A Novel High Performance and Energy Efficient NUCA Architecture for STT-MRAM LLCs With Thermal Consideration

Bi Wu, *Student Member, IEEE*, Pengcheng Dai, Yuanqing Cheng, *Member, IEEE*, Ying Wang, *Member, IEEE*, Jianlei Yang, *Member, IEEE*, Zhaohao Wang, *Member, IEEE*, Dijun Liu, and Weisheng Zhao, *Fellow, IEEE*

***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. Spin transfer torque magnetic memory (STT-MRAM) is proposed as a promising solution for the low power cache design due to its high integration density and ultralow leakage power. 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 nonuniform cache access (NUCA) design of the chip-multiprocessors with STT-MRAM-based last level cache (LLC). 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 dynamically based on the thermal distribution monitored by thermal sensors available on-chip, and adaptively migrates write intensive data among different thermal regions considering the thermal gradient. Compared to the conventional NUCA design, our proposed design can save 41.2% write energy at most and 13.01% on average with negligible hardware overhead.**

***Index Terms*—Cache, data migration, low power, spin transfer torque magnetic memory (STT-MRAM), thermal gradient.**

Manuscript received June 28, 2018; revised October 7, 2018 and January 10, 2019; accepted January 28, 2019. Date of publication February 4, 2019; date of current version March 18, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61401008, Grant 61704005, and Grant 61571023, in part by the Beijing Natural Science Foundation under Grant 4192035, in part by the International Collaboration Project under Grant B16001, in part by the National Key Technology Program of China under Grant 2017ZX01032101, in part by the Special Foundation of Beijing Municipal Science and Technology Commission under Grant Z161100000216149, and in part by the State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences under Grant CARCH201602. This paper was recommended by Associate Editor H. Li. *(Bi Wu and Pengcheng Dai contributed equally to this work.) (Corresponding author: Weisheng Zhao.)*

B. Wu, P. Dai, Z. Wang, and W. Zhao are with the Fert Beijing Institute, BDBC, Beihang University, Beijing 100191, China, and also with the School of Microelectronics, Beihang University, Beijing 100191, China (e-mail: bi.wu@buaa.edu.cn; zhaohao.wang@buaa.edu.cn; weisheng.zhao@buaa.edu.cn).

Y. Cheng is with the School of Microelectronics, Beihang University, Beijing 100191, China (e-mail: yuanqing@ieee.org).

Y. Wang is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangying2009@ict.ac.cn).

J. Yang is with the Fert Beijing Institute, BDBC, Beihang University, Beijing 100191, China, and also with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: jerryyangs@gmail.com).

D. Liu is with the China Academy of Information and Communications Technology, Beijing 100191, China (e-mail: liudj@datanggroup.cn).

Digital Object Identifier 10.1109/TCAD.2019.2897707

I. 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 [1]. As a result, a large shared last level cache (LLC) is beneficial and thought indispensable to overcome the "memory wall" issue. For example, the Intel Xeon-Phi deploys as large as 30 MB 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 cache memory array size increases the worst-case capacitive loads of bitlines/wordlines, and deteriorates the cache performance. Thus, organizing on-chip cache, especially LLC cache, into many banks and connecting them with network-on-chips (NoCs) is an effective way to improve the performance of multicore and many-core processors [3], [4]. Different from conventional uniform cache access (UCA) architecture, LLCs of chip-multiprocessors (CMPs) are commonly designed with non-UCA (NUCA) architecture, which can be classified as static NUCA (S-NUCA) and dynamic NUCA (D-NUCA) [5].

In addition, the growing cache capacity makes conventional SRAM-based cache suffer from the severe leakage power in the deep submicrometer regime [6]. To deal with this problem, several emerging nonvolatile memory technologies, such as PCRAM [7], ReRAM [8], and Spin transfer torque magnetic memory (STT-MRAM) [9], are proposed as promising alternatives for future cache and main memory designs. Among them, STT-MRAM has fast access speed, high endurance, and process compatibility with CMOS technology, and has become a attractive candidate of LLC designs [10]. However, STT-MRAM write operation is a time and energy consuming

procedure, and many techniques are proposed to reduce the write energy [11], [12].

As a common concern, a surge in the power consumption makes the thermal issue a challenging problem of multicore or many-core processor designs. The induced high temperature and severe thermal gradient threaten the chip reliability and aggravate leakage power significantly. Unfortunately, the thermal issue is either completely ignored in the traditional SRAM-based NUCA design [13] or taken as an undesirable characteristic. However, compared to its SRAM counterpart, STT-MRAM has a unique thermal property: with the temperature rising, the write latency and energy decrease significantly. Note that the write operation of STT-MRAM is a energy hungry and time-consuming procedure. The thermal property of STT-MRAM and the on-chip thermal gradient 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 called "Thermosiphon."[1] The proposed Thermosiphon architecture exploits the spatial thermal variation of the STT-MRAM LLC for write latency and energy improvements. Specifically, first of all, depending on the thermal distribution on-chip, the multibank LLC can be dynamically partitioned into the hot region and the cool region, and cache banks in different thermal regions can self-adjust their write pulses appropriately instead of conforming to the conservative write pulse setting, so that the write performance and energy efficiency can be improved. Second, with a thermal-aware data migration policy, Thermosiphon can migrate data among thermal regions to approach optimal performance and energy consumptions. Our main contributions are as follows.

1) We divide the STT-MRAM LLC into different thermal regions according to the thermal distribution on-chip. Each region sets its cache write latency and energy appropriately for performance and energy efficiency. To the best of our knowledge, it is the first work to exploit the on-chip thermal gradient to optimize write energy and performance of STT-MRAM-based LLCs.
2) 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 write operations can benefit from performance improvements and energy reductions brought by the high temperature without hurting the spatial locality in the NUCA cache.
3) Experimental results on both single-core and multicore processors validate the effectiveness and efficiency of our proposed technique. Thermosiphon can reduce the write energy by 13.01% on average and 41.2% at most compared to the conventional NUCA architecture design. Additionally, the performance can also be improved by 22.7% at most due to the reduction of data swappings induced by data migration.

[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.

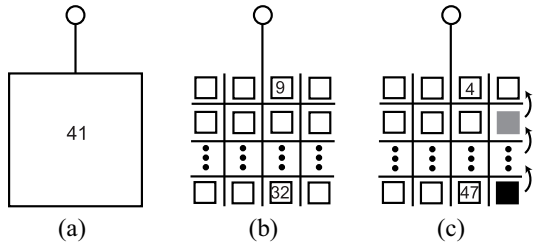


Fig. 1. Typical cache architectures (note that the numbers in the square specifies the access latencies of different cache banks) [5]. (a) UCA. (b) S-NUCA. (c) D-NUCA.

The rest of this paper is organized as follows. Section II presents preliminaries of STT-MRAM and NUCA design. Section III describes the motivation of this paper by examining the unbalanced thermal distribution on-chip and the thermal behavior of STT-MRAM access operation. Section IV details the proposed thermal aware NUCA design, and investigates the design tradeoff involved. The experimental results are given in Section V. Section VI presents the related work, and Section VII concludes this paper.

II. Preliminaries of NUCA and STT-MRAM

A. NUCA Architecture Design

Fig. 1 shows some commonly used cache organizations. The number of banks and the average access time for each kind of cache organizations are listed in the figure as well [5]. The UCA structure [shown in Fig. 1(a)], until very recently, is commonly adopted for on-chip cache design. It assumes that access latencies of different cache locations are uniform and determined by the delay to access the furthest bank. However, as the cache capacity increases, the cache access latency of UCA architecture degrades cache performance dramatically, and NUCA architecture is proposed to improve the cache performance [5]. NUCA design can be divided into S-NUCA and D-NUCA. In S-NUCA, which is shown in Fig. 1(b), the cache block position is fixed with some static address mapping scheme. The cache access latency depends on the distance of the cache bank from the core, which improves cache access performance effectively. However, this one-time data mapping can not explore the full potential of cache performance improvement. As a result, D-NUCA is proposed, which is shown in Fig. 1(c). In this case, the cache 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 their frequently used cache blocks, and can be more adaptive compared to S-NUCA.

