig. 1 illustrates the commonly used cell structure consisting of 1 transistor(T) and 1 Magnetic Tunnel Junction(MTJ). MTJ is the data storage device in the memory cell. It is a sandwich-like structure with two ferromagnetic layers and one barrier in between. The MTJ can have different resistance value depending on its magnetization, which can be used to store data. In this work, we focus on the perpendicular MTJ(P-MTJ) since it has better scalability than its in-plane counterpart [25].

During the write operation, word line is enabled and a write voltage is applied between bit line and source line to generate the switching current to flip the MTJ state. According to the polarity of the switching current, a '0' or '1' can be written. As for the read operation, word line is enabled, and a read voltage is applied between bit line and source line. The MTJ state can be sensed by comparing sense currents flowing through the data cell and the reference cell. Then, a '0' or '1' can be read out by the sense amplifier.

3. MOTIVATION

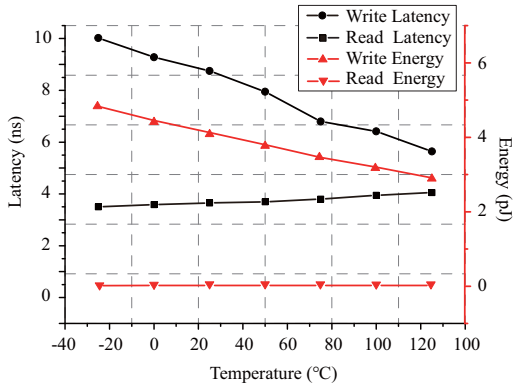

Figure 2: 1T1MTJ cell read/write latencies and energy consumptions under different temperatures.

3.1 Evaluation of the thermal effect on STT-MRAM access operation

As we have mentioned, with the chip power density increasing, on-chip temperature and thermal gradient elevate rapidly. It is necessary to investigate the thermal impact on STT-MRAM when concering both write performance improvement and energy efficiency. As temperature rises, the transistor driving current reduces due to degraded carrier mobility. Meanwhile, temperature can affect MTJ's switching probability, which can be expressed as the following formula [7]:

$$P = 1 - e^{-\frac{t_p}{\tau}} \tag{1}$$

where P is the switching probability of MTJ. t_p is the duration of the write voltage pulse. τ is computed as follows:

$$\tau = \tau_0 e^{\Delta(1-\frac{V}{V_{c0}})} \tag{2}$$

where τ_0 is the thermal attempt time at 0K, V is the magnitude of the applied voltage pulse, V_{c0} is the critical switching voltage of MTJ, and Δ is the energy barrier of MTJ which can be calculated with the formula [1]:

$$\Delta = \frac{H_K M_S}{k_B T} V_{ol} \tag{3}$$

Where V_{ol} is the MTJ volume, M_s is the saturation magnetization, k_B is the Boltzmann constant and T is the temperature. Formula (3) indicates that the thermal stability of MTJ decreases when temperature increases. In other words, it is easier to switch MTJ in higher temperature.

Based on the MTJ thermal model from [3], we can obtain the read/write latencies and energy consumptions under different temperatures with circuit simulations. As shown in Fig. 2, it indicates that the write energy and latency decrease rapidly as temperature increases. The write energy reduces from 4.3pJ at 0°C to 3.1pJ at 100°C (reduced by 27.9%), and the write latency reduces from 9.4ns at 0°C to 6.5ns at 100°C. Compared to the thermal effect on the write operation, temperature has negligible impact on the read operation. This thermal property provides us an opportunity to optimize the write energy of STT-MRAM.

3.2 Evaluation of the thermal distribution on-chip

As mentioned above, the continuous increasing of integration density on-chip escalates the "power wall" problem. The thermal issue is becoming an imminent challenge for the multi-processor design [19]. Moreover, when different tasks run on the CMP, the striking differences among running tasks result in significant thermal gradient across the whole chip.

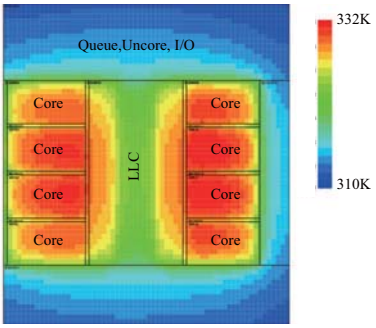

Figure 3: The thermal map of eight-core Intel Haswell architecture.

