Quantifying Architectural Impact of Liquid Cooling for 3D Multi-Core Processors[1]

Hyung Beom Jang*, Ikroh Yoon, Cheol Hong Kim***, Seungwon Shin**, and Sung Woo Chung***

Abstract—**For future multi-core processors, 3D integration is regarded as one of the most promising techniques since it improves performance and reduces power consumption by decreasing global wire length. However, 3D integration causes serious thermal problems since the closer proximity of heat generating dies makes existing thermal hotspots more severe. Conventional air cooling schemes are not enough for 3D multi-core processors due to the limit of the heat dissipation capability. Without more efficient cooling methods such as liquid cooling, the performance of 3D multi-core processors should be degraded by dynamic thermal management. In this paper, we examine the architectural impact of cooling methods on the 3D multi-core processor to find potential benefits of liquid cooling. We first investigate the thermal behavior and compare the performance of two different cooling schemes. We also evaluate the leakage power consumption and lifetime reliability depending on the temperature in the 3D multi-core processor.**

Index Terms—**Multi-core, 3D integration, thermal management, liquid cooling, architectural investigation**

Manuscript received Dec. 6, 2011; revised Jan. 9, 2012.
* Department of Computer and Radio Communication Engineering, Korea University, Seoul 136-713, Korea.
E-mail: {kuphy01, swchung}@korea.ac.kr
** Department of MechanicalEngineering, Hongik University, Seoul 121-791, Korea.
*** School of Electronics and Computer Engineering at Chonnam National University, Gwangju 500-757, Korea.
·Sung Woo Chung is the corresponding author of this paper.

I. INTRODUCTION

As technology scales down and integration densities continue to increase, interconnect delay has become a dominating factor of performance in recent microprocessors. Moreover, power consumption of the microprocessor is also increased due to long wire length. As one of the most efficient solutions to reduce interconnect delay as well as power consumption, three-dimensional (3D) integration has drawn much attention since it enables faster on-chip communication and lower power consumption by drastically reducing global wire length [31-33, 35, 45]. Moreover, increasing the number of stacked dies makes the 3D integration technique more beneficial.

Currently, even in 2D planar processors, thermal problems are serious [2, 19, 21], which cause unexpected functional errors or permanent damages. Thus, most 2D planar processors already have thermal management schemes, such as Dynamic Voltage and Frequency Scaling (DVFS), decode throttling and I-cache toggling [4]. Unfortunately, 3D processors have more severe thermal problems than 2D planar processors. It comes from the fact that the temperature of dies increases more significantly due to higher heat density as more dies are stacked in 3D processors. The longer heat dissipation path from the die to the heat sink is another reason for creating worse thermal problems in 3D processors. For these reasons, the 3D integration technique makes existing thermal hotspots worse and creates new thermal hotspots [45]. Reducing the localized temperature at hotspots is more important than reducing the average temperature across a processor [30], since Dynamic Thermal Management (DTM) [38] based on localized

temperature rather than the overall temperature degrades performance.

According to previous works on the thermal analysis of 3D processors [12, 17], thermal-aware techniques should be considered to sustain the temperatures of 3D processors below the thermal emergency. There have been many studies focused on the thermal optimization of 3D processors through floorplanning methods [1, 9, 10, 13, 15, 37]. The heat dissipation capability of functional units can be improved via the thermal-aware floorplanning methods, resulting in temperature reduction of 3D processors. Though the floorplanning approaches alleviate thermal problems, they are not efficient enough to eliminate thermal problems in 3D processors. As another approach, inserting thermal Through Silicon Via (TSV) into the vertically integrated processor was proposed to reduce on-chip temperature [44]. In this method, a large number of thermal TSVs are required to dissipate the heat flux from hot regions. In addition, thermal TSVs cannot be inserted directly into the hot regions of 3D processors since most of hot areas are occupied by macro blocks or functional units. As a result, thermal TSVs do not always reduce the peak temperature of a 3D processor below the thermal emergency.

The conventional air cooling method has a limited heat dissipation capability. If the 3D integration technique is applied in designing multi-core processors with conventional air cooling, the thermal problems are expected to be more serious, since the power density increases and heat dissipation distance becomes longer as multiple dies are stacked vertically. To mitigate thermal problems in 3D processors with more aggressive cooling methods, the interlayer liquid cooling methods using water as a coolant have been investigated by several researchers [6, 8, 22]. Especially, Brunschwiler et al. examined the heat dissipation capability of direct liquid cooling at practical operation conditions based on the most recent liquid cooling scheme proposed by IBM Corporation [6]. Additionally, the scalable interlayer cooling concept, using the correlation-based prediction method, was proposed to identify the optimal liquid cooling structure through the numerical modeling method [6]. Chen et al. demonstrated that inter-layer liquid cooling and die spacing resolved the thermal problem in the chip with a total heat dissipation of 25 W/cm^2 ~ 50 W/cm^2 [8]. Koo et al. also showed that the layer of integrated microchannel cooling with water as a coolant reduced heat density down to 135 W/cm^2 in a 3D processor with the maximum circuit temperature of 85 ºC [22].

As far as we know, [20] is the first study on the architectural effects of the liquid cooling scheme on a 3D processor, while the above-mentioned studies on liquid cooling are classified as mechanical engineering analyses. However, in [20], the performance impacts of the air cooling scheme and the liquid cooling scheme were not presented, since it takes too much time to evaluate the performance of each application with the Fluent Package (ICEPAK) simulation engine while considering the DTM triggers. In this paper, we analyze the impact of the liquid cooling scheme on 3D multi-core processors compared to the conventional air cooling scheme in the perspective of temperature, performance, leakage, and reliability, based on the most recent liquid cooling scheme proposed by IBM Corporation [6]. Especially, the performance impact in accordance with the cooling schemes is evaluated by using the temperature-aware DVFS technique [23]. In addition, we perform more thermal evaluations with representative applications to investigate the benefits of the liquid cooling scheme.

The rest of this paper is organized as follows. Section 2 explains related works on 3D integration and liquid cooling. Section 3 describes the evaluated 3D processor incorporated with the liquid cooling scheme. In Section 4, we provide a detailed description of our evaluation environments and modeling methodology. Section 5 analyzes our evaluation results on temperature, performance, leakage, and reliability. Section 6 concludes this paper.