B. 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. [14]. Fig. 2 illustrates the

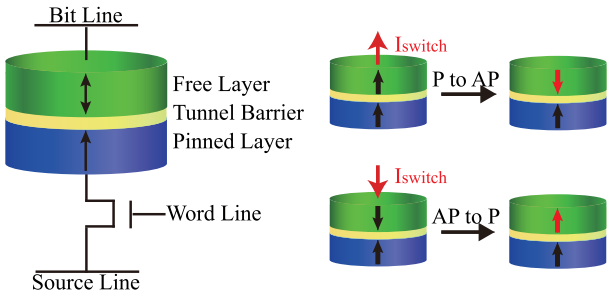


Fig. 2. 1T1MTJ STT-MRAM cell structure.

commonly used cell structure consisting of 1 transistor (T) and 1 magnetic tunnel junction (MTJ). MTJ is the data storage device in a memory cell. It is a sandwich-like structure with two ferromagnetic layers and one barrier in between. A MTJ has different resistances depending on its magnetizations, which can be used to store data. In this paper, we focus on the perpendicular MTJ since it has better scalability than its in-plane counterpart [15], [16].

During the write operation, the word line is enabled and a write voltage is applied between the bit line and the 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, the word line is enabled, and a read voltage is applied between the bit line and the 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.

III. Motivation

A. Evaluation of the Thermal Effect on the STT-MRAM Cache Access

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 from the write performance and the energy consumption perspectives. Both the access transistor and MTJ device can be significantly affected by the temperature variation on-chip. Since the read voltage is much smaller than the supply voltage, the transistor works at linear region during the read operation. Then, driving current of the access transistor can be calculated as [17]

$$I_{\rm ds} = \mu C_{\rm ox} \frac{W}{L}\left(V_{\rm gs} - V_{\rm th} - \frac{V_{\rm ds}}{2}\right)V_{\rm ds} \tag{1}$$

where $V_{\rm ds}$ is the voltage difference between drain and source of the access transistor. $V_{\rm gs}$ is the gate-source voltage. $V_{\rm th}$ is the threshold voltage. W is the transistor width. L is the channel length of transistor. $C_{\rm ox}$ is gate oxide capacitance per unit area. μ is the carrier mobility which can be calculated by the following expression:

$$\mu(T) = \mu(T_r)\left(\frac{T}{T_r}\right)^{-k_\mu} \tag{2}$$

where $\mu(T_r)$ is the carrier mobility at room temperature and k_μ is a fitting parameter. According to the above equations, we can derive that as the temperature increases, the driving current reduces due to the degraded carrier mobility.

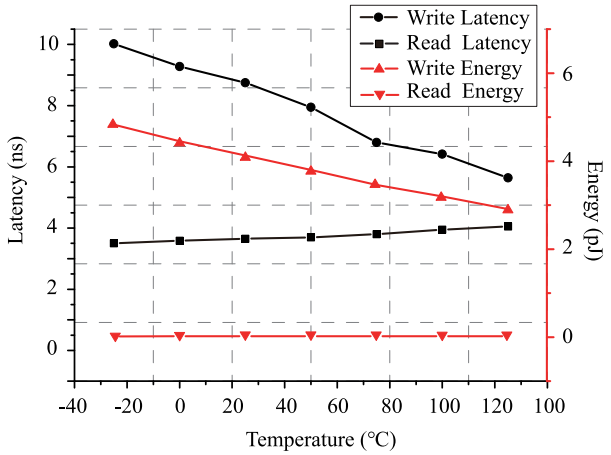


Fig. 3. 1T1MTJ cell read/write latencies and energy consumptions with different temperatures.

Meanwhile, temperature can affect an MTJ's switching probability, which can be expressed as [18]

$$P = 1 - e^{-\frac{t_p}{\tau}} \tag{3}$$

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

$$\tau = \tau_0 e^{\Delta\left(1-\frac{V}{V_{c0}}\right)} \tag{4}$$

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 the MTJ, and Δ is the energy barrier of the MTJ which can be calculated [19]

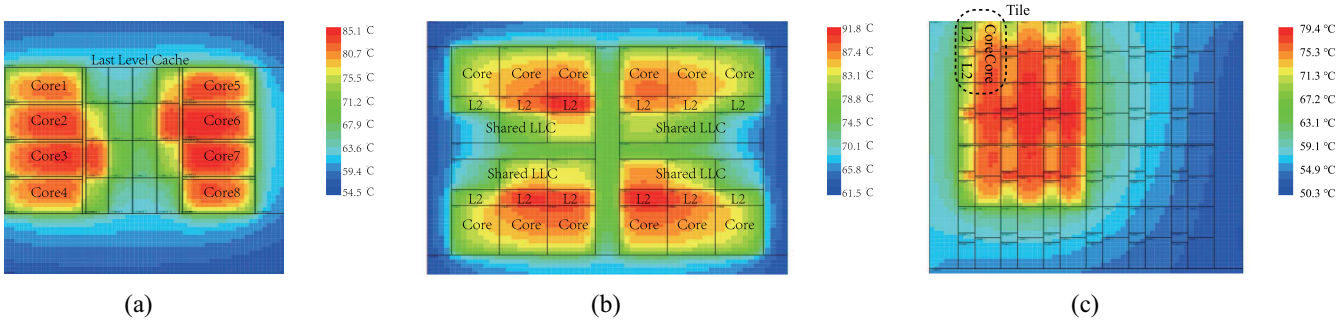
$$\Delta = \frac{H_K M_S}{k_B T} V_{ol} \tag{5}$$

where V_{ol} is the MTJ volume, M_s is the saturation magnetization, k_B is the Boltzmann constant, and T is the temperature. Equation (5) indicates that thermal stability of an MTJ decreases when temperature increases. In other words, it is easier to switch an MTJ in higher temperature.

Based on the MTJ thermal model from [20], we can obtain read/write latencies and energy consumptions under different temperature with SPICE simulations. As shown in Fig. 3, it indicates that the write energy and latency decrease rapidly as temperature increases. The write energy reduces from 4.3 pJ at 0 °C to 3.1 pJ at 100 °C (reduced by 27.9%), and the write latency reduces from 9.4 ns at 0 °C to 6.5 ns at 100 °C. Compared with 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 and performance of STT-MRAM.

B. 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 multiprocessor design [22]. Moreover, when different



(a) (b) (c)

Fig. 4. Thermal map of: (a) Intel Haswell architecture; (b) IBM POWER8 architecture; and (c) 48-core IA-32 processor.

TABLE I
CONFIGURATIONS OF PLATFORM ADOPTED IN THERMAL SIMULATIONS

Processor	Cores	Cache Levels	Sim Target	Peak Power
IBM POWER8 [23]	12	3 Level	Shared L3	192W
48-Core IA-32 [24]	48	2 Level	Shared L2	201W
Intel Haswell [22]	8	2 Level	Shared L2	140W

tasks run on a CMP, the striking differences among running tasks result in significant thermal gradient across the whole chip.

To illustrate this point, we performed thermal simulations on three different CMPs (refer to Table I) to capture the on-chip thermal distributions using Hotspot [22] (refer to Section V for the thermal simulation setup). The thermal map of 8-core Intel Haswell architecture [22] is shown in Fig. 4(a). As shown in the figure, LLC cache banks are located between two columns of cores. Each column has four cores. I/O interface and cache controller modules are located on the top and bottom side. The chip thermal design power is 140 W. When all eight cores are running at the peak power, we observe that core 3 and core 6 are working under the peak temperature which can approach 85.1 °C (assume the room temperature is 27 °C). Then, the on-chip thermal gradient can exceed 30.6 °C, which coincides with the measurement from [25]. Due to the lateral thermal propagation, the LLC region also suffers from severe thermal gradient.

For the thorough thermal evaluation, we also took IBM POWER8 [23] and the 48-core IA-32 processor [24] as another two examples. The thermal map of IBM POWE8R architecture is shown in Fig. 4(b). The POWER8 processor has 12 cores with 192 W peak power. The figure shows an example scenario that only central cores are running at the peak power while other cores are running with relatively lower power. We can observe that the thermal gradient can be as high as 30.3 °C as well. The thermal map of 48-core IA-32 architecture is shown in Fig. 4(c). The processor has 48-cores organized into 24 tiles with 201 W peak power. The figure illustrates the scenario that only couple of cores on the left side are running at the peak power. The thermal gradient is nearly 30 °C. Note that the real thermal distributions on-chip strongly depend characteristics of running applications. Nonetheless, it is expected that the thermal gradient will aggravate further due to the shrinking of technology node and the "dark silicon" problem [26]. In the following sections, we take the Intel Haswell architecture as an example to present our idea. Normally, the aggravating temperature and thermal gradient are undesirable for SRAM-based cache as they can increase the leakage power and threaten the chip reliability significantly. However, considering the thermal property of STT-MRAM write operation, we can take advantage of the thermal gradient on-chip to reduce the write latency and energy of LLC which will be detailed in Section IV.

