# TESLA: Using Microfluidics to Thermally Stabilize 3D Stacked STT-RAM Caches

Majed Valad Beigi and Gokhan Memik
Department of EECS, Northwestern University, Evanston, IL
majed.beigi@northwestern.edu and memik@eecs.northwestern.edu

*Abstract*— **In this work, we develop a 3D architecture that utilizes STT-RAM for the last level cache (LLC). 3D integration enables large LLCs to be stacked onto a die. However, 3D architectures suffer from higher operating temperatures due to increased power densities. The elevated temperatures can adversely impact the STT-RAM performance and reliability. The objective of this paper is to address the limits of integrating STT-RAM in 3D chip stacks from a thermal perspective and propose a novel stacking structure that minimizes heat-induced problems. Specifically, we analyze the system-level impact of increased temperatures and propose a novel technique to dynamically adjust the flow rate of the liquid interlayer cooling at run time to reduce the STT-RAM temperature and alleviate temperature-induced problems that cause the performance degradation and prevent overcooling the STT-RAM die and minimize the pump energy consumption. Evaluation results reveal that our approach achieves up to 19.1% performance improvement and 14.6% power reduction over an architecture that does not include an insulating layer.**

***Keywords: 3D stacking; Non-Volatile Memories; Temperature; Cooling Systems***

## I. INTRODUCTION

3D integration [5, 15] of large last level caches (LLCs) stacked onto a multicore die has become an attractive alternative for overcoming memory system problems such as large off-chip access latencies, lack of scalability, and high static energy consumption. Emerging non-volatile memory (NVM) technologies [5, 11, 19, 27] have been proposed to alleviate the bottlenecks of existing technologies due to their higher density, better scalability, and lower leakage power. Among the proposed NVM technologies, STT-RAM is an attractive alternative for on-chip SRAM due to its higher density and, lower leakage power compared to the traditional SRAM-based cache architectures and lower access delays compared to other alternative technologies [5, 26]. Also, it is compatible to be integrated with 3D stacking using conventional CMOS logic [16]. Even though the 3D-STT-RAM based cache architectures have many advantages, they suffer from two main challenges. The first one is the high STT-RAM write latency/energy overhead [5, 26]. Hybrid caches [5, 10, 26], which include both STT-RAM and SRAM, have been proposed as a potential solution to alleviate the impact of high write latency/energy of STT-RAM structures. The second challenge is caused by die stacking: die stacking typically increases the power density and causes elevated operating temperatures [13, 29]. Moreover, the TSVs and thermal vias play a crucial role in providing pathways for thermal and electrical communications between different layers [12]. Elevated temperatures affect both electrical and mechanical characteristics of STT-RAM and can cause performance degradation [5, 7, 13]. For example, an increase in the temperature in STT-RAM is reported to cause up to 25% reduction in the sense margins during a read operation and degrades the write speed by up to 20% [7]. Figure 1 shows the latency of cache hit, cache miss, and cache write of a 96 MB STT-RAM in 32nm technology for different temperatures. The results are obtained using DESTINY tool [18]. We normalize the results to the latencies at 40$^{o}$C. As illustrated in the figure, the cache hit, cache miss, and write latencies increase by 18.2%, 9.5%, and 8.3%, respectively, for the 96 MB STT-RAM with 24-way associativity when the component reaches 110$^{o}$C. These results show that temperature can have a significant impact on performance. Although the performance and power consumption in 2D and 3D stacked hybrid caches have been evaluated in the above mentioned works, none of the previous studies consider the impact of temperature on the STT-RAM performance in the system.

The solution for the temperature-related problems we propose is rather simple: thermally decouple the 3D-stacked logic die from the STT-RAM LLC die by introducing a cooling layer between them to maintain higher thermal stability and easier thermal decoupling. To address the challenges of heat removal in 3D-stacked logic, innovative cooling solutions such as air, porous silicon, and liquid cooling have been proposed [12, 21, 29]. Among those techniques, interlayer liquid cooling is an attractive solution due to the higher heat removal capability. Previous works have also shown that the flow rate of the cooler can be tailored dynamically [9, 12, 21]. This helps to minimize the vertical thermal gradient across the stacks when power dissipation varies (e.g., from the processor die to the STT-RAM die). Meanwhile, adjusting the flow rate to the needed value results in reduced pumping power and increased microchannel lifetime. First, pumping power increases exponentially with the increase of flow rate [12, 28]. Second, the microchannel is relatively fragile since the length of a channel wall is about 50 µm [21]. Using higher flow rate requires a higher pressure drop between the inlet and outlet [28]. Hence, this may damage microchannels and shorten their lifetime.