II. RELATED WORKS

1. 3D Integration Techniques

1) Structure of 3D Integration: Researches for 3D integration techniques are categorized into two major groups: (1) Multi-layer Buried Structures (MLBS) and (2) die-bonding techniques [42]. In MLBS, the process of building active device layers is repeated on a single wafer to compose multiple dies before processing all the metal routing layers. 3D integration techniques enable

mixing dissimilar process technologies, such as high-speed CMOS and high-density DRAM by using MLBS, which make it possible to stack many heterogeneous dies [25, 32]. While MLBS require changes in the manufacturing process, die-bonding 3D integration techniques insert metal vias to bond the two planar dies based on the conventional 2D manufacturing process. Various bonding materials have been studied [36]. Die-bonding 3D integration techniques are classified into three different topologies for interfacing between multiple planar dies; face-to-face (F2F), face-to-back (F2B), and back-to-back (B2B). In this paper, we just consider the face-to-face bonding technique because it provides a dense interface between adjacent dies and enables various combinations for 3D processor organizations.

2) Through-Silicon Vias (TSVs) in 3D Integration: In the 3D integration structure, through-silicon vias (TSVs) are usually vertically used for electrical interconnections across multiple layers. In the past, multiple layers were connected by wiring at their edges. However, TSVs replace the edge-wiring interconnection by the vertical interconnection, leading to significant reduction of the total wire length. It has been reported that the vertical latency for traversing the height of a 20-layer stack is just 12 ps [28]. Therefore, the interconnection using TSVs enables fast on-chip communications for 3D processors. Though TSVs are wider than typical metal wires, TSVs have very short height, since each wafer is thinned to only tens of microns. After fabrication, each wafer is thinned to only 10-100 μm in thickness [3, 14], and then TSVs that have pitches of only 4-10 μm 0 are etched through a bulk silicon. Lastly, thermo-compression is used to bond individual layers together to form the 3D integration structure [27]. In this paper, we consider TSVs for the vertical interconnections between two different dies. The number of TSVs and the location of TSVs are explained in sub-section 3.1.

3) Thermal Management in 3D Processors: Various techniques have been proposed for the thermal management of 3D processors. One of them is the DTM technique that has been widely used for 2D planar processors. In [4], thermal control techniques, such as DVFS and decode throttling were applied to 2D planar processors for DTM. In DVFS, the clock frequency is dynamically scaled along with the supply voltage to reduce processor power, resulting in temperature reduction. On the other hand, in decode throttling, processor core is throttled by restricting the flow of instructions for reducing power and temperature. For 3D processors, the thermal control techniques used for 2D processors can be adopted. Sun et al. proposed the 3D-Wave scheme to tackle the temperature problems in 3D multi-core processors [41]. They balanced the power consumption of each core by scaling the supply voltage to the optimal point, resulting in power reduction. Additionally, the migration algorithm was proposed to find the appropriate cores for detected hot tasks by considering the physical location of cores in a 3D processor. Through this scheme, they reduced the peak temperature of each core efficiently.

Another technique for managing the temperatures of 3D processors is the temperature-aware job allocation proposed by Coskun et al. [11]. They introduced the adaptive temperature-aware job allocation algorithm, which adjusts and balances application loads considering the location and the thermal behavior of each core. In the latter scheme, less intensive loads are assigned to cores that can be easily heated up due to their location by using the thermal index reflecting the location of each core. In addition, the scheduler assigns tasks considering the average temperature of each core; the average temperature reflects the thermal behavior based on the history window of each core. They reported that their technique reduces hotspots with little performance degradation. Moreover, this technique combined with the conventional DVFS technique enables more significant temperature reduction.

On the other hand, Cong et al. proposed the thermal-aware floorplanning method [9] based on the simulated annealing algorithm that has been widely used for floorplanning. By adding bucket structures to a 2D planar processor, they represented the vertical temperature information of a 3D processor. Considering vertical temperature information, they utilized the simulated annealing algorithm to find the optimal placement in 3D spaces. Hung et al. proposed another 3D floorplanning method considering the wafer-bonding technique [16]. Besides considering the area and the wire length, they also took into account the maximum power density and the vertical heat flow to explore the thermal-aware floorplanning for a 3D processor. They modified the B*-

tree floorplanning model originally proposed for 2D planar processors. Puttaswamy and Loh introduced a thermal-aware 3D microprocessor architecture called by ‘thermal herding’ [34]. They reported that placing hot dies near a heat spreader and a heat sink is helpful to easily emit heat flux.

Previously proposed thermal-aware techniques for 3D processors should be combined with DTM to avoid the thermal emergency, resulting in inevitable performance degradation. Considering this situation, a paradigm shift from the conventional air cooling scheme to the liquid cooling scheme is necessary for 3D processors.

2. Liquid Cooling Techniques

Conventionally, 2D planar high-performance micro-processors have relied on the air cooling scheme to avoid the thermal emergency. Fig. 1(a) depicts the conventional air cooling scheme in a 3D processor, where a heat spreader is used for dissipating heat flux. However, the vertical integration of multiple layers needs more cooling capacity than the conventional air cooling scheme. For 3D processors, the liquid cooling scheme is one of the most efficient cooling methods due to its superior heat dissipation efficiency. Generally, liquid cooling techniques are categorized into two major types; one is indirect cooling and the other one is direct cooling.

1) *Indirect Liquid Cooling:* In the indirect cooling technique, a coolant does not directly have contact with electronic components. Instead, intermediate structures, such asmicrochannels, are commonly used. Microchannels are integrated into a substrate or a heat sink and then inserted between layers. Heat is transmitted by the intermediate structure through a coolant [12, 22, 29].

2) *Direct Liquid Cooling:* Different from the indirect cooling technique, a dielectric coolant is allowed to pass between layers, as shown in Fig. 1(b). In other words, a coolant directly absorbs the heat flux from each layer of a 3D processor. In this case, a hermetic sealing of electrical TSVs is necessary to use water as a coolant [6]. In general, direct liquid cooling offers higher heat dissipation capability than indirect liquid cooling, resulting in more significant temperature reduction in 3D processors. Direct liquid cooling also provides more balanced thermal behavior within a die than the conventional air cooling scheme.

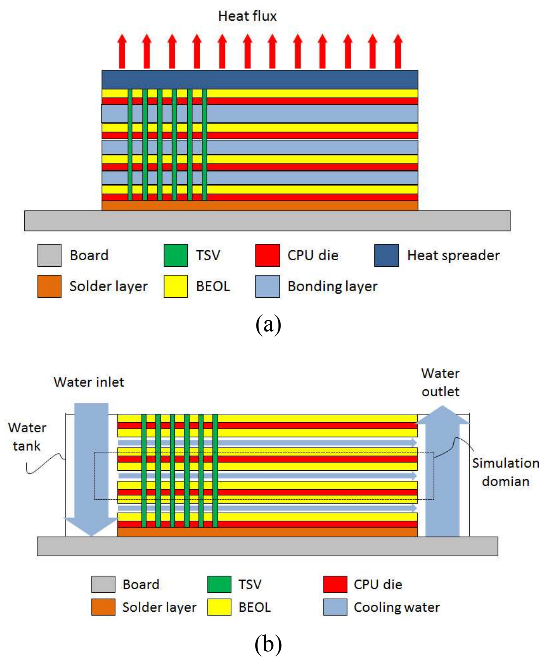


Fig. 1. Conceptual schematic of (a) air cooling, (b) liquid cooling.