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

A. Introduction to the Baseline Case

As shown in Fig. 5(a), the CMP considered in this paper, which is similar to the Intel Haswell architecture, has eight cores with private SRAM L1 cache and shared STT-MRAM-based L2 cache. The large LLC cache 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. Eight banks in a half-row constitute a bankset. For the S-NUCA architecture, the cache block is fixed in the bankset while the cache block can migrate within the bankset in the bankset with a specific data migration policy [21]. Fig. 5(b) illustrates a widely used D-NUCA data migration policy, called "gradual promotion" [21]. For example, if a data block in bank 8 is accessed, it will migrate toward 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 call this policy the gradual promotion policy.

Although S-NUCA does not incur data migration, which can avoid long write latency of STT-MRAM, D-NUCA can improve performance by migrating active data near to the core. To validate the performance benefit of the gradual promotion scheme, we compared performance of PARSEC applications using gradual promotion scheme with those using S-NUCA scheme as shown in Fig. 6 (refer to Section V for the detailed architecture simulation setup). It shows both quad-core and eight-core simulation results. It indicates that although gradual promotion incurs extra write latency, the access latency

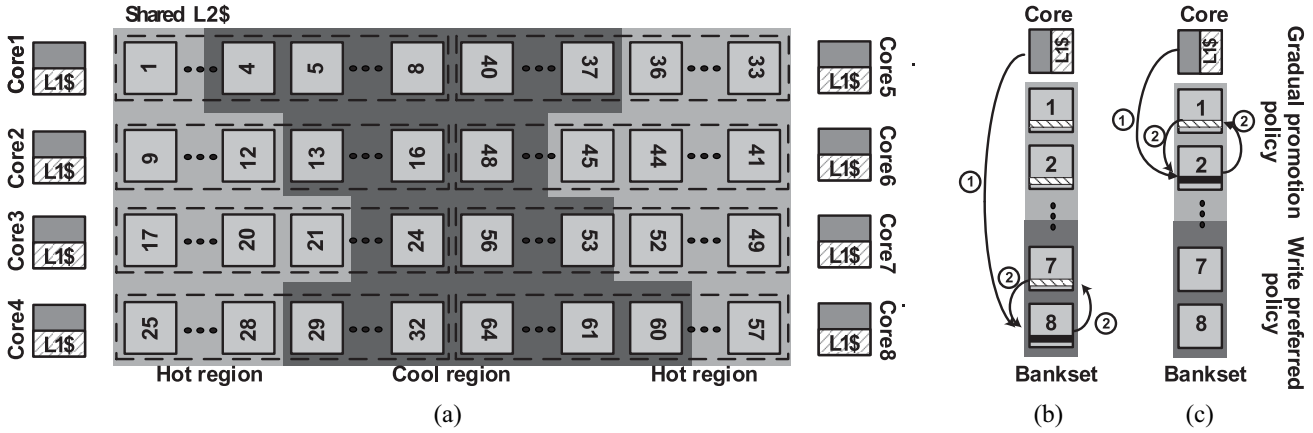


Fig. 5. (a) Example of the NUCA architecture to illustrate data migration policies. (b) Gradual promotion policy [21]. (c) Data migration policy of Thermosiphon design.

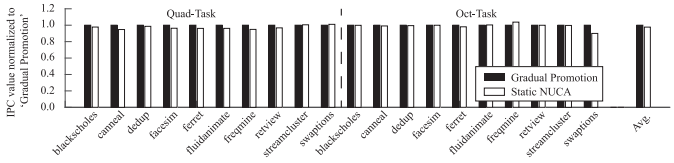


Fig. 6. IPC value comparisons of gradual promotion and S-NUCA for quad-core and eight-core cases when running PARSEC applications.

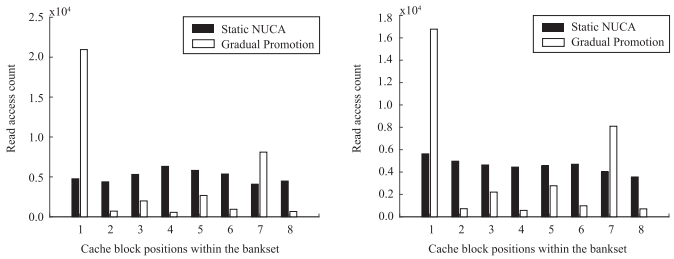


Fig. 7. Average read hit counts of different cache banks within a bankset for quad-core and eight-core cases when running PARSEC benchmarks.

of frequently accessed data can be reduced effectively by promoting them near to the core. The D-NUCA policy can improve IPC value by 2.38% compared to S-NUCA policy on average. Fig. 7 plots the average read hit counts of gradual promotion and S-NUCA for PARSEC benchmarks after three million instruction execution. It illustrate that gradual promotion policy can increase the read hit rate of the bank near to the requesting core dramatically, which contributes the IPC improvements of gradual promotion policy. The evaluation results for SPEC2000 and SPEC2006 benchmarks show the similar trends. Therefore, we set D-NUCA adopting the gradual promotion data migration policy as our baseline.

B. Overview of Our Proposed Thermal Aware NUCA Architecture

The working mechanism of the proposed NUCA design is illustrated in Fig. 8. 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) [27]. In the thermal evaluation module (TEM), there is a tuning table which stores the write pulse width and latency settings under different temperatures, which can be precharacterized by prototype measurements. During the application running, the thermal information is fed to TEM and the TEM will search the tuning table to find the appropriate write pulse configuration for the cache bank in LLC. The LLC temperature information and the corresponding write pulse settings 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 different thermal region types affects the complexity of TEM and the tuning table. Increasing the number of different thermal region types and the write pulse width tuning levels complicates the bank controller and the write driver design. In this paper, we only consider to divide each bankset into two thermal regions, i.e., the hot region and the cool region, and the tuning table contains 16 entries covering -20 °C to 130 °C to approach a reasonable tradeoff of hardware overhead and energy savings. In this paper, we adopted the write pulse tuning circuit proposed in [28]. As shown in Fig. 8, the write pulse tuning circuit can dynamically select an optimized pulse width (voltage) for an STT-RAM LLC bank according to the temperature. The write pulse can be generated by the simple delay circuit, which is integrated into the write driver circuitry.

C. On-Line Thermal Evaluation and Dynamic Thermal Region Division

Since our proposed Thermosiphon scheme is based on the thermal distribution on-chip, it is crucial to obtain the temperature of each cache bank efficiently. In this paper, we assume that all cache banks are classified as belonging to the hot region or the cool region. Banks in the hot region have temperature higher than the thermal threshold, so they have lower write energy and latency as well. Similarly, banks in the cool region suffer from larger write energy and latency.

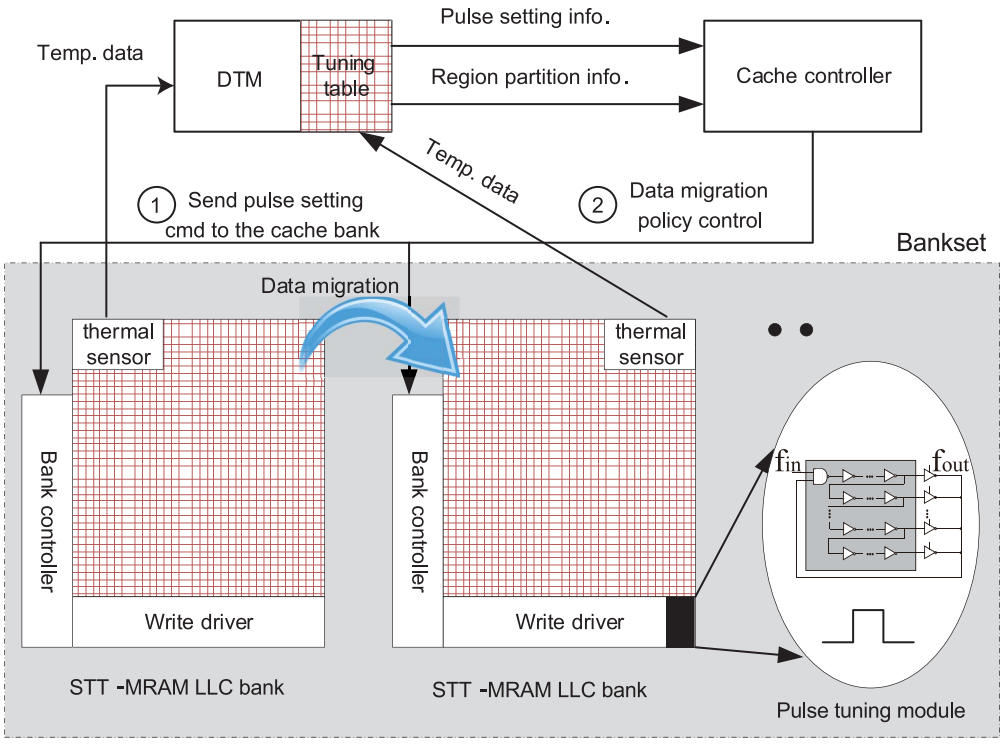


Fig. 8. Overview of our proposed NUCA design.