To illustrate this point, we perform thermal simulations for a CMP platform similar to Intel Haswell architecture [19] (refer to Section 5 for the thermal simulation setup.) and plot the thermal map in Fig. 3 when all eight cores running at the peak power. We can observe that the peak temperature on-chip can exceed 59°C while the temperature is roughly the ambient temperature (27°C in the paper)when all cores are idle. Therefore, the thermal gradient can exceed 30°C, which coinciding with the measurement from [6]. Due to the horizontal thermal propagation, severe thermal gradient can also be observed within the LLC region. Normally, the aggravating working temperature and thermal gradient are undesirable for SRAM-based cache as they can increase the leakage power and threaten the chip reliability. However, considering the thermal property of STT-MRAM write operation investigated in the above section, we can take advantage of the thermal gradient on-chip to reduce the write latency and energy of LLC.

4. THERMOSIPHON: A NOVEL THERMAL AWARE NUCA DESIGN FOR STT-MRAM BASED LLC

4.1 Introduction to the baseline case

First, we use an example to illustrate the commonly used data migration policy in existing D-NUCA architectures. As shown in Fig. 4(a), the CMP is assumed to have 8 cores with private SRAM L1 cache and shared STT-MRAM based L2 cache. The architecture is similar to Intel Haswell architecture. The large L2 cache, which acts as the LLC, is split into 64 banks and is implemented with STT-MRAM technology to reduce the leakage power. The cache banks are interconnected with a mesh NoC. 8 banks in a half-row constitute a bankset. After the block is firstly loaded into the initial entry according to the insertion policy, the data block can be migrated in the same bankset with a specific data migration policy [8].

Fig. 4(b) illustrates a widely used D-NUCA data migration policy, called "gradual promotion" [8]. For example, if a

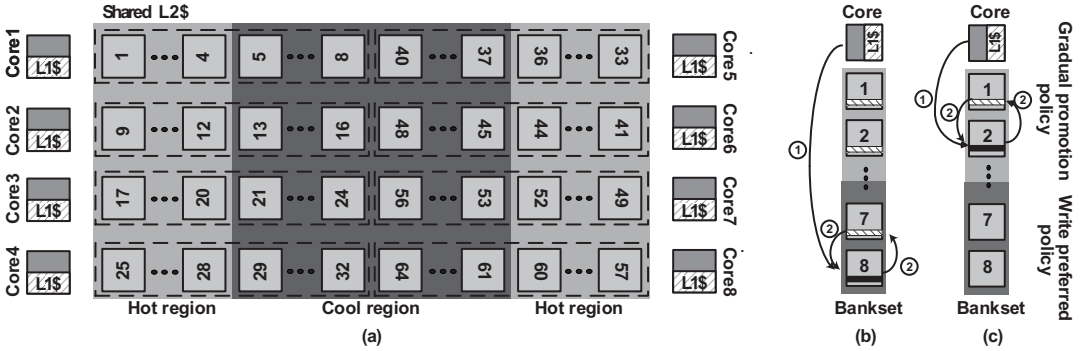

Figure 4: (a) An example of the NUCA architecture to illustrate data migration policies. (b) The "gradual promotion" policy [8]. (c)The data migration policy of "Thermosiphon" design.

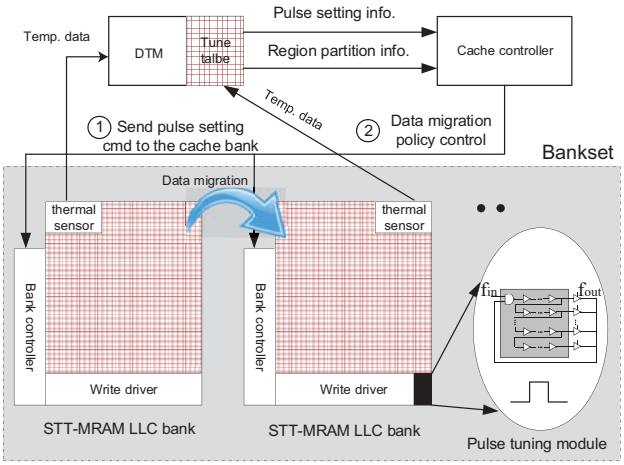

Figure 5: The overview of our proposed NUCA design.

data block in bank 8 is accessed, it will migrate towards the requesting core by one step within the bankset (i.e., migrate to bank 7). Consequently, the frequently accessed data may migrate to the neighborhood of the requesting core gradually. Each migration incurs one data swapping and associated write operations. In the paper, we adopt the "gradual promotion" as our baseline.