III. 3D Processor Incorporated with the Liquid Cooling Scheme

1. Structure Overview of the 3D Processor

Fig. 2 shows the processor layout for a conventional planar processor sharing an L2 cache and the corresponding layout for the 3D processor. For the 2D planar processor, we model a processor based on the

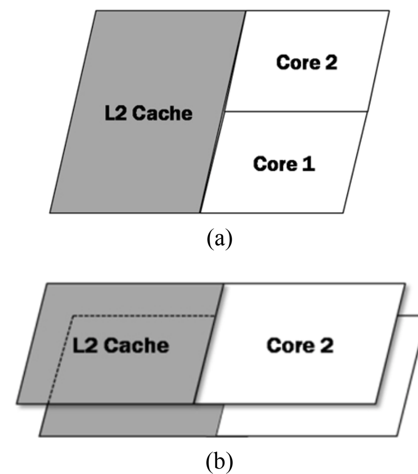


Fig. 2. (a) Planar layout of cores and an L2 cache, (b) 3D implementation of cores with split L2 caches.

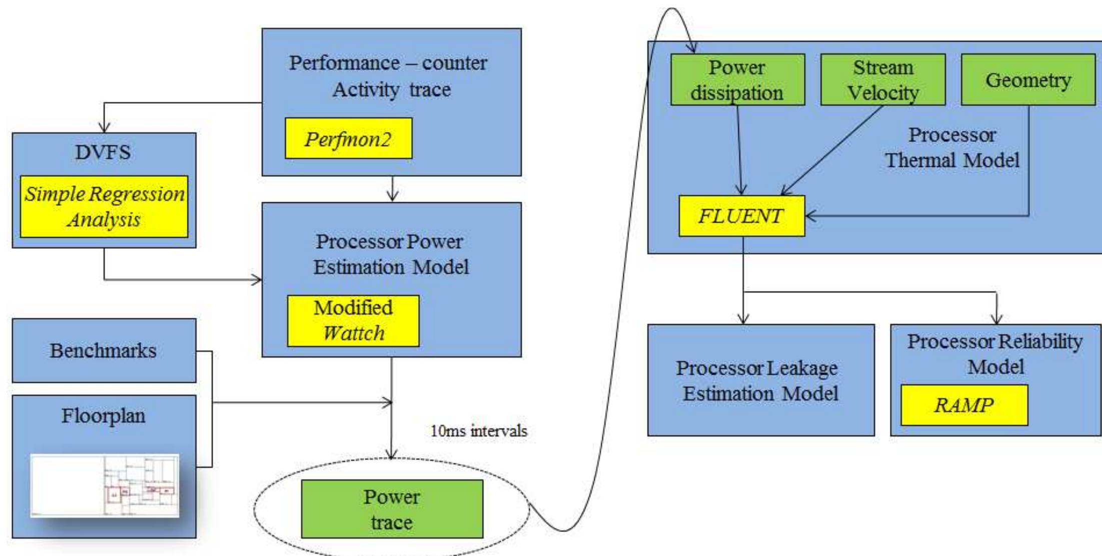


Fig. 3. Overall evaluation methodology.

Intel Core 2 microarchitecture according to [34]. For the 3D processor, we partition and stack one core and half of the L2 cache on each of the die [32, 42]. Though there are alternatives to stack the cores and L2 caches, we consider the 3D processor which has one core and half of the L2 cache on each of the die. As the processor technology is advanced, the power density is also increased resulting in more severe thermal problems. Therefore, to consider the processor which has high power density, we compose the 3D processor as shown in Fig. 2(b). The majority of the 3D processor layout is identical to the layout of the 2D planar processor except for total area reduction by the partitioned implementation of two cores on each die. The area used for each core (one die) is 27.4 mm^2, and the area per L2 cache (one die) is 27.4 mm^2. The total area per each layer is 54.8 mm^2. The 3D integration technique reduces the total area by approximately 50% and also reduces wire delay. The thickness of each layer in the evaluated 3D processor is modeled based on [3].

In our 3D processor, TSVs are also considered. The diameter of each TSV is set to 10 μm and the space requirement between TSVs is 10 μm. We assume that 64 TSVs are used for vertical interconnections and these TSVs are concentrated around the L2 cache.

2. Liquid Cooling Scheme

We adopt the direct liquid cooling structure proposed by IBM Corporation [6]. The direct liquid cooling scheme is depicted in Fig. 1(b), where four dies are stacked symmetrically with TSVs and the coolant is pumped in-between individual dies. The direct liquid cooling scheme is scalable to multiple-die stacked 3D processors. In our evaluation, we consider the two-die stacked 3D processor. Our two-die stacked 3D processor is equipped with the coolant containment, and the insulated channel is used for a path of the coolant. We use water as a coolant due to its superior heat dissipation capability. We assume that the coolant is fed at a fixed flow rate from a pump to the channel.

IV. EVALUATION METHODOLOGY

Fig. 3 shows our overall evaluation methodology. To compare the architectural effects of two different cooling schemes (the conventional air cooling scheme and the liquid cooling scheme) in the perspective of temperature, performance, leakage and reliability, we integrate architectural performance, power, and reliability tools with a mechanical thermal tool.

We first evaluate the power consumption of each functional unit. Based on the results of power consumption, we obtain the temperature of each functional unit (thermal behavior) from a thermal simulator. From the temperature results, we also estimate the processor leakage power by using the widely used leakage model. Lifetime reliability is estimated by an

architectural reliability simulator. Additionally, to compare the impact of DTM triggers on our 3D processor, we use the temperature-aware DVFS technique. Details of the power model, thermal model, leakage estimation model, and reliability model will be explained in the following sub-sections.

From 26 SPEC2000 benchmarks [50], we select ten applications (*art*, *bzip2*, *crafty*, *gcc*, *gzip*, *mcf*, *mesa*, *parser*, *twolf*, and *vortex*) that are representative applications [38]. We run the same application on both cores simultaneously to accurately evaluate the impact of the liquid cooling scheme on our 3D processor.

1. Power Modeling

To evaluate the power consumption of each functional unit, we use Wattch [5] and perfmon2 [46]. We first modified Wattch based on the empirical power model proposed by Isci et al. [18] to read the performance counts representing the activity factor of each functional unit. We also modify Wattch to configure the Core 2 processor since Wattch was originally implemented to operate with Pentium 4 processors.