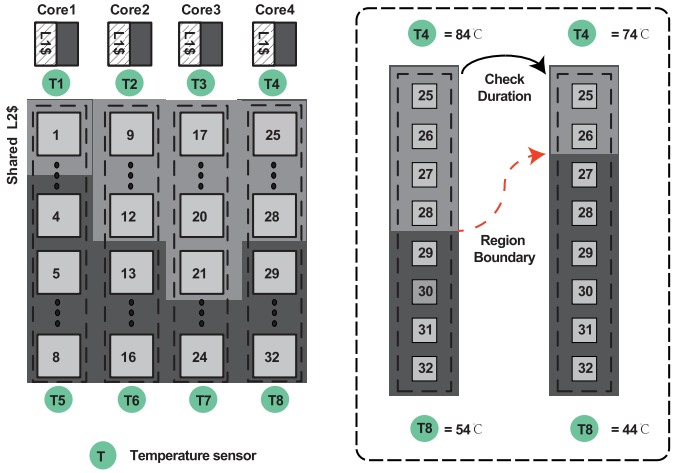


Fig. 9. Example of placements of thermal sensors.

Since thermal sensors are commonly available in the contemporary processors, we can utilize them to evaluate cache bank temperature. We assume thermal sensors are only available in a few locations as shown in Fig. 9. Then, the temperature of any bank n in a bankset (e.g., bank 0∼7) can be calculated as follows:

$$T_n = T_1 - (n-1)\frac{T_1 - T_5}{7} \tag{6}$$

where T_n is the temperature of bank n, T_1 and T_5 are the temperature of the top and the bottom thermal sensors in the bankset, respectively. The formula is based on the observation that cache bank temperature decreases with the distance away from the processor core as shown in Fig. 4.[2]

With the help of thermal sensors, we can dynamically divide LLC banks into different thermal regions. The temperature threshold for the thermal region division can be expressed as follows:

$$T_{\text{Threshold}} = \frac{\max\{T_1, T_2, T_3, T_4\} - \min\{T_5, T_6, T_7, T_8\}}{2}. \tag{7}$$

Based on the temperature, cache banks are classified as either hot banks or cool banks. For example, as shown in the right part of Fig. 9, the hot/cool region boundary of one bankset can be updated dynamically according to thermal sensing results.

D. 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, data read accesses are more frequent than write accesses. Moreover, since read performance largely determines cache access performance, it should be assigned with a higher migration priority. As a result, read intensive data may occupy banks in the hot region most of the time. Note that hot banks are usually located in the vicinity of active cores. Then, few write intensive data have the opportunity to migrate into the hot region. Moreover, a data migration

[2]In addition to the above thermal evaluation method, some other more accurate on-line thermal evaluation techniques such as [29] can be adopted as well. Note that our thermal-based data migration policy is orthogonal to the thermal evaluation method.

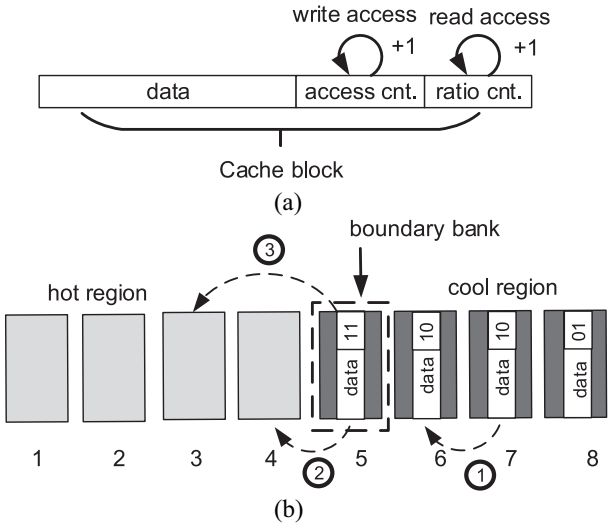


Fig. 10. (a) Access counter and the ratio counter in the cache block. (b) Illustration of data migration in the cool region.

occurs in each cache access because the newly accessed data migrate toward the requesting core by one step each time except that the cache block has already been located in the nearest bank to the core. The migration introduces extra write overheads which are more undesirable for write energy consuming STT-MRAM LLCs. 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 on-chip without degrading read performance significantly.

To explore the tradeoff involved in the thermal-aware NUCA design, we consider two extremes first. One extreme is always promoting the touched data block toward the requesting core. It results in read intensive data clustering in the neighborhood of the core. Thus, write operations 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 toward the requesting core with a higher priority. Therefore, we can reap more thermal benefits by clustering write-intensive data in the hot region. However, read performance may degrade dramatically in this case. Thus, there is a tradeoff 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. As shown in Fig. 10(b), block 1 may be occupied by read intensive data with a high possibility. Since write latency and energy are roughly the same in the same thermal region,[3] write intensive block can still reap the thermal benefit without residing in block 1 as long as it is in the hot region [i.e., the light gray area in Fig. 5(a)] as well. So 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 write intensive data toward the hot region with a higher possibility. Each cache block has two counters, i.e., the access counter and the read/write ratio counter as shown in Fig. 10(a). The former one indicates the access history of the cache block. Meanwhile, to elevate the possibility of write intensive data being migrated into the hot region, 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, write access to the block increases the access counter by 1 directly. Therefore, the net effect is to set the priority of data migration for write intensive data higher than that of read intensive data. A higher ratio means a write intensive block is assigned a higher probability to be migrated to the hot region. Another benefit of the proposed policy is that data migration may not always incur data swappings as long as the access counter does not change. Then, the write overhead induced by data migration can also be reduced significantly.

With the access counter and the ratio counter, banks in the cool region can be classified as the boundary bank and the nonboundary bank. The former one 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. 10(b). As for cache accesses in a nonboundary bank [case ① in Fig. 10(b)], cache controller reads the access counter in the cache block, and compare it with previous cache blocks.[4] The block is then migrated to the position next to the block, whose access counter value is larger than that of the block being migrated. For example, if an write 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, cache controller tries to find whether that block is just evicted from the hot-region or not. In the former case, cache controller makes the data migration similar to the nonboundary block case (case ② in Fig. 10). If the boundary block is just evicted out from the hot region, it will migrate by one additional step (case ③). The reason is that if it 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 brought into the hot region as well. By migrating toward the core one step further (i.e., migrate to bank 3), the undesirable ping-pong effect can be eliminated effectively.

The access counter and the ratio counter need to be reset in two cases. First, when one of the counters overflows, the access counter of each block is reset, which is similar to the "pseudo-LRU" cache replacement counter. This policy is to keep the leading-edge block in the previous position. Meanwhile, it gives opportunities to other blocks to be migrated into the hot region. Second, when a new block is loaded into the LLC, all the block counters are reset. Since the new block has the higher possibility to be visited again,

[3]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 V.

[4]The previous blocks are located in banks within the same bankset but nearer to the hot region.

TABLE II
CMP ARCHITECTURE CONFIGURATION

Processor	8-core @ 3.3 GHz, out-of-order, alpha
L1-Cache	I-cache 128KByte 8-way set associative
	D-cache 128KByte 8-way set associative
L2-Cache	16 MByte, 8-way set associative, shared
	MOESI cache coherence protocol
Main Memory	4 GByte DDR3 DRAM

TABLE III
THERMAL SIMULATION PARAMETERS

Bulk Si thickness	150μm
Cu metal layer thickness	0.42μm
Si thermal conductivity	$100.0W/(m-K)$
Heat sink thermal conductivity	$400.0W/(m-K)$
HotSpot grid resolution	64×64
Ambient temperature	27°C

the reset can make the new block be migrated to the boundary block directly if the block is touched again.