4.2 The overview of our proposed NUCA design

The working mechanism of the proposed NUCA design is illustrated in Fig. 5. In order to take advantage of the thermal gradient within the LLC, we can split it into different regions based on the temperature distribution. The thermal distribution information can be collected by thermal sensors embedded on-chip or through temperature prediction techniques widely used for dynamic thermal management (DTM) as shown in Fig 5. In the DTM module, there is a lookup table which stores the write pulse width and latency settings under different temperatures, which can be pre-characterized by prototype measurement. During the application running, the thermal information is fed to DTM and the DTM will search the lookup table to find the appropriate write pulse configuration for the cache bank in LLC. The LLC temperature information and the corresponding write pulse setting are then input to the cache controller. The controller can control the write operation and data migration based on the thermal information. Note that the number of entries in the lookup table determines the number of the write pulse tunning levels. Increasing the tuning levels complicates the bank controller and the write driver design. Therefore, we only consider to divide each bankset into two thermal regions, i.e., the hot region and the cool region. The worst-case write pulse is used in the cool region, and a shorter write pulse is used in the hot region according to the settings in the lookup table. The detailed write pulse tuning circuit design has been investigated in many literatures like [23], and is out of the scope of this paper.

4.3 An adaptive data migration policy in Thermosiphon

Through the thermal region partitioning, we expect that data blocks migrated into the hot region can obtain the write performance and energy benefits. Unfortunately, if we use the conventional "gradual promotion" policy directly, the thermal benefit we can reap may be very limited. The reason is analyzed as follows. Normally, the data read frequency is much higher than the write frequency. Moreover, since the read performance largely determines the cache access performance, it should be responded with a higher priority. As a result, the read intensive data may occupy banks in the hot region most of the time since hot banks usually locate in the neighborhood of active cores. Few write intensive data can have the opportunity to migrate into the hot region. Moreover, a data migration occurs in each cache access because the newly accessed data will migrate towards the requesting core by one step each time except that the cache block has already located in the nearest bank to the core. The migration introduces extra write overhead. To deal with these problems, it is necessary to propose a novel thermal-aware NUCA design to make more write operations benefit from the thermal gradient without degrading the read performance negatively.

To explore the tradeoff involved in the thermal-aware NUCA design, we consider two extremes firstly. Considering the performance optimization, one extreme is always promoting the touched data block to migrate towards the requesting core. It results in read intensive data clustering in the neigh-

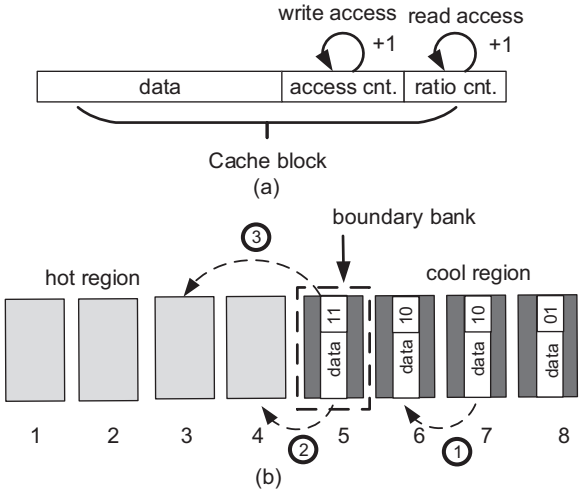

Figure 6: (a) The access counter and the ratio counter in the cache block. (b) The illustration of data migration in the cool region.

borhood of the core. The write operations will mostly happen in the cool region resulting higher write energy and latency. From the write energy reduction perspective, another extreme is to migrate the data block being written towards the requesting core with higher priority. Therefore, we can reap more thermal benefits by clustering the write-intensive data in the hot region. However, the read performance may degrade dramatically in this case. Thus, there is a trade-off between the two extremes for performance and energy optimizations. Our proposed thermal aware NUCA design explores to obtain the sweet point by adopting different data migration policies in different thermal regions.