In this paper, we propose TESLA, a STT-RAM based 3D cache design that encases the STT-RAM die in a thermal cooling interlayer to reduce and stabilize its temperature, while minimizing the spatial and temporal thermal coupling between logic and STT-RAM LLC components. In this work, we first try to stack various interlayer cooling methods such as liquid cooling with uniform flow rate and porous silicon interlayer in order to explore their impact on thermally decoupling the processor and STT-RAM layers. We also note that, it is not energy-efficient to use a uniform coolant flow rate based on the worst-case conditions, since it would cause a waste in the pump power consumption. Similarly, using higher flow rate based on the worst-case condition may damage microchannels and shorten their lifetime. Hence, we propose a novel technique to dynamically adjust the flow rate of the liquid interlayer cooling at run time. Our goal is to prevent overcooling the STT-RAM die. We then evaluate the performance impact of thermal decoupling on the STT-RAM LLC die running a range of scientific workloads, under realistic physical constraints.

The rest of this paper is organized as follows. Next section discusses the TESLA. Section III describes our experimental methodology. Section IV presents the evaluation results and finally, a summary is provided in Section V.

## II. TESLA

TESLA alleviates the heating of the LLC layer by thermally decoupling the 3D-stacked processor die from the LLC die. This is achieved by placing a liquid cooling layer between them (Figure 2). However, using a cooling interlayer with a uniform coolant flow rate based on the maximum temperature (i.e., worst-case condition) is not energy-efficient. Hence, TESLA aims to predict the necessary flow rate of interlayer cooling at run time to minimize the pump energy consumption and reduce the temperature efficiently. Figure 3 shows the TESLA overview. The design of our predictor is based on collecting information of cache banks to predict their future temperature in order to assign an appropriate flow rate for the cooling layer. TESLA consists of a monitoring

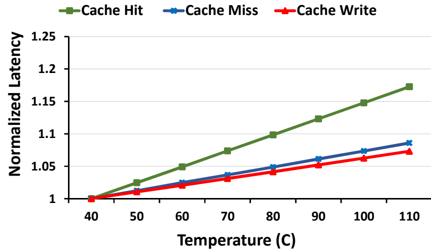


Fig. 1. Normalized cache hit, miss, and write latencies of STT-RAM for various temperatures.

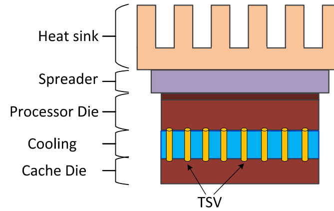


Fig. 2. TESLA 3D structure.

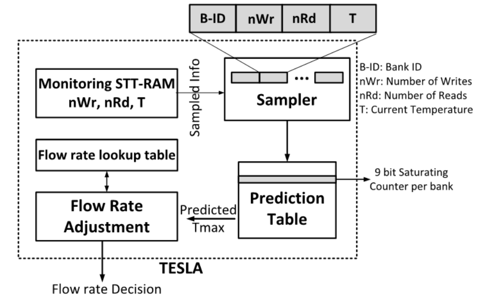


Fig. 3. TESLA overview.

component, a sampler, a prediction table, and a flow rate adjustment controller. As shown in Figure 3, TESLA first monitors the temperature and the number of read/write accesses to each bank to provide the ability to adapt the flow rate controller and cache access decisions. TESLA gathers the temperature of banks by using a thermal sensor in each bank. It also uses an $n$ bit saturating counter for collecting the number of read and write accesses per bank. In order to obtain the appropriate number for $n$, we explore the IPS (i.e., instruction per second) of the executed applications. Our evaluations revel that the best number for $n$ is 32 based on the rate we are going to collect our samples. The total storage of counters for 16 banks is 128B.