TABLE IV
BENCHMARK SUITES USED IN OUR ARCHITECTURAL SIMULATIONS

SPEC2000	Apsi, Art, Crafty, Mesa
	Swim, Twolf, Vortex, Vpr,
SPEC2006	Bzip2, CactusADM, DealII, Gromacs
	Hmmer, Libquantum, mcf, sjeng
PARSEC	blackscholes, canneal, dedup,facesim
	ferret, fluidanimate, freqmine
	retview, streamcluster,swaptions

TABLE V
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	$3.68 \times 10^3 T$
Oxide barrier thickness	0.85nm
Free layer thickness	33.55nm
Gyromagnetic ratio	$1.76 \times 10^7 rad/(s \cdot T)$

V. EXPERIMENTAL RESULTS

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

A. Experiment Setup

The CMP architecture used in our simulation consists of 8 Alpha 21264 cores as shown in Fig. 5(a). Each core has 128 KB private instruction and data cache, respectively. L2 cache is shared by all cores. L2 cache banks are interconnected with a mesh NoC. We assume one cycle/hop of the interconnect delay. The detailed architectural setup for simulations is tabulated in Table II. We extended the gem5 simulator [30] to model the proposed NUCA architecture. The gradual promotion was 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 widths in different thermal regions. These two schemes were used for comparison in our experiments.

In order to obtain thermal distribution of the CMP, we extracted the floorplan and power consumption information of Intel Haswell processor which is very similar to architecture used in our simulations. Hotspot [22] was used for the thermal simulations. The parameter settings for Hotspot simulations are shown in Table III. Since the thermal constant lies at hundreds of milliseconds [31], we set 100 ms as a reasonable thermal checking interval, which can guarantee that the temperature rise do not exceed 1 °C during the interval. At each thermal checking interval, the bankset was divided into two regions dynamically as mentioned above. Depending on the bank temperature, write pulse widths of the two regions can be determined by the pulse width tuning table as shown in Fig. 8. The write pulse width was set based on the lowest bank temperature in the specific thermal region.

The cell read/write energy and latency were obtained from the HSPICE simulations on the 40-nm perpendicular STT-MRAM technology using the model developed in [20]. The MTJ parameters used for the simulation are shown in Table V. Then, simulation results were fed into NVSim [32] to obtain the write energy and latency of L2 STT-MRAM cache. We assume that L2 cache is 16 MB, and the bank capacity is 256 KB.

The experiments were performed on the single-task/quad-task/eight-task scenario, respectively. The benchmark suites used in simulations include SPEC2000, SPEC2006 as multiprogram applications, and PARSEC as multithread applications as shown in Table IV. We constructed 16 single-task cases, four combinations of SPEC benchmarks for quad-task multiprogram simulations and four combinations of SPEC benchmarks for eight-task multiprogram simulations, which are tabulated in Table VI. Notice that, the single-task denotes the case that we only ran one benchmark on the eight-core processor to evaluate the effectiveness of our method on a single core. Quad-task represents the case that we assign four benchmarks (they are

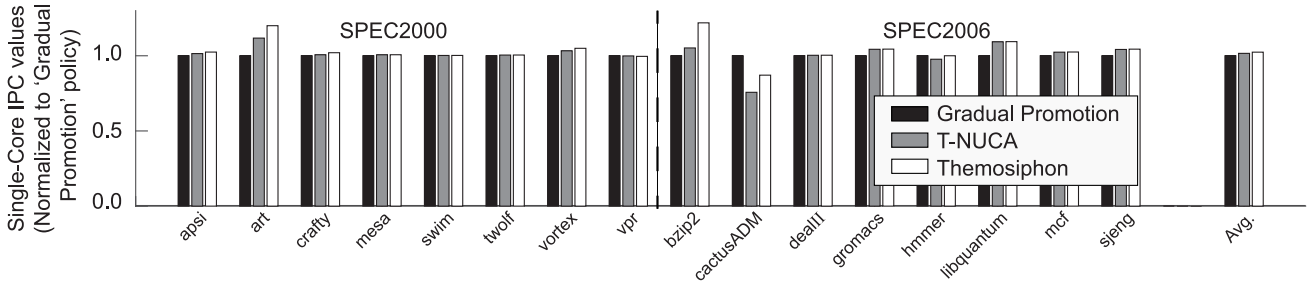


Fig. 11. Single-task IPC comparisons of three different NUCA designs: the gradual promotion NUCA, T-NUCA, and Thermosiphon.

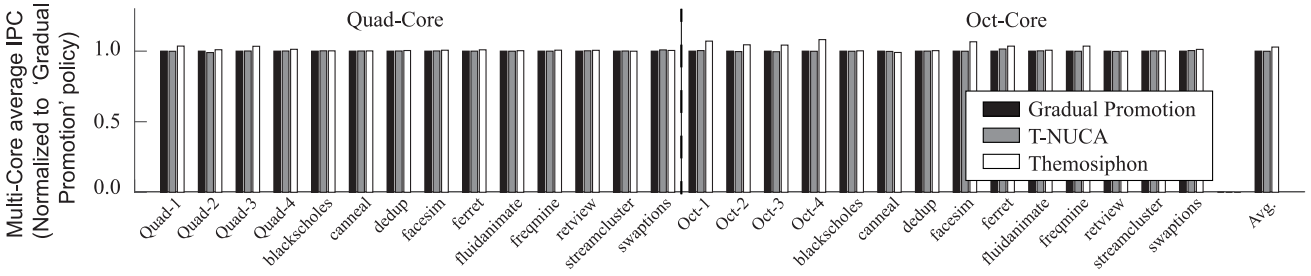


Fig. 12. Multitask IPC comparisons of three different NUCA designs: the gradual promotion NUCA, T-NUCA, and Thermosiphon.

TABLE VI
QUAD-TASK AND PCT-TASK BENCHAMRK GROUPS CONSTRUCTED FROM THE SPEC2000 & SPEC2006 BENCHMARK SUITES

Group	Benchmark
Quad-1	DealII, CactusADM, Sjeng, Libquantum
Quad-2	DealII, Hmmer, Mcf, Gromacs
Quad-3	Libquantum, Sjeng, CactusADM, Mcf
Quad-4	Sjeng, Gromacs, DealII, Mcf
Oct-1	Swim, Crafty, Gap, Mesa, Art, Bzip2, Applu, Mgrid
Oct-2	Vortex, Gzip, Apsi, Art, Mgrid, Vpr, Swim, Gcc
Oct-3	Vpr, Gap, Applu, Gzip, Mesa, Vortex, Bzip2, Apsi
Oct-4	Hmmer, Libquantum, CactusADM, Gromacs DealII, Sjeng, Mcf, Bzip2

shown in Table VI correspond to the multiprogram scenario) or four threads (multithread scenario) on the eight-core processor. Eight-task is the case that we assign a benchmark to every core (the benchmark groups are shown in Table VI and correspond to the multiprogram scenario) or one thread to every core (multithread scenario). In our simulations, 200 million instructions were fast-forwarded to warm up the cache, and then 30 million instructions were executed to generate the simulation statistics. L2-cache access statistics obtained from gem5 were used to estimate the overall write energy consumption induced by both normal write accesses and data migrations.

B. Performance Analysis

IPC performance comparisons for the single-task case are plotted in Fig. 11, and those for quad-task and eight-task cases are plotted in Fig. 12. The results are normalized to the baseline, i.e., gradual promotion policy. As shown in Fig. 11(a), left bars denote IPC values of single-task simulation results, and the rightmost bar represents the geometric mean. Among the three NUCA designs, Thermosiphon performs the best, and can improve performance by 2.65% on average. For "bzip2," T-NUCA scheme can improve 4.95% performance compared to the baseline while Thermosiphon can obtain 22.7% improvement compared to the baseline. Considering the multitask simulations, IPC comparisons when running multiprogram and multithread applications are shown in Fig. 12, the IPC improvement of T-NUCA sharply drops to 0.5% at most (Oct-1), and Thermosiphon can still improve performance by 2.89% on average compared to the baseline. The reason is that Thermosiphon can provide more opportunities for write intensive data blocks in the cool region to be migrated in the hot region. Therefore, the write latency can be reduced accordingly. Meanwhile, read performance does not degrade significantly since the gradual promotion policy is adopted in the hot region. Moreover, the adoption of ratio counter also reduces data swappings as mentioned before, which also contributes to the write performance improvement.