In the hot region, to optimize the cache performance, read intensive data should be placed as near to the core as possible. For example, block 1 may be occupied by the read intensive data with a high possibility. Since the write latency and energy are roughly the same in the same thermal region[2], the write intensive block can still reap the thermal benefit without residing in block 1 as long as locates in the hot region (i.e., the light gray region in Fig. 4(a)). With the above analysis, we adopt the "gradual promotion" policy in the hot region to obtain better cache performance.

In the cool region, we propose a counter based data migration policy to promote the write intensive data towards the hot region with a higher probability. Each cache block has two counters, i.e., the access counter and the read/write ratio counter as shown in Fig 6(a). The former one indicates the access history of the cache block. Meanwhile, to elevate the possibility of the write intensive data to be migrated into the hot region, the the ratio counter is used to adjust the weight between a read and a write operation. For example, if the ratio counter is 3 bits, the counter will increase by 1 for each read access. If the counter overflows, it will increase the access counter by 1. On the other hand, the write access to the block will increase the access counter by 1 directly. Therefore, the net effect of the ratio counter is to set the priority of data migration for the write intensive data over the read intensive data. A higher ratio implies that the write intensive blocks have a higher probability to be migrated to the hot region. A side effect of the proposed policy is that data migration may not always incur the data swapping as long as the access counter does not change. Then, the write overhead induced by data migration can also be reduced accordingly.

With the access counter and the ratio counter, cache access in the cool region can be classified as the border bank access and the non-boundary bank access. The boundary bank is the nearest bank in the cool region to the hot region within the same bankset. For instance, bank 5 is a boundary bank in Fig. 6(b). As for the cache access in the non-boundary bank (case ① in Fig. 6(b)), the cache controller will read the access counter in the cache block, and compare it with the previous cache blocks[3]. The block will be migrated to the position next to the block, whose access counter value is larger than that of the block being accessed. For example, if an access hits bank 7 and the access counter value of the data block becomes 10, the block in bank 7 will swap with the block in bank 6. For the boundary block hit, the cache controller will try to find whether that block is just evicted from the hot-region or not. In the former case, the cache controller will make the data migration similar to the non-boundary block access (case ② in Fig. 6). If the boundary block is just evicted out from the hot region, the boundary block will migrate by one additional step, which locates in front of the bank in the boundary of the thermal region (case ③). The reason is that if the boundary block being considered migrates into the last position in the hot region (i.e., bank 4), it will evict the original block in that position, which may be just migrated into the hot region as well. By migrating towards the core one step further (i.e., migrate to bank 3), the undesirable ping-pong effect can be eliminated.

The access counter and the ratio counter need to be reset in two cases. First, when one of the access counters overflows, the access counter of each block will be reset, which is similar to the "pseudo-LRU" cache replacement counter. This policy is to keep the leading-edge block in the boundary position as the next migration candidate. Meanwhile, it gives the chance to other blocks to be migrated into the hot region. Second, when a new block is loaded into the LLC, all the block counters will be reset. Since the new block has the higher possibility to be visited again, the reseting can make the new block be migrated to the boundary block directly if the block is reused.

5. EXPERIMENT RESULTS

To validate the effectiveness of the proposed NUCA design in terms of write power and performance, we perform extensive simulations described as follows.

5.1 Experiment Setup

The CMP architecture used in our simulation consists of 8 Alpha 21264 cores as shown in Fig. 4(a). Each core has 32

[2]Note that although the write pulse width and amplitude are the same for the banks in the same thermal region, they are not exactly the same due to interconnect delay when accessing different banks. We have taken this point into account in the experiments of Section 5.

[3]The previous blocks locate in the banks nearer to the hot region and reside in the same bankset with the block being accessed.