The Intel Core 2 processor provides three fixed-function performance counters and two general-purpose performance counters for counting events. Each performance counter is related with the configuration register, which determines the event to be counted. The perfmon2 enables both configuring the configuration register and reading the performance counters. Among 129 events, we utilize 17 events for estimating the activity factor of each functional unit. We use the power values of each functional unit from [26]. Fig. 4 shows the floorplan of one die for the evaluated 3D processor.

We also consider the power consumption of the flowing coolant. The flowing coolant power is mainly dependent on the coolant flow rate, since the pressure drop of the coolant is highly different from the coolant flow rate. Consequently, the flowing coolant power consumption is approximately proportional to the cube of the coolant flow rate, and it is calculated as follows [43]:

$$P_{flowing_coolant} = \Delta Pressure \cdot Q \qquad (1)$$

where ΔPressure is the pressure drop calculated from the numerical simulation and Q is the volumetric flow rate. In our evaluation, the power consumption for the forced flowing coolant is only 0.823 W due to the low volumetric flow rate (6.55×10^{-3} m^3/hr), which is enough to absorb the heat flux of the two-die 3D processor. Note that the flowing coolant power consumption used in our evaluation is acceptable compared to the flowing coolant power consumption in [7]. The processor leakage power estimation model used for our 3D processor is explained in sub-section 4.4.

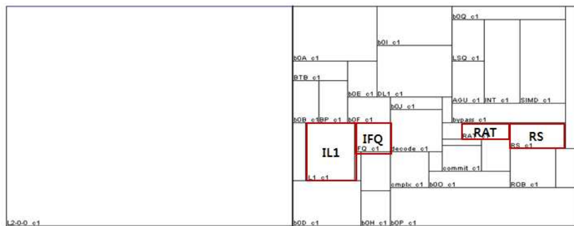


Fig. 4. Floorplan of a die in the 3D processor.

2. Thermal Evaluation Modeling

To investigate the heat dissipation capability of the 3D processor incorporated with the liquid cooling scheme, we use the Fluent Package (ICEPAK) simulation engine [51] that analyzes the detailed heat flows in the mechanical engineering community. Each of the die composed of one core and half of the L2 cache is mounted at the board by the Flip-chip Ball Grid Array (FBGA) type and the components of the package (498-solderball, Flip-chip under-fill, and Flip-chip substrate) are modeled as a single combined layer that has an equivalent thermal resistance. The other thermal parameters for the 3D processor are obtained from [3]. For thermal evaluations, 4.3 million and 8.1 million non-uniform grid elements are used for modeling the conventional air cooling scheme and the liquid cooling scheme, respectively. To more accurately model the liquid cooling scheme by considering the coolant flow, we need to increase the number of grid elements. However, for the air cooling scheme, we do not need to increase the number of grid elements, since the 4.3 million non-uniform grid elements are enough to model the conventional air cooling scheme without diminishing the accuracy.

In the air cooling scheme, the coolant material is the ambient air of 293.15K temperature passing through a

heat spreader and a heat sink. We assume that the constant convective heat transfer coefficient is 15000 W/m^2K, which is sufficiently high enough to represent a well-designed air cooling scheme at the upper side of a heat spreader. On the other hand, in the evaluated liquid cooling scheme, we assume that the coolant (293.15K water) flows in one-dimension through in-between individual dies at a constant mass flow rate. Additionally, the heat flux of all functional units on each die is considered individually, since heat flux of each functional unit is different depending on its corresponding power consumption.

3. Performance Evaluation Modeling

For DTM, we adopt the temperature-aware DVFS technique proposed by [23]. The temperature-aware DVFS technique finds the correlation between the activity factor of one specified functional unit and its corresponding temperature. Since temperature is proportional to power and power is also proportional to voltage squared, temperature is consequently proportional to the voltage squared. Based on these observations, a simple regression analysis is conducted between the temperature and activity factor multiplied by voltage squared. Since there is no available 3D processor, the temperature is simulated from the Fluent Package explained in the previous sub-section. Additionally, in our 3D processor model, there are two dies dissipating heat flux. Thus, we consider the activity factors of two vertically aligned functional units for a simple regression analysis.

DVFS is implemented by using 4-step frequency and voltage pairs in the Intel Core2 processor [47, 48]. Thus, both frequency and voltage should be adjusted simultaneously. The voltage value which corresponds to each frequency level is acquired based on NHC (Notebook Hardware Control) [49]. Table 1 shows these frequency and voltage pairs. We assign the lowest frequency and voltage pairs (1 GHz, 0.95 V) as step 1 and the highest frequency and voltage pairs (2 GHz, 1.237 V) as step 4 as shown in Table 1.

Though we are able to investigate the temperatures of all functional units, we just focus on the temperature of the L1 instruction cache since it is the hottest functional unit in our evaluation and it also shows the most significant temporal thermal variation. We set the emergency temperature to 358.15K (85 ºC) as introduced in [38]. The trigger value X, in our simple regression formula, is determined corresponding to the emergency temperature of 358.15K. Each DVFS step has different trigger values for the access ratio of L1 instruction cache (the L1 instruction cache is the hottest functional unit in our evaluation). By modifying the Linux 2.6.25 kernel, we adjust the DVFS step referring to the activity factor in the Intel Core2 processor. When the L1 instruction cache access ratio is bigger than the trigger value, the DVFS step in Table 1 is lowered by two steps. However, when the L1 instruction cache access ratio is less than 80% of the trigger, the step is raised by one. In the other cases, the DVFS step is not changed.

Table 1. DVFS Steps

Step	Frequency(MHz)	Voltage(V)
Step 4	2000	1.237
Step 3	1667	1.137
Step 2	1333	1.05
Step 1	1000	0.95

4. Temperature-Dependent Leakage Modeling

We also examine the leakage gain from the liquid cooling scheme compared to the air cooling scheme. In general, the leakage power mainly consists of the sub-threshold leakage power and gate leakage power [24].

For estimating the sub-threshold leakage power, we consider two types of leakage; one is leakage in the logic circuits including functional units and the other is leakage in SRAM-based units such as caches and register files. The leakage power of logic circuits is calculated as the product of the number of gates and the average sub-threshold leakage current per gate. On the other hand, the leakage power of SRAM-based units is the sum of the SRAM memory cells' leakage power and their peripheral circuits' leakage power. The gate leakage power is calculated from the current between gates and channels and the gate direct tunneling current – including the tunneling current between a gate and a substrate.