C. Write Energy Analysis

Write energy comparisons of three NUCA designs are presented in Fig. 13. Fig. 13(a) illustrates write energy comparisons, including single-task, quad-task, and eight-task cases. The total energy comparisons, considering both read and write energy, are shown in Fig. 13(b). All results are normalized to the baseline. As shown in the figures, Thermosiphon performs the best in terms of write energy and overall energy consumptions. As for write energy, Thermosiphon can save 13.01% on average, 41.2% at most ("swaptions") compared with the baseline. Meanwhile, T-NUCA can only achieves about 2.79%

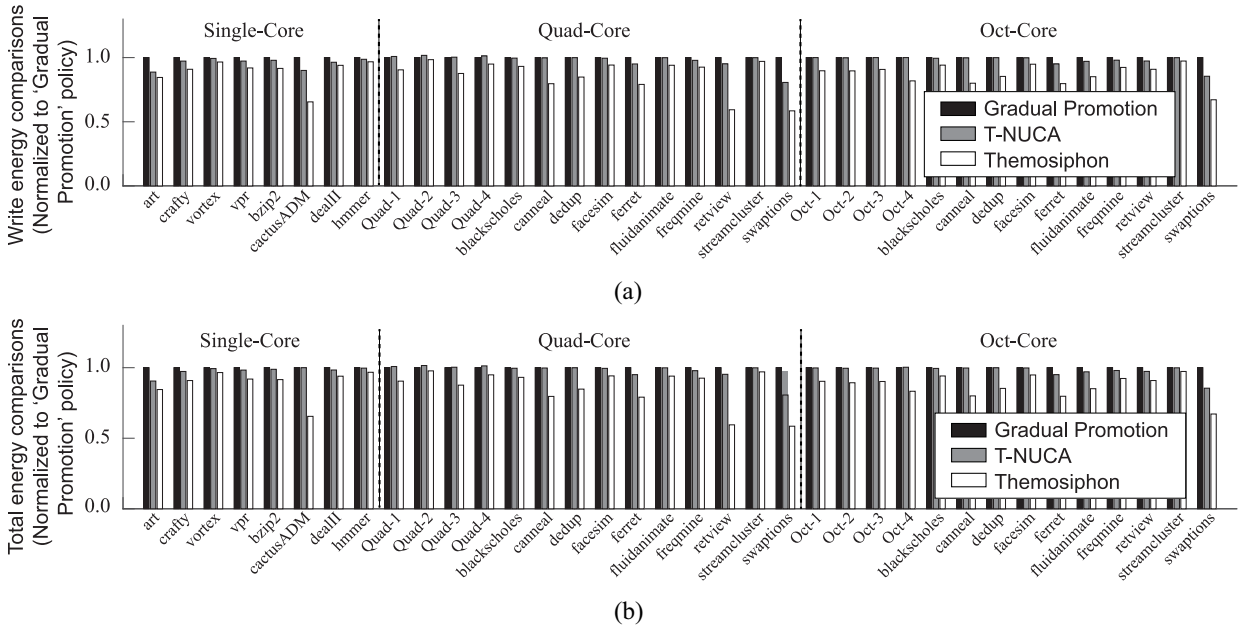


Fig. 13. Energy comparisons of three NUCA designs. (a) Write energy comparisons. (b) Over all energy comparisons.

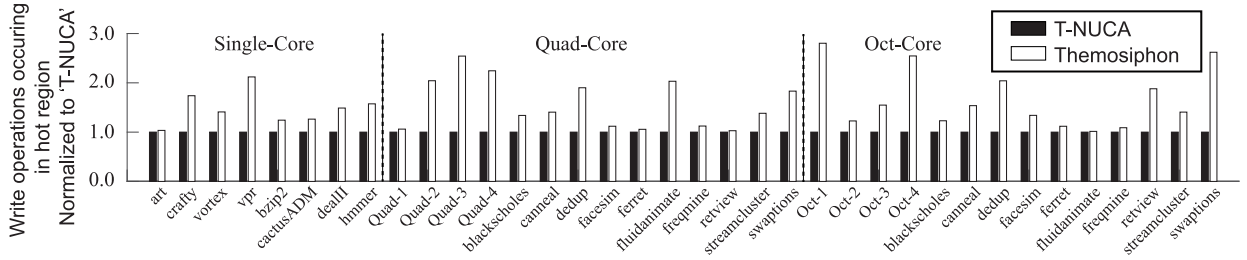


Fig. 14. Comparisons of the write activities in the two thermal regions. It implies that more write operations occur in the hot region for Thermosiphon.

write energy reduction on average. Considering the overall energy consumption as shown in Fig. 13(b), Thermosiphon also performs the best. The reason is that Thermosiphon can migrate a write intensive block into the hot region with a 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 occupy the hot region most of the time. For those write infrequent or read intensive applications, without the help of ratio counter, most of write operations still remain in the cool region, which prevents us from obtaining the benefit brought by the hot region.

D. Write Activities in Different Thermal Regions

To validate the effectiveness of our proposed design further, we analyze write operations occurring in different thermal regions. Write operations can be divided into two categories. The first one is normal write accesses (including instruction/data loading operations). The other one is data swapping induced write operations. Considering data migration, the two blocks involved may be located in different thermal regions. By comparing percentages of write operations occurring in different thermal regions, we can verify whether more write operations are performed in the hot region. As shown in Fig. 14(a), 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 Thermosiphon design.

E. 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 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 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 access characteristic of the cache block can not be captured accurately. To explore the design space of the counter design, we obtained IPC values and write energy with different counter configurations, and plotted the average value over all benchmarks in Fig. 15. As shown in Fig. 15(b), we observe 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,

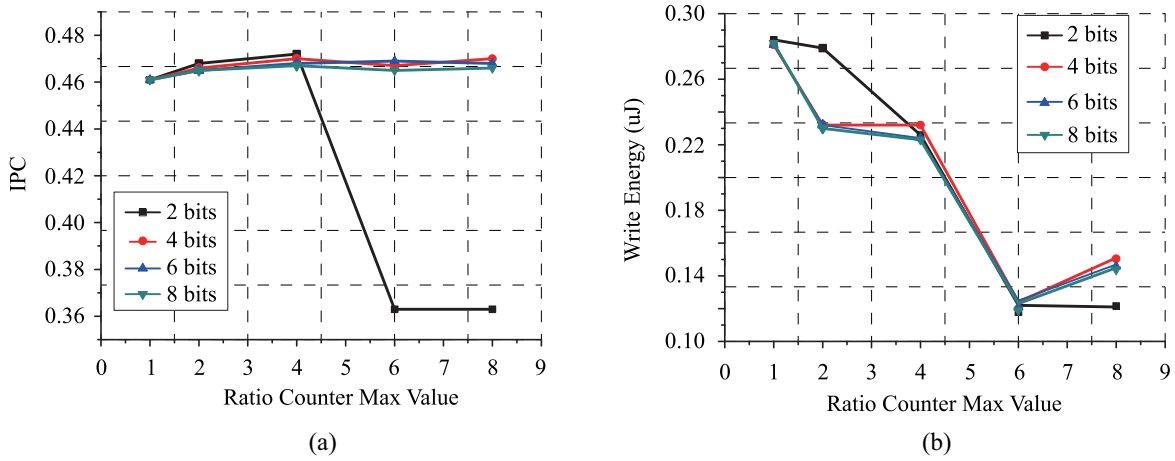


Fig. 15. Design space exploration of the access/ratio counter (the legend represents the access counter bit): (a) from the IPC perspective and (b) from the energy consumption perspective.

we derived that the ratio 6 is optimal as well, which is shown in Fig. 15(a). So the ratio counter has 3 bits. Similarly, the optimal access counter bits should be 4.

F. Hardware Overhead Analysis of the Proposed Thermosiphon

As for the temperature sensors and DTM module, they commonly exist in modern microprocessors [33]. These facilities can be used to measure on-chip temperature and assist the thermal management. In this paper, we just took advantage of these existing facilities to obtain the bank temperature, and partitioned LLC bankset into different thermal regions. Additionally, we adopted the write pulse width adjustment circuit proposed in [34]. It only incurs <0.005% area overhead. Considering the access and ratio counters, since the access counter is commonly accompany with the cache block in CMPs to record the access history for the cache block replacement [30], our proposed scheme can use it directly. The extra hardware overhead is the addition of a ratio counter to each block. As the maximum count value of the ratio counter is 6 as mentioned above, which is equivalent to 3 bits hardware implementation. The hardware overhead is 3 bits/64 bytes = 1.3%. Compared to the write energy and performance improvements, the hardware overhead is acceptable.