Table 1: The CMP architecture configurations

Processor	8-core @ 3.3 GHz, out-of-order, alpha
L1-Cache	I-cache 32KByte 8-way set associative D-cache 32KByte 8-way set associative
L2-Cache	16 MByte, 8-way set associative, shared MOESI cache coherence protocol
Main Memory	4 GByte DDR3 DRAM
SPEC2000	Crafty, Galgel, Equake, Lucas Applu, Mesa, Mgrid, Fma3d Swim, Ammp, Apsi, Art Bzip2, Gzip, Gcc, crafty, Gap Vortex, Twolf, Vpr, Mgrid

KB private instruction and data cache respectively. The L2 cache is shared by all cores. The L2 cache banks are interconnected with a mesh NoC. We assume one cycle/hop of the interconnect delay. The detailed architectural setting is tabulated in Table 1. We extended the gem5 simulator [13] to model the proposed NUCA architecture. The "gradual promotion" is used as the baseline in our simulations. Another thermal aware NUCA design denoted as "T-NUCA", adopts "gradual promotion" policy in both hot and cool regions, but uses different write pulse settings in different thermal regions. These two scheme are used for comparisons in our experiments.

In order to obtain the thermal distribution of the CMP, we extracted the floorplan and power consumption information of the Intel Haswell processor which is very similar to our simulation target. Hotspot [19] is used for the thermal simulation. Based on the thermal simulation results, we assume that the temperature difference of the hot and the cool region is 30℃.

Table 2: MTJ parameters used in our simulations

Symbol	Value
diameter	40nm
Magnetic anisotropy	$5 \times 10^5 A/m$
Magnetic damping constant	0.03
Saturation magnetization	$3.68 \times 10^3 A/m$
Oxide barrier thickness	0.85nm
Free layer thickness	33.55nm
Gyromagnetic ratio	$1.76 \times 10^7 rad/(s \cdot T)$

The write energy and latency of L2 STT-MRAM cache are obtained by NVSim [21]. The L2 cache is 16MB, and the bank capacity is 256KB. The cell read/write energy and latency are obtained from the HSPICE simulations on the 40nm perpendicular STT-MRAM technology using the model developed in [3]. The MTJ parameters are shown in Table. 2.

SPEC2000 benchmark suite is used in the simulations. We construct 4 combinations of these benchmarks for evaluations as shown in Table. 3. Each group contains 8 benchmarks. We ran one instance of these workloads per core to simulate the multi-programmed case. 2 million instructions is fast-forwarded to warm up the cache and then ten million instructions are executed to generate the simulation statistics. The L2-cache access statistics acquired from gem5 are

Table 3: The benchmark groups constructed from the SPEC2000 benchmark suite

Group	Benchmark
int	Gzip, GCC, Crafty, Bzip2 Gap, Vortex, Vpr, Twolf
float	Ammp, Swim, Applu, Mesa Mgrid, Art, Apsi, Fma3d
Hybrid-1	Swim, Ammp, Applu, Mesa Gap, Vortex, Vpr, Twolf
Hybrid-2	Mgrid, Art, Apsi, Fma3d Gzip, GCC, Crafty, Bzip2

used to estimate the overall write energy consumption including the write operations caused by normal write accesses and data migrations.

5.2 Performance Analysis

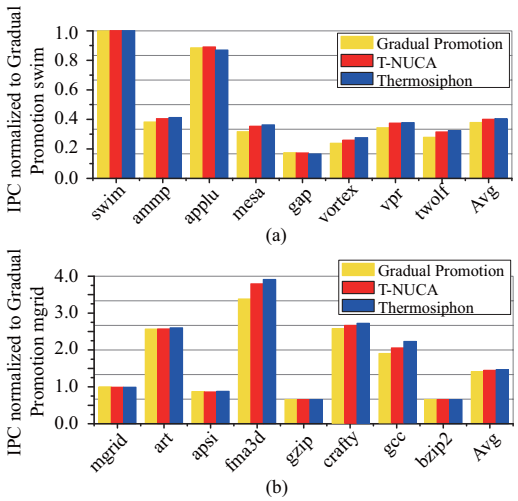

Figure 7: IPC comparisons of three different NUCA designs: the "gradual promotion" NUCA, "T-NUCA" and "Thermosiphon".