In addition, we determine the sampling rate based on a typical impeller pump [1] used in TESLA. This pump usually takes around 250-300ms to complete the transition to a new flow rate. The time delay in adjusting the flow rate may result in over-/under-cooling until the flow rate transition is complete [9]. Hence, for optimal operation, we need to predict the temperature into the near future and adjust the flow rate control on time to meet the heat removal requirement. In our experiments, we use a sampling rate of 100ms, and predict 500ms into the future. Our results show that collecting a single activity level for the whole interval results in low accuracy. Hence, within each sampling interval, we collect the number of reads, writes, and current temperature $m$ times. Our evaluations show that if $m$ is 128, the accuracy of the prediction is higher than 80%. Increasing the value of $m$ to above 128 does not significantly improve the accuracy of the prediction. Hence, we select the value of 128 for $m$. Predictions are made by the prediction table, which consists of the current temperature, the energy value of each bank, and saturating counters per bank. The locations of the banks cause changes in the temperature of each bank (e.g., stacked on a hot core). The cores show different behaviors depending on the workload and the location of the core. Hence, vertical heating caused by these cores results in increasing the temperature of the bank. In our prediction scheme, we consider the current temperature of the bank as an indication of its location. Moreover, some of the banks are more frequently accessed, which causes further temperature imbalance. It should be noted that the write accesses in STT-RAM has higher impact on the temperature than the reads. Specifically, the write energy per access ($E_{write/access}$) in STT-RAM is approximately 4 times that of a read ($E_{read/access}$). Hence, we need to distinguish the number of read and write accesses ($N_{read}$ and $N_{write}$). As a result, we calculate the energy value of a bank as $E_{write/access} \times N_{write} + E_{read/access} \times N_{read}$. In each sampling interval, we collect the number of reads, writes, and temperature 128 times. Hence, the total storage overhead is 1.5KB per bank. For each entry, if a bank has the maximum temperature compared to other banks, we increment the saturating counter of the bank by $w_T$. Similarly, if a bank has the maximum energy across banks we increment the counter by $w_E$. More specifically, since the location of a bank has a higher impact on the temperature compared to the number of accesses, we set the value of $w_T$ to be 2 and $w_E$ to be 1. In this case, we select a 9-bit saturating counter in our prediction table (18B storage overhead).

Once we check all the entries in our prediction table during the sampling interval, we select the maximum temperature of the bank that has the maximum counter value as our predicted temperature. The reason behind this is as follows: a bank which is hotter and highly accessed is more probable to keep its state in near future. Our experimental results, which we describe later, also confirm this intuition. The predicted value will be send to the flow rate adjustment component and all bank counters in the prediction table will be set to zero. Hence, the input to the controller is the predicted maximum temperature, and the output is the flow rate for the next interval. The flow rate adjustment component uses its flow rate lookup table to assign the flow rate. The lookup table includes the flow rate (per cavity) that should be applied when the maximum temperature is observed for the 3D system. We gathered these data based on existing analysis [9]. The lookup table is indexed by temperature values, and each line holds a flow rate value. At runtime, depending on the predicted maximum temperature, we pick the appropriate flow rate from the table. Note that the runtime overhead of using a look-up table based controller is negligible, considering that the cost is only limited to a lookup from a small-sized table.

TABLE I. System Configuration

| | |
|---|---|
| Core | 16-core, 4.0 GHz, out-of-order |
| SRAM caches | I-L1/D-L1: private, 32KB/32KB, 4-way, LRU, 128B line |
| LLC | SRAM: shared, 512KB per core, 2-way, LRU, 128B line<br>STT-RAM: shared, 6MB per core, 24-way, LRU, 16 banks, 128B line |
| DRAM | 4GB, 128-entry write buffer, 200-cycle |
| Network Router | 2-stage wormhole switched, virtual channel flow control, 6 VCs per Port, 5 flit buffer depth, 8 flits per Data Packet, 1 flit per address packet |
| Network Topology | 3D network, each layer is an 4×4 mesh, each node in layer 1 has a router, processor, private L1 cache, and SRAM bank, each node in layer 2 has an STT-RAM bank |