VI. Related Work

The prevalence of multicore processor demands a steady growth of on-chip cache capacity to bridge the gap between processor throughput and the off-chip memory bandwidth. Kim *et al.* [5] proposed NUCA architecture to address the performance bottleneck of conventional UCA cache. After that, a lot of researchers studied the data mapping, insertion and interbank data migration strategies to fully exploit the potential of NUCA. For example, Beckmann and Wood considered the cache block migration and replacement policy in D-NUCA of multicore systems [35]. In addition, Qiu *et al.* [36] studied the data migration policy in a hybrid STT-MRAM cache. Beside migration strategy, there are also some literatures on the optimal placement of cache blocks in order to improve data locality without introducing data migration overhead [37].

Although STT-MRAM has many advantages working as an emerging memory technology, its write operation is energy and time consuming thus attracting a lot of research efforts to solve this issue. Zhou *et al.* [11] proposed a write early termination technique to reduce write energy of STT-MRAM. The experimental results indicated that the total write energy can be reduced by more than 33%. Similarly, Park *et al.* [38] took advantage of the stochastic long-tail nature of STT-MRAM write operation, and shut down the bitline as soon as the cell was detected to have the desirable switching. Sun *et al.* [39] proposed a multiretention level STT-MRAM cache design and associated adaptive data refresh policy to reduce the energy overhead and improve cache access performance. Chi *et al.* [40] explored the data encoding scheme for STT-MRAM. By mapping frequent occurring data patterns to the energy efficient resistive states, the write energy of STT-MRAM can be reduced dramatically. Bishnoi *et al.* [41] observed the asymmetric write of STT-MRAM, i.e., writing 1 is more energy-consuming than writing 0, and provided different write timings for the two writing scenarios for the write energy reduction. Zeng and Peir [42] proposed to select a replacement block near the LRU position, which has the most similar content to the missed block to reduce the total switch bits. The proposed method achieves 20% switching bits reduction. Wu *et al.* [43] exploited the data redundancy within the write back data from upper level cache to STT-MRAM LLCs and proposed a compare-and-write technique to eliminate the redundant write back data for write energy optimization.

Different from the above related work, this paper exploits the thermal gradient on-chip, and seeks to increase energy efficiency and performance of STT-MRAM LLCs taking advantage of STT-MRAM's thermal property.

VII. Conclusion

As the modern processor enters into the multicore 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 this paper, we take advantage of the thermal property of STT-MRAM write operation to reduce the energy consumption and improve cache access performance in LLCs. With the thermal consideration, we propose a thermal aware NUCA design–Thermosiphon, which adopts different data migration policies in different thermal regions. The experimental results show that compared with the baseline, our proposed NUCA design can improve the performance by 2.65% on average for the single-task case, 2.89% for the multitask case on average. Meanwhile, Thermosiphon can reduce the write energy by 13.01% on average and 41.2% at most with negligible hardware overhead.

ACKNOWLEDGMENT

The authors would like to thank Prof. G. Sun and Dr. C. Zhang for helpful discussions and suggestions.

REFERENCES

[1] M. P. Jagtap, "Era of multi-core processors," *Power*, vol. 2, pp. 87–94, Mar. 2009.

[2] A. Sodani *et al.*, "Knights landing: Second-generation Intel Xeon Phi product," *IEEE Micro*, vol. 36, no. 2, pp. 34–46, Mar./Apr. 2016.

[3] N. Muralimanohar, R. Balasubramonian, and N. Jouppi, "Optimizing NUCA organizations and wiring alternatives for large caches with CACTI 6.0," in *Proc. IEEE/ACM Int. Symp. Microarchitect. (MICRO)*, Chicago, IL, USA, Dec. 2007, pp. 3–14.

[4] Y. Wang, H. Li, Y. Han, and X. Li, "A low overhead in-network data compressor for the memory hierarchy of chip multiprocessors," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 6, pp. 1265–1277, Jun. 2018.

[5] C. Kim, D. Burger, and S. W. Keckler, "An adaptive, non-uniform cache structure for wire-delay dominated on-chip caches," in *Proc. ACM Archit. Support Program. Lang. Oper. Syst. (ASPLOS)*, San Jose, CA, USA, Oct. 2002, pp. 211–222.

[6] Y. Wang, Y. Han, H. Li, and X. Li, "VANUCA: Enabling near-threshold voltage operation in large-capacity cache," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 3, pp. 858–870, Mar. 2016.

[7] S. Raoux *et al.*, "Phase-change random access memory: A scalable technology," *IBM J. Res. Develop.*, vol. 52, nos. 4–5, pp. 465–479, Jul. 2008.

[8] S. Gaba, P. Knag, Z. Y. Zhang, and W. Lu, "Memristive devices for stochastic computing," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Melbourne, VIC, Australia, Jul. 2014, pp. 2592–2595.

[9] M. Wang *et al.*, "Field-free switching of a perpendicular magnetic tunnel junction through the interplay of spin-orbit and spin-transfer torques," *Nat. Electron.*, vol. 1, no. 2018, pp. 582–588, 2018.

[10] 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 *Proc. ACM Int. Conf. Great Lakes Symp. VLSI (GLSVLSI)*, New York, NY, USA, May 2013, pp. 55–60.

[11] P. Zhou, B. Zhao, J. Yang, and Y. Zhang, "Energy reduction for STT-RAM using early write termination," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, San Jose, CA, USA, Dec. 2009, pp. 264–268.

[12] W. Wen, Y. Zhang, Y. R. Chen, Y. Wang, and Y. Xie, "PS3-RAM: A fast portable and scalable statistical STT-RAM reliability/energy analysis method," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 11, pp. 1644–1656, Nov. 2014.

[13] X. Jia *et al.*, "Understanding how non-uniform distribution of memory accesses on cache sets affects the system performance of chip multiprocessors," in *Proc. IEEE Int. Symp. Parallel Distrib. Process. Appl. Workshops*, Busan, South Korea, Jul. 2011, pp. 266–272.

[14] L. Song *et al.*, "STT-RAM buffer design for precision-tunable general-purpose neural network accelerator," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 4, pp. 1285–1296, Apr. 2017.

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

[16] M. X. Wang *et al.*, "Current-induced magnetization switching in atom-thick tungsten engineered perpendicular magnetic tunnel junctions with large tunnel magnetoresistance," *Nat. Commun.*, vol. 9, no. 1, pp. 671–678, 2018.

[17] C.-P. Wan and B. J. Sheu, "Temperature dependence modeling for MOS VLSI circuit simulation," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 8, no. 10, pp. 1065–1073, Oct. 1989.

[18] 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," *J. Appl. Phys.*, vol. 114, no. 11, 2013, Art. no. 114506.

[19] 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 *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Baltimore, MD, USA, Mar. 2009, pp. 1–4.

[20] B. Wu, Y. Cheng, J. Yang, A. Todri-Sanial, and W. Zhao, "Temperature impact analysis and access reliability enhancement for 1T1MTJ STT-RAM," *IEEE Trans. Rel.*, vol. 65, no. 4, pp. 1755–1768, Dec. 2016.

[21] J. Lira, C. Molina, R. N. Rakvic, and A. González, "Replacement techniques for dynamic NUCA cache designs on CMPs," *J. Supercomput.*, vol. 64, no. 2, pp. 548–579, 2013.

[22] W. Huang *et al.*, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 14, no. 5, pp. 501–513, May 2006.

[23] B. Sinharoy *et al.*, "IBM POWER8 processor core microarchitecture," *IBM J. Res. Develop.*, vol. 59, no. 1, pp. 380–393, 2015.

[24] J. Howard *et al.*, "A 48-core IA-32 message-passing processor with DVFS in 45nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, Feb. 2010, pp. 108–109.

[25] F. J. Mesa-Martinez, E. K. Ardestani, and J. Renau, "Characterizing processor thermal behavior," in *Proc. ACM Archit. Support Program. Lang. Oper. Syst. (ASPLOS)*, Santa Cruz, CA, USA, Mar. 2010, pp. 193–204.