In this paper, we take into account both the sub-threshold leakage and gate leakage of the logic circuits and SRAM-based units to reflect the thermal effects on the leakage power. We extract the leakage power values of 65nm technology [52] for the Intel Core 2 processor.

5. Reliability Evaluation Modeling

For evaluating the lifetime reliability of our processor model, we use the application-aware architecture-level methodology called the Reliability Aware Microprocessor (RAMP) [40]. RAMP dynamically tracks lifetime reliability depending on application behaviors. This methodology represents the processor lifetime reliability in terms of the Mean Time to Failure (MTTF) or the expected lifetime of a processor, and it calculates an instantaneous MTTF based on the current temperature and the utilization of each functional unit. In this paper, we evaluate the lifetime reliability of the 3D processor with liquid cooling compared to that with air cooling in the perspective of the MTTF represented by the Failures in Time (FIT), which means the number of failures per 109 device operating hours.

There are four models for evaluating the reliability of a processor in RAMP [40]. The first model is electromigration (EM) that occurs in aluminum and copper interconnects due to the mass transport of conductor metal atoms in interconnects. EM is exponentially dependent on temperature. The MTTF from EM is as follows [40]:

$$MTTF_{Electro\ Migration} \propto (J - J_{critical})^{-N} e^{\frac{E_{aEM}}{kT}} \tag{2}$$

where J is the current density in the interconnect, $J_{critical}$ is the critical current density needed for EM, E_{aEM} is the activation energy for EM, k is the Boltzmann's constant, and T is absolute temperature in Kelvin. N is the constant depending on the interconnect metal.

The second model is the stress migration (SM) that occurs when the metal atoms in interconnects migrate. It is caused by the different thermal expansion rates of different materials in the device. It is proportional to temperature changes. The MTTF from SM in RAMP is as follows [40]:

$$MTTF_{Stress\ Migration} \propto |T_0 - T|^{-N} e^{\frac{E_{aSM}}{kT}} \tag{3}$$

where T is the operating temperature and T_0 is the metal deposition temperature in Kelvin. N and E_{aEM} are constant values depending on materials and k is the Boltzmann's constant.

The third model is the time-dependent dielectric breakdown (TDDB), or the gate oxide breakdown. The gate dielectric wears down with time and fails when a conductive path is formed in the dielectric. The advent of thin and ultra-thin gate oxides coupled with the aggressive scaling of supply voltage accelerates the TDDB failure rates. The TDDB model highly relies on voltage and temperature. The MTTF from TDDB at specific temperature (T) and voltage (V) is given as follows [40]:

$$MTTF_{\substack{Time-Dependent \\ Dielectric\ Breakdown}} \propto (\frac{1}{V})^{(a-bT)} e^{\frac{(X+\frac{Y}{T}+ZT)}{kT}} \tag{4}$$

where $a, b, X, Y,$ and Z are constant parameters based on data in [40], and k is the Boltzmann's constant.

The last model in RAMP is thermal cycling (TC). All parts of the device experience with fatigue damages when there is a thermal difference. The accumulated damages due to temperature variations eventually cause failures. Thus, TC causes this kind of failure in the package and the die interface, such as solder joints. The MTTF due to TC is given as follow [40]:

$$MTTF_{Thermal\ Cycling} \propto (\frac{1}{T - T_{ambient}})^q \tag{5}$$

where q is the value of the Coffin-Manson exponent constant, T is the average temperature of the structure and $T_{ambient}$ is the ambient temperature.

In addition to the existing four failure models in RAMP, we also consider the negative bias temperature instability (NBTI) for the emerging critical failure model. NBTI takes place when the gate is negatively biased with respect to the source and the drain. The equation used to determine the MTTF at the temperature, T from NBTI [39] is:

$$MTTF_{\substack{Negative\ Bias \\ Temperature\ Instability}} \propto \left[\left(ln\left(\frac{A}{1+2e^{\frac{B}{kT}}}\right) - ln\left(\frac{A}{1+2e^{\frac{B}{kT}}} - C\right)\right) \times \frac{T}{e^{\frac{-D}{kT}}}\right]^{\frac{1}{\beta}} \tag{6}$$

where A, B, C, D, and β are constant values based on [39], and k is the Boltzmann's constant value.

We evaluate the lifetime reliability of the 3D processor depending on the cooling schemes by using the above five models.

V. Evaluation Results and Analysis

1. Temperature

Fig. 5(a) and Fig. 5(b) show the steady state thermal profiles of the 3D processor with the conventional air cooling scheme and the liquid cooling scheme, respectively, in the case of *gcc* application. As shown in the figures, the L1 instruction cache (IL1) is turned out to be a hotspot with the air cooling scheme; the L1 instruction cache is known as the hottest functional unit in the Intel Core 2 microarchitecture [34]. However, when we use the liquid cooling scheme, the temperature of IL1 is reduced drastically. In our evaluation, the coolant flows from left (IL1) to right (rename table (RAT), instruction fetch queue (IFQ), and reservation station (RS) in Fig. 4) absorbing the heat flux of IL1. Thus, RAT, IFQ, and RS are cooled down with the relatively warm coolant that already went through IL1, resulting in less thermal reduction. If they were placed at the upper side or lower side of IL1 in Fig. 4, then their temperature would be further reduced.

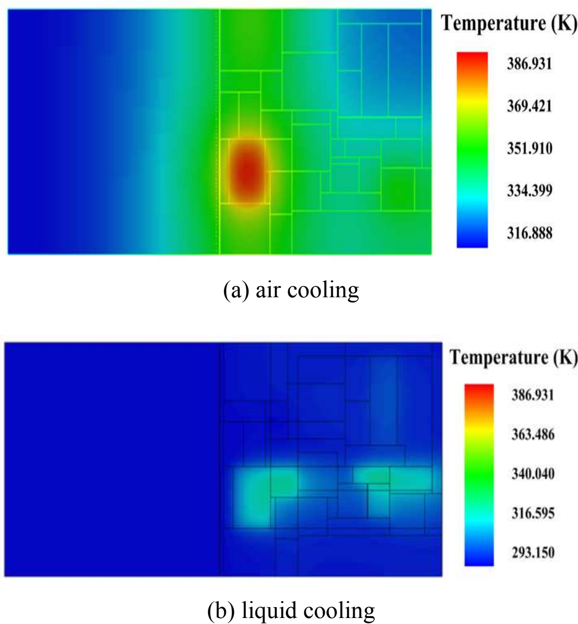


Fig. 5. Steady state thermal profiles of the 3D processor (a) air cooling, (b) liquid cooling.