We assume a microchannel cavity layer for liquid flow, which is presented by Sridhar et al. [23]. We use uniformly distributed microchannels. In addition to dynamic flow rate adjustment, we also evaluate a cooling interlayer in TESLA with an equivalent fluid flow rate through each channel in the same layer to explore its impact. TESLA builds an insulation layer between processor and LLC dies to increase the thermal resistivity of the heat path from the LLC layer to the heat sink. Hence, TESLA prevents unwanted thermal crosstalk between the processor and LLC dies. We consider an oxidized macro porous Si [17] as TESLA insulation layer. Porous Si has 100x lower thermal conductivity than Si [19] and is 150 µm thick. We evaluate these structures in Section IV. The 3D TSV model parameters based on [5, 6] are used in our work. The TSV pitch size is reported to be <10 µm and we assume a 1024-bit bus. Also, TSVs are distributed uniformly between layers to keep the uniform thermal coupling between layers [30]. Heat sink is located at the top of the processor stack. The LLC banks are stacked on bottom of the processor layer. The reason that the heat sink is located over the processor layer is to reduce its temperature as much as possible. We assume 3.5 mm$^2$ area for each core at 32 nm technology based on a scaled version of the Rock core [25]. We also estimate 5.08 mm$^2$ and 1.51 mm$^2$ for the 6MB STT-RAM and

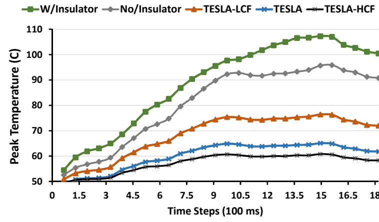


Fig. 4. Transient analysis of temperature fluctuations in the STT-RAM die for the bt application.

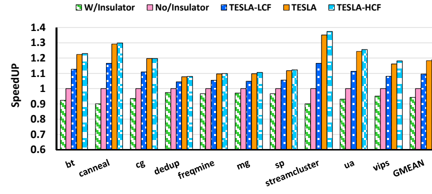


Fig. 6. Performance improvement for various workloads

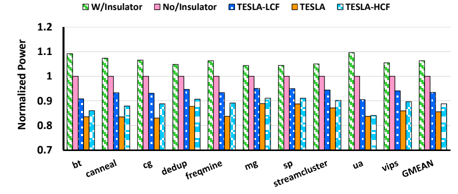


Fig. 7. Total power consumption for various workloads

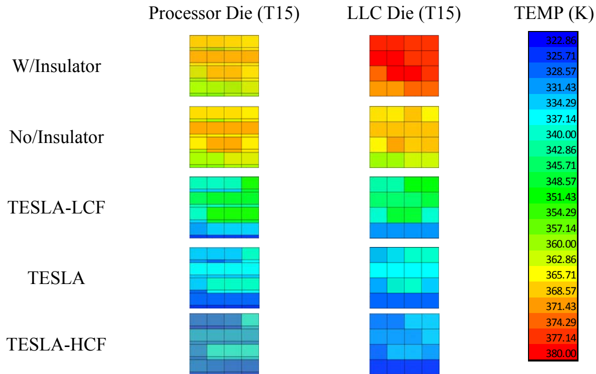


Fig. 5. Case study: Impact of thermal cooling on the LLC (i.e., STT-RAM) and Processor (Core+SRAM ) layer temperature for bt.

512KB SRAM, respectively. These parameters, which are conservative estimations based on [16], are generated by CACTI [2] and DESTINY. Based on estimates from CACTI and DESTINY, a 6MB STT-RAM bank has an area roughly equal to the area of one core with 512KB SRAM cache bank. In order to alleviate the impact of high write latency/energy of STT-RAM, we use 6MB STT-RAM in the cache layer and 512KB in the processor layer together as a hybrid LLC. We use an adaptive placement and migration policy based on the approach proposed by Wang et al. [26]: we put the write burst blocks to SRAM and the remaining blocks to STT-RAM.

## III. EXPERIMENTAL METHODOLOGY

To evaluate the impact of TESLA on a realistic multicore system, we model a 16-core processor fabricated using 32 nm SOI CMOS process based on Gem5 simulator [3] integrated with DRAMSim 2.0 [20]. Table I lists the system configuration. We collect runtime statistics from full-system simulations, and use them to calculate the power consumption using McPAT [14]. We calibrate the dynamic power numbers collected from McPAT based on the power dissipation data published for scaled version of the Rock core. Our cache power model adds the static and dynamic power of the caches used by a workload in the simulation. The static power is obtained from CACTI. We also use DESTINY for calculating the read/write latency and energy of our STT-RAM LLCs. The benchmarks we use in this study are chosen from NAS [4] and PARSEC [8].