[26] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Proc. IEEE Annu. Int. Symp. Comput. Archit. (ISCA)*, San Jose, CA, USA, Jun. 2011, pp. 365–376.

[27] S. Sharifi, D. Krishnaswamy, and T. S. Rosing, "PROMETHEUS: A proactive method for thermal management of heterogeneous MPSoCs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 7, pp. 1110–1123, Jul. 2013.

[28] 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 *Proc. ACM/IEEE Int. Symp. Low Power Electron. Design (ISLPED)*, New York, NY, USA, Jul. 2012, pp. 197–202.

[29] N. Allec, Z. Hassan, L. Shang, R. P. Dick, and R. Yang, "ThermalScope: Multi-scale thermal analysis for nanometer-scale integrated circuits," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, San Jose, CA, USA, Nov. 2008, pp. 603–610.

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

[31] J. Choi *et al.*, "Thermal-aware task scheduling at the system software level," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Portland, OR, USA, Aug. 2007, pp. 213–218.

[32] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 7, pp. 994–1007, Jul. 2012.

[33] C. M. Jha, Ed., *Thermal Sensors (Principles and Applications for Semiconductor Industries)*. New York, NY, USA: Springer, 2015.

[34] S. Wang *et al.*, "MTJ variation monitor-assisted adaptive MRAM write," in *Proc. IEEE Design Autom. Conf. (DAC)*, Austin, TX, USA, Aug. 2016, pp. 1–6.

[35] B. C. Beckmann and D. A. Wood, "Managing wire delay in large chip-multiprocessor caches," in *Proc. IEEE/ACM Int. Symp. Microarchit. (MICRO)*, Washington, DC, USA, Dec. 2004, pp. 319–330.

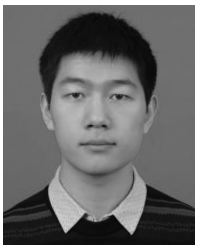
[36] K. Qiu, M. Zhao, Q. Li, C. Fu, and C. J. Xue, "Migration-aware loop retiming for STT-RAM-based hybrid cache in embedded systems," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 3, pp. 329–342, Mar. 2014.

[37] N. Hardavellas, M. Ferdman, B. Falsafi, and A. Ailamaki, "Reactive NUCA: Near-optimal block placement and replication in distributed caches," *ACM SIGARCH Comput. Archit. News*, vol. 37, no. 3, pp. 184–195, 2009.
[38] J. Park, T. Zheng, M. Erez, and M. Orshansky, "Variation-tolerant write completion circuit for variable-energy write STT-RAM architecture," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 4, pp. 1351–1360, Apr. 2016.
[39] Z. Sun *et al.*, "Multi retention level STT-RAM cache designs with a dynamic refresh scheme," in *Proc. IEEE/ACM Int. Symp. Microarchit. (MICRO)*, Porto Alegre, Brazil, Dec. 2017, pp. 329–338.
[40] P. Chi, C. Xu, X. Zhu, and Y. Xie, "Building energy-efficient multi-level cell STT-MRAM based cache through dynamic data-resistance encoding," in *Proc. IEEE Int. Symp. Qual. Electron. Design*, Santa Clara, CA, USA, Mar. 2014, pp. 639–644.
[41] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. Tahoori, "Asynchronous asymmetrical write termination (AAWT) for a low power STT-MRAM," in *Proc. Conf. Design Autom. Test Europe (DATE)*, Dresden, Germany, Mar. 2014, pp. 1–6.
[42] Q. Zeng and J.-K. Peir, "Content-aware non-volatile cache replacement," in *Proc. Parallel Distrib. Process. Symp.*, Orlando, FL, USA, Jun. 2017, pp. 92–101.
[43] B. Wu *et al.*, "Write energy optimization for STT-MRAM cache with data pattern characterization," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, Hong Kong, Jul. 2018, pp. 1–6.

Bi Wu (S'15) received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, and the M.S. degree from Beihang University, Beijing, China, where he is currently pursuing the Ph.D. degree in electrical engineering.

His current research interests include circuit level and architecture level design and optimization of STT-MRAM, SOT-MRAM, and the corresponding reliability analysis and improvement.

Mr. Wu was a recipient of the China National scholarship for doctoral students, which is awarded by the Ministry of Education of China, in 2017.



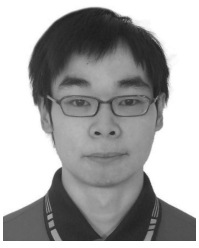
Pengcheng Dai received the B.S. degree from Beihang University, Beijing, China, where he is currently pursuing the M.S. degree in electrical engineering.

His current research interests include usage of MRAM in computer architecture and architecture of embedded devices.

Yuanqing Cheng (S'11–M'13) received the Ph.D. degree from the Key Laboratory of Computer Systems and Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

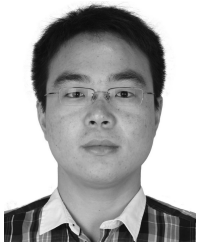
After spending one year of post-doctoral study with LIRMM, CNRS, Paris, France, he joined Beihang University, Beijing, as an Assistant Professor. His current research interests include reliable design for 3-D integrated circuits, as well as computing architecture design for emerging technologies, including spintronic and carbon nanotube devices.

Dr. Cheng is an ACM Member.



Ying Wang (M'14) received the B.S. and M.S. degrees in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2007 and 2009, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2014.

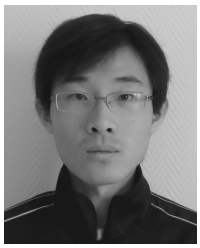
He is currently an Assistant Professor with ICT, CAS. His current research interests include computer architecture and very large scale integration design, specifically memory systems, on-chip interconnects, resilient and energy-efficient architecture, and machine learning accelerators.



Jianlei Yang (S'12–M'16) received the B.S. degree in microelectronics from Xidian University, Xi'an, China, in 2009 and the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2014.

He joined Beihang University, Beijing, in 2016, where he is currently an Associate Professor with the School of Computer Science and Engineering. From 2014 to 2016, he was a Post-Doctoral Researcher with the Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA. From 2013 to 2014, he was a Research Intern with Intel Labs China, Beijing, and Intel Corporation, Santa Clara, CA, USA. His current research interests include spintronics and neuromorphic computing systems.

Dr. Yang was a recipient of the First Place prize from the TAU Power Grid Simulation Contest in 2011, the Second Place prize from the TAU Power Grid Transient Simulation Contest in 2012, the IEEE ICCD Best Paper Award in 2013, the IEEE ICESS Best Paper Award in 2017, and the ACM GLSVLSI Best Paper Nomination in 2015.



Zhaohao Wang (S'12–M'16) received the B.S. degree in microelectronics from Tianjin University, Tianjin, China, in 2009, the M.S. degree in microelectronics from Beihang University, Beijing, China, in 2012, and the Ph.D. degree in physics from the University of Paris-Saclay, Paris, France, in 2015.

His current research interests include the modeling of nonvolatile nano-devices and design of new nonvolatile memories and logic circuits.

Dijun Liu received the B.S. and M.S. degrees in electrical engineering from the China University of Petroleum (Huadong), Dongying, China, in 1993 and 1998, respectively, and the Ph.D. degree in geophysics from the China University of Petroleum (Beijing), Beijing, China, in 2001.

He is currently a Chief Scientist of the China Academy of Telecommunications Technology, Beijing, and a Professor with the School of Electronic and Information Engineering, Beihang University, Beijing. His current research interests include high performance SoC design and software define radio communication.

Dr. Liu is the Director of the China Institute of Communications and the Chairman of China Communications Integrated Circuit Committee.

Weisheng Zhao (M'06–SM'14–F'19) received the Ph.D. degree in physics from the University of Paris-Sud, Paris, France, in 2007.

From 2009 to 2014, he was a Tenured Research Scientist with CNRS, Paris. Since 2014, he has been a Youth 1000 Plan Distinguished Professor with Beihang University, Beijing, China. He is also the principal inventor of four international patents. He has authored or co-authored over 150 scientific papers (e.g., *Nature Electronics*, *Nature Communications*, and IEEE TRANSACTIONS). His current research interests include the hybrid integration of nano-devices with CMOS circuit and new nonvolatile memory (40-nm technology node and below) like MRAM circuit and architecture design.

Dr. Zhao is an Associate Editor of the IEEE TRANSACTIONS ON NANOTECHNOLOGY and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I: REGULAR PAPERS.