The IPC performance comparisons are plot in Fig. 7. The results are normalized to the baseline, i.e.,"gradual promotion" policy. As shown in Fig. 7(a), the left bars denote the IPC values of eight cores, and the rightmost bar represents the geometric mean of the measurements. Among the three NUCA designs, "Thermosiphon" performs the best in terms of IPC. On average,"T-NUCA" scheme can improve 5.8% compared to the baseline when running the hybrid-1 benchmark combination."Thermosiphon" can obtain 7% improvement compared to the baseline. As for the hybrid-2 benchmark group, "T-NUCA" can improve the IPC by 2.5%, and "Thermosiphon" can improve 3.9% IPC performance compared to the baseline. The reason is that "Thermosiphon" can provide more opportunities for the write intensive data block in the cool region to be migrated in the hot region. Therefore, the write latency can be reduced accordingly. Meanwhile, the read performance will degrade significantly since the "gradual promotion" policy is adopted in the hot region. Moreover, the ratio counter also reduces the data swapping as mentioned before, which also contributes to

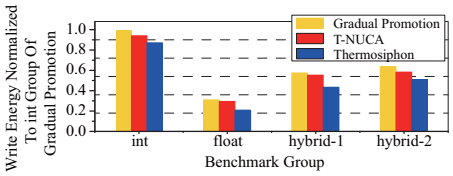

Figure 8: Write energy comparisons of three NUCA designs considered in the paper.

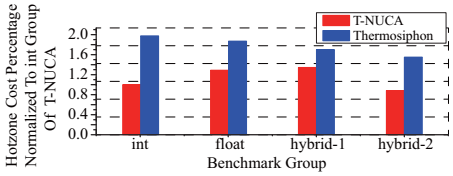

Figure 9: The comparisons of the write energy consumed in the hot region. It implies that more write operations occurring in the hot region for "Thermosiphon".

the write performance improvement. As for the other two benchmark groups, we obtain the similar results which are omitted due to the lack of space.

5.3 Write Energy Analysis

The write energy comparisons of three NUCA designs are presented in Fig. 8. The results are normalized to the baseline. As shown in the figure, "Thermosiphon" performs the best in terms of the write energy consumption. It can save 22.5% write energy on average while "T-NUCA" can reduce write energy by 5.7% compared to the baseline. The reason is that "Thermosiphon" can make a write intensive block to be migrated into the hot region with the higher probability. Considering the "T-NUCA" scheme, although the write operation in the hot region can obtain the thermal benefit, the energy reduction is limited because read intensive blocks will occupy the hot region most of the time.

To validate the effectiveness of our proposed design further, we analyze the total write energy consumption occurring in the hot region. The write operation in the hot region can be divided into two categories. The first one is the normal write access (including data loading operations). The other one is the data swapping induced write operations. Due to the region partitioning, the two blocks involving in the data migration may locate in different thermal regions. By comparing the percentage of the write energy occurring in the hot region over the total write energy, we can verify whether more write operations can obtain the benefit by the data migration. As shown in Fig. 9, we can observe that the "Thermosiphon" scheme makes more write operations occur in the hot region compared to the "T-NUCA" scheme, which validates the effectiveness of our NUCA design.

5.4 The access counter and the ratio counter design space exploration

As we discussed previously, the number of bits of the access/ratio counter is an important parameter for our NUCA design. If the bit count of the access counter is too large, the overflow will occur too late to migrate the new data block into the hot region, which may result in access operations

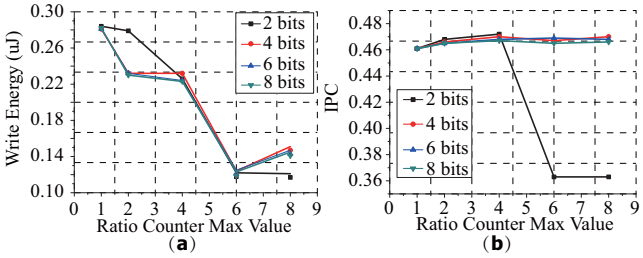

Figure 10: The design space exploration of the access/ratio counter(The legend represents the access counter bit) (a) from the energy consumption perspective. (b) from the IPC perspective.

mostly occurring in the cool region. On the other hand, If we set the bit count too small, the counter will be refreshed so frequently that the access characteristic of the cache block can not be captured. To explore the design space of the counter design, we obtained the IPC and the write energy with different counter configurations. As shown in Fig. 10, we observed that when the ratio is 6, which implies 6 read hits are equivalent to one write hit in the cool region, we can obtain the optimal write energy consumption. From the IPC perspective, we derived that the ratio 6 is optimal as well. Although other three counter configurations can also achieve the optimal IPC value, 3 bits for the ratio counter is the best choice considering the hardware overhead. Similarly, the optimal access counter bits should be 4 as shown in Fig. 10.