We model the thermal characteristics of a 3D stacked architecture where the LLC die sits under the logic die using the 3D Interlayer Cooling Emulator (3D-ICE) [24]. The physical parameters for the thermal model are based on [5]. We model a multicore system that employs dynamic thermal management by throttling the voltage and the frequency of the chip to keep it within safe operational temperatures (below 90$^o$C) [22]. We estimate the temperature of the chip using the 3D-ICE. The estimated temperature is then used to refine the STT-RAM cache latencies. We also adjust the voltage and frequency of the logic die based on the stable-state power and temperature estimates, and repeat the process until the system reaches a stable state. The ambient temperature is fixed at 45$^o$C. STT-RAM layer is sensitive to vertical heating from the processor die (Figure 1). The exact interaction depends highly on the thermal decoupling solutions (cooling/insulation). Therefore, the thermal decoupling that TESLA advocates will be more important when better cooling/insulation solutions are employed. To evaluate the impact of TESLA across cooling solutions we model microchannel model based on [23]. We assume the height, channel length, and wall length to be 100 µm, 50 µm, and 50 µm, respectively [21]. We also assume uniformly distributed microchannels and a coolant volumetric heat capacity of 4.172e-12 J/µm$^3$K.

Three approaches we consider in this paper depend on the flow rate. In the first one, we assume an equivalent fluid with Low Constant Flow rate (TESLA-LCF) through each channel in the same layer (10 ml/min). In the second one, we use the variable flow rate mechanism we described in Section II (TESLA). Third one, High Constant Flow rate (TESLA-HCF), has a higher constant flow rate (32 ml/min). The fluid pump and valve consumes 1.3 W and 3.6 W per tier for 10 ml/min and 32 ml/min flow, respectively. The power consumption increases almost linearly according to the volumetric fluid flow [21]. We model high-aspect ratio TSVs with 10 µm diameter, located and etched within 100 µm wide microchannel walls as in [21, 24]. For the insulation layer, we consider a 150 µm thick Porous Si with 0.156 (W/mK) thermal conductivity.

## IV. EXPERIMENTAL RESULTS

### A. *Temperature Evaluation*

In this section, we evaluate the thermal shielding effect of the insulating/cooling layer by observing the temperature variation in the LLC die resulting from temperature fluctuations in the processor die. Figure 4 shows the transient analysis of the peak temperature of the STT-RAM die for the bt from NAS suite. As depicted in the figure, TESLA-HCF reduces temperature more than other techniques. This result is directly related to higher flow rate we used in this design. TESLA also reduces the temperature of the LLC efficiently. In addition, adjusting flow rate at run-time causes TESLA to be more energy-efficient compared to TESLA-HCF (we will describe this impact in Section IV.C). As we discuss further in the following section, the small difference between TESLA and TESLA-HCF ($<$ 4$^o$C at any time during simulation) does not have a significant impact on the performance of STT-RAM for TESLA, while TESLA-HCF consumes higher power. Compared to No/Insulator architecture, TESLA reduces temperature up to 30$^o$C because of the thermal shielding effect of the cooling layer. As we discussed in Figure 1, this change in temperature has a significant impact on the performance of the STT-RAM. More specifically, this difference increases cache hit, miss, and write latency by 7.6%, 3.9%, and 3.4%, respectively. Overall, TESLA allows for faster LLC access because it shields the STT-RAM layer from the temperature fluctuations occurring in the processor layer. Moreover, as seen in the figure, using Porous Si insulator to avoid vertical heating from the processor die to the LLC has a negative impact on the LLC die. Although insulation minimizes the heat flow from the processor layer to the STT-RAM, it also increases the thermal resistance from the STT-RAM to the heat sink and traps the heat within the LLC die.

Figure 5 shows the temperature distribution of the processor and STT-RAM during the simulation (at t=1.5 second) of the bt

application for various architectures. As depicted in the figure, TESLA can efficiently reduce the temperature of both processor and LLC die. We must also point that the processor and LLC layers show a similar behavior, except for the W/Insulator mechanism; the thermal shielding provided by Porous Si insulator avoid vertical heating from LLC to the processor die. As a result, the processor die is cooler than the LLC.