Additionally, the liquid cooling scheme also shows smaller thermal fluctuation than the air cooling scheme in the 3D processor. In the steady state thermal profiles shown in Fig. 5, the air cooling scheme shows high spatial thermal fluctuation. However, in the case of the liquid cooling scheme, the overall temperature is drastically reduced and there is no hotspot, leading to performance improvement due to the reduced number of DTM invocations.

Fig. 6 shows the peak temperature of the 3D processor with two different cooling schemes. As shown in Fig. 6, the liquid cooling scheme reduces the peak temperature of the 3D processor compared to the air cooling scheme by 72 degrees, on average. Due to the superior heat dissipation capability of the liquid cooling scheme, thermal violations that the temperature of the processor goes over the pre-defined emergency temperature do not occur at all, resulting in no invoked DTM operation. Therefore, we do not include the result of the liquid cooling scheme with the DTM technique in Fig. 6. On the other hand, DTM operations implemented by the DVFS technique reduce the temperature of the 3D processor incorporated with the air cooling scheme by 39 degrees, on average. Peak temperature reduction by the DTM technique is only about 45% compared to that by the liquid cooling scheme. From our simulation results, we can determine that the peak temperature of the 3D processor with the liquid cooling scheme is efficiently controlled compared to that with the conventional air cooling scheme even if the DTM technique is adopted. Note that the air cooling scheme with the DTM technique

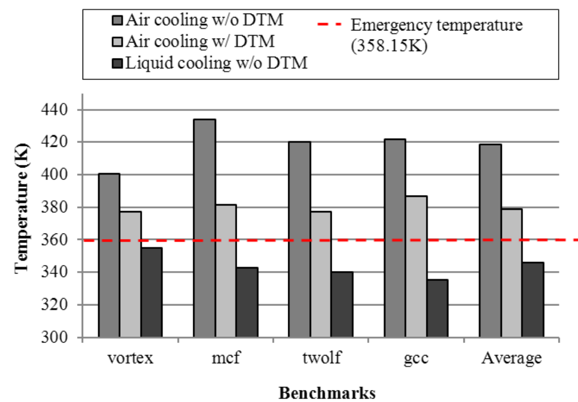


Fig. 6. Peak temperature comparison.

goes over the emergency temperature only for a moment, as shown in Fig. 7. In general, a moderately high temperature does not seriously hurt the processor reliability unless it goes over the emergency temperature for a long time. However, the air cooling scheme with the DTM technique should suffer performance degradation to control the temperature. On the other hand, the liquid cooling scheme always keeps the peak temperature of the 3D processor below the emergency temperature, eliminating the need for DTM operations.

We also evaluate the temporal temperature change of IL1 depending on the cooling schemes. As shown in Fig. 7(a), when *mcf* application is executed, the liquid cooling scheme reduces the temperature of IL1 by 75 degrees compared to the air cooling scheme, on average. Even though the DTM technique is applied to the air cooling scheme, temperature reduction does not reach that of the liquid cooling scheme. In our evaluation, the DTM technique reduces temperature by 33 degrees with the air cooling scheme, on average. Fig. 7(b) shows the temporal temperature change with *twolf* application. The DTM technique reduces the temperature of IL1 in the air

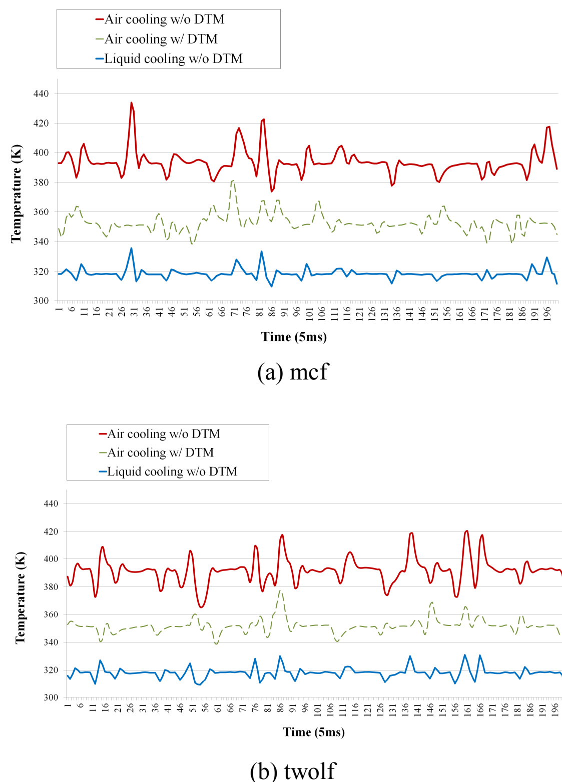


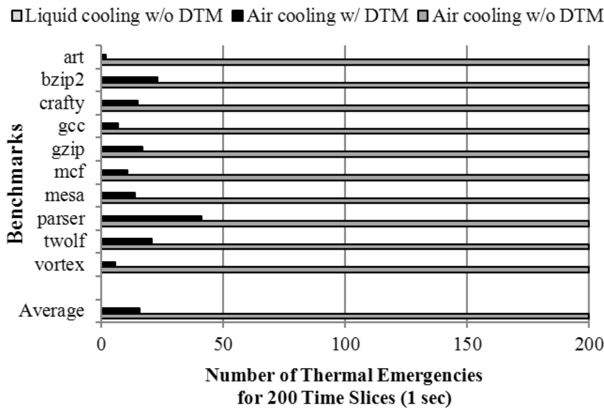
Fig. 7. Temperature comparison of the hottest functional unit (IL1) between the conventional air cooling scheme and the liquid cooling scheme.

cooling scheme by 33 degrees, on average. However, the temperature of IL1 is reduced by as much as 73 degrees with the liquid cooling scheme.

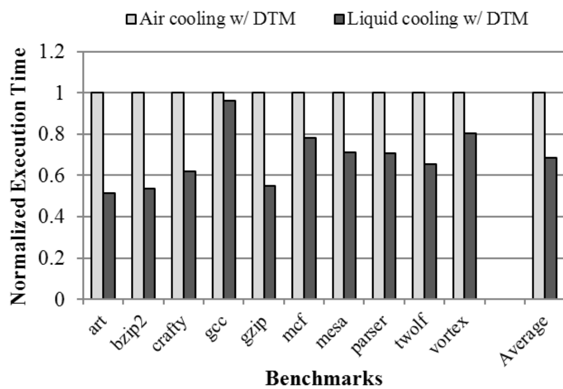
Though the results are not shown in the figure, we also investigate temperature reduction by the liquid cooling scheme for another two hottest functional units (IFQ and RS) in addition to IL1. In the case of *gcc* application, the temperatures of IFQ and RS are reduced by 35 degrees and 31 degrees, respectively, on average. In *gzip* application, as much as 31 degrees and 26 degrees are reduced, respectively, on average. There are also temperature reductions of IFQ and RS by as much as 54 degrees and 39 degrees, respectively, on average, with *twolf* application. In *mcf* application, the temperature of IFQ is reduced by as much as 55 degrees and the temperature of RS is decreased as much as by 41 degrees, on average. Our experimental results also show that the volumetric flow rate used in the liquid cooling scheme, which is 6.55×10^{-3} m^3/hr, is enough to reduce the temperature of the hottest functional units below the emergency temperature (358.15K).