6. RELATED WORK

The on-chip multicore scaling are demanding a steady grow of on-chip cache capacity to bridge the gap between processor throughput and the off-chip memory bandwidth. Kim et al. resorted to non-uniform cache architecture (NUCA) cache design to address the worsened performance of growing UCA cache. [5]. After that a lot of researchers studying the data mapping, insertion and inter-bank migration strategy to fully exploit the potential of NUCA. For example, Beckmann and Wood first time considered the cache block migration and replacement policy in dynamic NUCA of multi-core systems [5]. Beside migration strategy, there are also some literatures on the optimal placement of cache blocks in order to harvest locality without inducing data moving overhead [14]. In contrast to dynamic NUCA that manages the deterministic performance variation caused by spatial locality, this work discover a new factor of thermal issue that causes performance non-uniformity in STT-MRAM cache and seeks to exploit this opportunity to increase cache efficiency.

Since STT-RAM has many advantages for working as future memory technologies, there are a lot of researchers trying to solve the issue of expensive write operation in STT-MRAM. Sun investigated spin current induced MTJ switching and described switching behavior by LandauLifshitz-Gilbert equation accurately making it a base for following compact model development [9]. Zhao et al. developed a model for thermally assisted MTJ [20]. Zhang et al. developed a compact model for perpendicular anisotropy MTJ [25]. The thermal issue in cache design is also an

important research topic in manycore era. In contrast to this work, prior studies on thermal-aware cache design are putting more emphasis on the task management and mapping policy [22]. These work avoid the formation of hotspot that causes severe leakage increase in cache by means of D-VFS, task migration or power gating [23].

7. CONCLUSION

As the modern processor enters into the multi-core and many core era, cache capacity increases rapidly and NUCA architecture is introduced for the cache performance improvement. To mitigate the leakage power, STT-MRAM based LLC cache is promising to replace the conventional SRAM cache. At the same time, The soaring power consumption due to high integration density introduces severe thermal issue on-chip. In the paper, we take advantage of the thermal property of STT-MRAM write operation to reduce the energy consumption in LLCs. With the thermal consideration, we propose a thermal aware NUCA design - "Thermosiphon". The experimental results show that compared to the baseline, our proposed NUCA design can improve the performance by 7% at most, and reduce the write energy by 22.5% on average with only 1.3% extra hardware overhead.

8. ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China (NSFC) under grant No. 61401008, No.61602022, No.61504153, the International Collaboration 111 Project from the Ministry of Education and Foreign Experts under grant No.B16001 and the State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, under grant No.CARCH201602. The corresponding authors are Weisheng Zhao and Ying Wang.

We would like to thank Prof. Guangyu Sun and Dr. Chao Zhang for helpful discussions and suggestions.

9. REFERENCES

[1] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. De. Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances. In *IEDM*, pages 1–4. IEEE, 2009.

[2] A. Sodani, R. Gramunt, J. Corbal, H. S. Kim, K. Vinod, S. Chinthamani,S. Hutsell, R. Agarwal and Y. C. Liu. Knights Landing: Second-Generation Intel Xeon Phi Product. *IEEE Micro*, 36(2):34–46, 2016.

[3] B. Wu, Y. Cheng, J. Yang, A. Todri-Sanial and W. Zhao. Temperature Impact Analysis and Access Reliability Enhancement for 1T1MTJ STT-RAM. *IEEE Transactions on Reliability*, 65(4):1755–1768, 2016.

[4] C. Chappert, A. Fert and F. N. Van Dau. The emergence of spin electronics in data storage. *Nature materials*, 6(11):813–823, 2007.

[5] C. Kim, D. Burger and S. W. Keckler. An Adaptive, Non-uniform Cache Structure for Wire-delay Dominated On-chip Caches. In *ASPLOS*, pages 211–222. ACM, 2002.