### *B. Performance Evaluation*

Figure 6 shows the performance of the evaluated schemes for various workloads. The results are normalized to the No/Insulator architecture. As seen from the figure, W/Insulator architecture degrades the performance by 5.9%, on average compared to No/Insulator. Moreover, TESLA-HCF, TESLA and TESLA-LCF achieve superior improvements compared to the No/Insulator by 19.1%, 18.2% and 9.6%, on average, respectively. The reason for these improvements is the direct result of using the liquid cooling layer between the processor and LLC dies. Moreover, TESLA improves the performance by 8.6%, on average, over the TESLA-LCF due to adjusting the flow rate to reduce the temperature efficiently. Also, as we will explain in Section IV.D, the TESLA prediction scheme is accurate and hence, it predicts more appropriate flow rate for future temperature and thus reduces the impact of response time. Note that as depicted in the figure, TESLA's architecture achieves up to 37% performance improvements over the architecture without the insulating layer for streamcluster workload. The streamcluster workload is a read intensive benchmark; 96.2% of the total LLC accesses are read requests. As we explained in Figure 1, the temperature changes have a higher impact on the LLC read latency compared to the write latency. As a result, reducing the temperature for this benchmark can cause a significant performance improvement.

### *C. Power Evaluation*

Figure 7 shows the total power consumption of the architectures analyzed in this paper. As seen in the figure, TESLA-LCF, TESLA, and TESLA-HCF reduce the power consumption by 6.7%, 14.6%, and 11.4% compared to the No/Insulator architecture, respectively. TESLA architectures reduce the LLC temperature efficiently as shown in the Figure 4. This reduces the leakage power consumption. Evaluation results obtained by DESTINY tool reveals that the leakage power of 96 MB STT-RAM changes from 20W to 31.2W when the temperature reaches 110ºC from 40ºC. It should be noted that cache hit, miss, and write dynamic energy of STT-RAM do not increase significantly when temperature changes. On average, TESLA achieves 7.9% and 3.2% power reduction compared to TESLA-LCF and TESLA-HCF, respectively. Compared to TESLA-LCF, although in some cases TESLA uses higher flow rate and indeed additional pump power, its accurate prediction and appropriate flow rate adjustment results in handling the temperature efficiently. In comparison with TESLA-HCF, TESLA shows a similar temperature profile (Figure 4). However, its accurate predictor and flow adjustment controller reduce the cooling power by up to 31%; TESLA-HCF uses extra power (up to 3.6 W) to handle fluid pump to cool the chip.

### *D. Coverage and Accuracy*

One of the key factors in evaluating the effectiveness of TESLA is the coverage and accuracy of the LLC temperature predictor. As long as the temperature predictions are accurate, our approach can efficiently adjust the flow rate. Mispredictions can either be false positives (over-predicting the temperature and unnecessarily increasing the flow rate) and false negatives (under-predicting the temperature and causing temperature increase). TESLA has a correct prediction rate of 86.3%; i.e., TESLA's predictions are within 1ºC of the actual temperature 86.3% of the time. The false positive error rate is 4.9% on average.

## V. CONCLUSION

In this paper, we proposed TESLA, a STT-RAM based 3D cache design to stabilize and reduce the temperature of the STT-RAM die in order to avoid STT-RAM performance degradation due to temperature changes. Elevated temperatures due to vertical heating and activity in 3D stack degrade the read and write speed of the STT-RAM. In this work, we use an interlayer liquid cooling between the processor and LLC die to thermally decouple the vertical heating between layers to reduce the STT-RAM temperature and avoid performance degradation. We also propose a novel technique to dynamically adjust the flow rate of the liquid interlayer cooling at run time to minimize pump energy consumption. Evaluation results reveal that our approach achieves up to 19.1% performance improvement and 14.6% power reduction over an architecture that does not have an insulator.

## REFERENCES


[1] "Laing 12 volt DC pumps datasheets.," http://www.lainginc.com/pdf/DDC3 LTI USletter BR23.pdf.

[2] "CACTI 6.5," http: //hpl .hp.com: research /cacti/.

[3] "Gem5 simulator," http://gem5.org/.

[4] "The NAS parallel benchmarks.," http://www.nas.nasa.gov/publications/npb.html.

[5] M. V. Beigi, and G. Memik, "TAPAS : Temperature-aware Adaptive Placement for 3D Stacked Hybrid Caches," in MEMSYS, 2016.

[6] M. V. Beigi, and G. Memik, "Therma: Thermal-Aware Run-Time Thread Migration for Nanophotonic Interconnects," in ISLPED, 2016.

[7] X. Bi, H. Li *et al.*, "STT-RAM Cell Design Considering CMOS and MTJ Temperature Dependence," *IEEE Transactions on Magnetics,* vol. 48, no. 11, pp. 3821 - 3824, 2012.