2. Performance

In our evaluation, to sustain the temperature of the 3D processor below the thermal emergency, we adopt the DVFS technique for DTM. In the case of the conventional air cooling scheme, frequent DTM operations are invoked due to the low heat dissipation capability, resulting in performance degradation. In Fig. 8(a), the number of time slices when the temperature of the 3D processor is higher than the emergency temperature (358.15K) is presented while the application is executed during 200 time slices (1 time slice=5 ms). The DTM technique reduces the number of the thermal violations in the conventional air cooling scheme by 92.2%, on average. Though the DTM technique is employed to avoid thermal violations, all thermal violations cannot be avoided in our 3D processor. The reason is that the thermal violations in our 3D processor incorporated with the conventional air cooling scheme can occur even in the case of the lowest DVFS step (1 GHz, 0.95 V). In other words, the lowest DVFS step is not enough to avoid thermal violations due to the prior power-consuming operations. However, the thermal violation does not occur at all in the liquid cooling



(a) The number of time slices when the temperature is higher than the emergency temperature for 200 time slices (One time slice is 5ms).



(b) Normalized execution time of the 3D processor

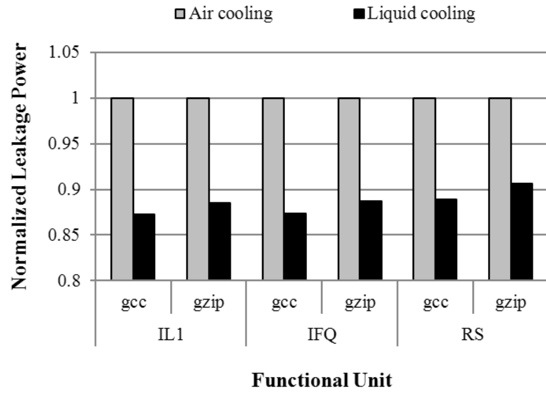
Fig. 8. Performance comparison between the conventional air cooling scheme and the liquid cooling scheme.

scheme. From these results, we can determine that the heat dissipation capability of the liquid cooling scheme is high enough to sustain the temperature of the 3D processor below the emergency temperature.

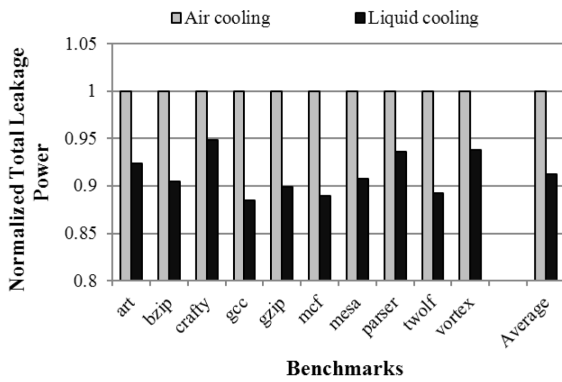
Fig. 8(b) shows the execution time (performance) that is highly related with the number of thermal violations. In Fig. 8(b), the execution time of the 3D processor when we use the liquid cooling scheme with the DTM technique is normalized to that of the air cooling scheme with the DTM technique. Note that the execution time of liquid cooling is the same as the execution time without DTM, since there is no DTM operation in the case of liquid cooling. Due to the high temperature of the 3D processor with the air cooling scheme, the DTM technique frequently lowers the DVFS level to the lowest one, resulting in longer execution time compared to the liquid cooling scheme. With the air cooling scheme, the performance degradation by DTM ranges from 4% (*gcc*) to 49% (*art*); the average performance degradation is 32%. In the case of *art* application, though the number of thermal violations in temperature is quite small (Fig. 8(a)), the performance is degraded by 49% (Fig. 8(b)). The reason is that the processor operates at the lowest DVFS step for most of the execution time to sustain the temperature. In *vortex* application, the number of time slices exceeding the thermal emergencies is similar to that of *gcc*. However, the performance is degraded by 20% when *vortex* application is executed, since the DTM technique frequently lowers the DVFS level to sustain the temperature of the 3D processor below the emergency temperature. On the other hand, when *gcc* application is executed, the 3D processor operates at a higher DVFS step for most of the execution time compared to when *vortex* application is executed. Thus, the performance is degraded by only 4% when *gcc* application is executed. From the evaluation results, we can determine that the liquid cooling scheme significantly reduces the temperature of the 3D processor with no performance degradation.

3. Leakage

Since the leakage power consumption is dependent on temperature, the temperature decreased by the liquid cooling scheme leads to leakage reduction. In this sub-section, we investigate the leakage consumption of the three hottest functional units (IL1, IFQ, and RS) and the overall processor leakage consumption. Fig. 9(a) shows the normalized leakage power for the three hottest functional units. In the case of *gcc* application, the liquid cooling scheme reduces the leakage power of IL1, IFQ, and RS by 12.8%, 12.7%, and 11.1%, respectively, on average. In *gzip* application, the liquid cooling scheme reduces the leakage power by as much as 11.5% (IL1), 11.3% (IFQ), and 9.4% (RS), respectively, on average. Accordingly, the liquid cooling scheme also reduces the total processor leakage power, as shown in Fig. 9(b). More than 10% total processor leakage reduction is observed with *gcc*, *gzip*, *mcf*, and *twolf* applications. The average total leakage reduction of ten applications is 8.8%.