[6] F. Mesa-Martinez, E. Ardestani and J. Renau. Characterizing Processor Thermal Behavior. In *ASPLOS*, pages 193–204. ACM, 2010.

[7] J. Kan, K. Lee, M. Gottwald, S. H. Kang and E. E. Fullerton. Low-temperature magnetic characterization of optimum and etch-damaged in-plane magnetic tunnel junctions. *Journal of Applied Physics*, 114(11):114506, 2013.

[8] J. Lira, C. Molina, R. N. Rakvic and A. GonzClez. Replacement techniques for dynamic NUCA cache designs on CMPs. *The Journal of Supercomputing*, 64(2):548–579, 2013.

[9] J. Z. Sun. Spin-current interaction with a monodomain magnetic body: A model study. *Physical Review B*, 62(1):570, 2000.

[10] L. Song, Y. Wang, Y. Han, H. Li, Y. Cheng and X. Li. STT-RAM buffer design for Precision-Tunable General-Purpose Neural Network Accelerator. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 25(4):1285–1296, 2017.

[11] M. J. Mao, H. Li, A. K. Jones and Y. Chen. Coordinating prefetching and STT-RAM based last-level cache management for multicore systems. In *GLSVLSI*, pages 55–60. ACM, 2013.

[12] M. P. Jagtap. Era of multi-core processors. *Power*, 2:2, 2009.

[13] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna and S. Sardashti. The gem5 simulator. *ACM SIGARCH Computer Architecture News*, 39(2):1–7, 2011.

[14] N. Hardavellas, M. Ferdman, B. Falsafi and A. Ailamaki. Reactive nuca: near-optimal block placement and replication in distributed caches. *ACM SIGARCH Computer Architecture News*, 37(3):184–195, 2009.

[15] N. Muralimanohar, R. Balasubramonian and N. Jouppi. Optimizing NUCA organizations and wiring alternatives for large caches with CACTI 6.0. In *MICRO*, pages 3–14. IEEE Computer Society, 2007.

[16] P. Zhou, B. Zhao, J. Yang and Y. Zhang. Energy reduction for STT-RAM using early write termination. In *ICCAD)*, pages 264–268. IEEE, 2009.

[17] S. Gaba, P. Knag, Z. Y. Zhang and W. Lu. Memristive devices for stochastic computing. In *ISCAS)*, pages 2592–2595. IEEE, 2014.

[18] S. Raoux, G. W. Burr, M. J. Breitwisch, C. T. Rettner, Y. C. Chen, R. M. Shelby, M. Salinga, D. Krebs, S-H. Chen, H-L. Lung and others. Phase-change random access memory: A scalable technology. *IBM Journal of Research and Development*, 52(4.5):465–479, 2008.

[19] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron and M. Stan. HotSpot: A compact thermal modeling methodology for early-stage VLSI design. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 14(5):501–513, 2006.

[20] W. S. Zhao, J. Duval, J. O. Klein and C. Chappert. A compact model for magnetic tunnel junction (MTJ) switched by thermally assisted spin transfer torque (TAS+ STT). *Nanoscale research letters*, 6(1):368, 2011.

[21] X. Y. Dong, C. Xu, N. Jouppi and Y. Xie. NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 31(7):994–1007, 2012.

[22] X. Y. Zhou, J. Yang, Y. Xu, Y. T. Zhang and J. H. Zhao. Thermal aware task scheduling for 3D multicore processors. *IEEE Transactions on Parallel and Distributed Systems*, 21(1):60–71, 2010.

[23] Y. H. Wang, C. Zhang, H. Yu and W. Zhang. Design of low power 3D hybrid memory by non-volatile CBRAM-crossbar with block-level data-retention. In *ISLPED*, pages 197–202. ACM, 2012.

[24] Y. Wang, Y. Han, H. Li and X. Li. VANUCA: Enabling Near-Threshold Voltage Operation in Large-Capacity Cache. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(3):858–870, 2016.

[25] Y. Wang, Y. Zhang, E. Y. Deng, J. O. Klein, L. Naviner and W. S. Zhao. Compact model of magnetic tunnel junction with stochastic spin transfer torque switching for reliability analyses. *Microelectronics Reliability*, 54(9):1774–1778, 2014.