[8] C. Bienia, S. Kumar *et al.*, "The PARSEC benchmark suite: Characterization and architectural implications," in PACT, 2009.

[9] A. K. Coskun, D. Atienza *et al.*, "Energy-efficient variable-flow liquid cooling in 3D stacked architectures," in DATE, 2010.

[10] M. Imani, S. Patil *et al.*, "Low Power Data-Aware STT-RAM based Hybrid Cache Architecture," in ISQED, 2016.

[11] M. Imani, A. Rahimi *et al.*, "A Low-Power Hybrid Magnetic Cache Architecture Exploiting Narrow-Width Values," in NVMSA 2016.

[12] S. G. Kandlikar, "Review and Projections of Integrated Cooling Systems for Three-Dimensional Integrated Circuits," *ASME J. Electron. Packag.,* vol. 136, no. 2, 2014.

[13] J. Kim, A. Paul *et al.*, "Spin-Based Computing: Device Concepts, Current Status, and a Case Study on a High-Performance Microprocessor," *In Proc. of the IEEE,* vol. 103, no. 1, 2015.

[14] S. Li, J. H. Ahn *et al.*, "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures," in MICRO, 2009.

[15] G. H. Loh, "3D-stacked memory architectures for multi-core processors," in ISCA, 2008, pp. 453-464.

[16] A. K. Mishra, X. Dong *et al.*, "Architecting On-Chip Interconnects for Stacked 3D STT-RAM Caches in CMPs," in ISCA, 2011.

[17] B. Mondal, P. Basu *et al.*, "Oxidized macro porous silicon layer as an effective material for thermal insulation in thermal effect microsystems," *In Proc. of Emerging Trends in Electronic and Photonic Devices Systems*, 2009.

[18] M. Poremba, S. Mittal *et al.*, "DESTINY: A Tool for Modeling Emerging 3D NVM and eDRAM caches," in DATE, 2015.

[19] B. Pourshirazi, and Z. Zhu, "Refree: A Refresh-Free Hybrid DRAM/PCM Main Memory System," in IPDPS, 2016.

[20] P. Rosenfeld, E. Cooper-Balis *et al.*, "Dramsim2: A cycle accurate memory system simulator," *Computer Architecture Letters (CAL),* vol. 10, no. 1, 2011.

[21] M. M. Sabry, A. K. Coskun *et al.*, "Energy-efficient multiobjective thermal control for liquid-cooled 3-d stacked architectures," *TCAD,* vol. 30, no. 12, 2011.

[22] K. Skadron, M. R. Stan *et al.*, "Temperature-aware microarchitecture: Modeling and implementation," *ACM Transactions on Architecture and Code Optimization (TACO),* vol. 1, no. 1, pp. 94-125, 2004.

[23] A. Sridhar, A. Vincenzi *et al.*, "Compact transient thermal model for 3D ICs with liquid cooling via enhanced heat transfer cavity geometries," in THERMINIC, 2010.

[24] A. Sridhar, A. Vincenzi *et al.*, "3D-ICE: Fast compact transient thermal modeling for 3D-ICs with inter-tier liquid cooling," in ICCAD, 2010.

[25] M. Tremblay, and S. Chaudhry, "A Third-Generation 65nm 16-Core 32-Thread Plus 32-Scout-Thread CMT SPARC Processor," *In Proc. of ISSCC*, 2009.

[26] Z. Wang, D. A. Jiménez *et al.*, "Adaptive Placement and Migration Policy for an STT-RAM-Based Hybrid Cache," in HPCA, 2014.

[27] C. Xu, D. Niu *et al.*, "Overcoming the Challenges of Crossbar Resistive Memory Architectures," in HPCA, 2015.

[28] Yue Hu, Shaoming Chen *et al.*, "Effective Thermal Control Techniques for Liquid-Cooled 3D Multi-Core Processors," in ISQED, 2013.

[29] Y. Zhang, L. Zheng *et al.*, "3-D Stacked TierSpecific Microfluidic Cooling for Heterogeneous 3-D ICs," *IEEE Trans. Components, Packaging and Manufacturing Technology,* vol. 3, no. 11, 2013.

[30] Y. Zhang, Y. Zhang *et al.*, "Thermal design and constraints for heterogeneous integrated chip stacks and isolation technology using air gap and thermal bridge," *In IEEE Trans. Compo. Packag. Manuf. Technol,* vol. 4, no. 12, 2014.