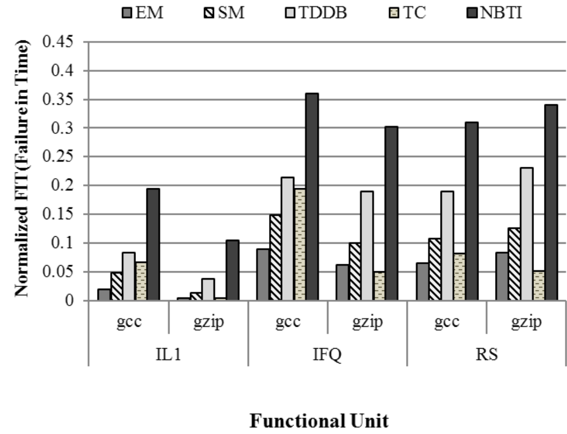


(a) Normalized leakage power for the three hottest functional units (L1 instruction cache (IL1), instruction fetch queue (IFQ), and reservation station (RS))

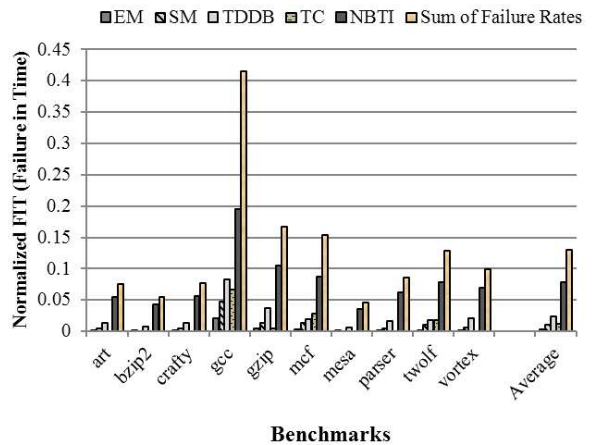


(b) Normalized total processor leakage power

Fig. 9. Leakage power comparison.



(a) Normalized lifetime reliability of the three hottest functional units (L1 instruction cache (IL1), instruction fetch queue (IFQ), and reservation station (RS))



(b) Normalized lifetime reliability of the hottest functional unit (L1 instruction cache (IL1)) for ten applications

Fig. 10. Normalized FIT values of the liquid cooling scheme for five different lifetime reliability models.

4. Lifetime Reliability

We evaluate the processor lifetime reliability in terms of its FIT values. Fig. 10 shows FIT values with the liquid cooling scheme that are normalized to those with the air cooling scheme. In Fig. 10(a), the normalized FIT values for five different lifetime reliability models in accordance with the three hottest functional units are presented.

In the case of *gcc* application, the liquid cooling scheme improves the lifetime reliability of IL1 by as much as 97.9%, 95.1%, 91.6%, 93.2%, and 80.5%, corresponding to EM, SM, TDDB, TC, and NBTI, respectively. There are also substantial lifetime reliability enhancements with the liquid cooling scheme for IFQ and RS. As shown in Fig. 10(a), the liquid cooling scheme shows substantial lifetime reliability improvements in the EM and SM models, since the lifetime reliability for these two models is mainly dependent on temperature. On the other hand, in the TDDB model, the lifetime reliability enhancement is relatively less since the thermal effect is limited when the voltage is not changed. Regarding the TC model, there is also significant lifetime reliability enhancement, since the reliability is highly dependent on the difference between the temperature of each functional unit and the ambient temperature. The liquid cooling scheme also enhances the lifetime reliability of the NBTI model due to its strong dependence on temperature.

In Fig. 10(b), the normalized lifetime reliability of the hottest functional unit (IL1) for ten applications is depicted. We evaluate the lifetime reliability based on the temperature and utilization of IL1. The liquid cooling scheme improves the lifetime reliability of IL1 by as

much as 99.6%, 98.8%, 97.6%, 98.7%, and 92.1%, on average, corresponding to EM, SM, TDDB, TC, and NBTI model, respectively. In *gcc* application, the temperature difference of IL1 between the conventional air cooling scheme and the liquid cooling scheme is not as much as the other applications. Thus, the lifetime reliability improvements are not as much as the other applications. Additionally, we also investigate the combined lifetime reliability of five different models. In Fig. 10(b), the rightmost bar per application represents the sum of failure rates that shows the combined lifetime reliability. In our evaluation, the sum of failure rates is improved ranging from 58% (*gcc*) to 95% (*mesa*) depending on applications and the average sum of failure rates is improved by as much as 86.9%.

VI. CONCLUSIONS

The 3D integration technique provides significant benefits in terms of area, wire length, and power consumption. However, the higher heat density of 3D integration requires more efficient cooling methods. In this paper, we investigate the architectural effects (temperature, performance, leakage, and reliability) of the direct interlayer liquid cooling method [6] for the 3D processor, where the dielectric coolant flows in-between individual dies. The evaluation results show that the liquid cooling scheme always sustains localized on-chip temperature below the thermal emergency (358K), which completely eliminates the need for DTM operations. Thus, performance degradation caused by DTM operations is eliminated. Temperature reduction also leads to more than 8% leakage reduction of the 3D processor. In addition, the lowered temperature due to the liquid cooling scheme also significantly improves the lifetime reliability.

We expect our architectural study to motivate more researches on 3D integrated processors and efficient cooling.

ACKNOWLEDGMENTS

This work was supported by the Smart IT Convergence System Research Center funded by the Ministry of Education, Science and Technology as Global Frontier Project (SIRC-2011-0031863) and by the Second Brain Korea 21 Project. Sung Woo Chung is the corresponding author of this paper.

REFERENCES

[1] B.–G. Ahn, J. Kim, W. Li, and J.–W. Chong, “Effective Estimation Method of Routing Congestion at Floorplan Stage for 3D ICs”, *Semiconductor Technology and Science, Journal of*, Vol. 11, No. 4, pp. 344-350, Dec., 2011.

[2] S.-Y. Bang, K. Bang, S. Yoon, and E.–Y. Chung "Run-Time Adaptive Workload Estimation for Dynamic Voltage Scaling", *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, Vol. 28, No. 9, pp. 1334-1347, Sep., 2009.

[3] B. Black, M. M. Annavaram, E. Brekelbaum, J. DeVale, L. Jiang, G. H. Loh, D. McCauley, P. Morrow, D. W. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. P. Shen and C. Webb, “Die-Stacking (3D) Microarchitecture,” *Microarchitecture, 2006, MICRO 2006, 39th Annual IEEE/ ACM International Symposium on*, pp. 469-479, Dec., 2006.

[4] D. Brooks and M. Martonosi. “Dynamic Thermal Managementfor High-Performance Microprocessors,” *High Performance Computer Architecture, 2001, HPCA 2001, 7th International Symposium on*, pp. 171-182, Jan., 2001.

[5] D. Brooks, V. Tiwari, and M. Martonoshi, “Wattch: A framework for architectural-level power analysis and optimizations,” *Computer Architecture, 2000, ISCA 2001, International Symposium on*, pp. 83-94, Jun., 2000.

[6] T. Brunschwiler, B. Michel, H. Rothuizen, U. Kloter, B. Wunderle, H. Oppermann and H. Reichl, “Forced Convective Interlayer Cooling in Vertically Integrated Packages,” *Thermal and Thermomechanical Phenomena in Electronic Systems, 2008, ITHERM 2008, 11thIntersociety Conference on*, pp. 1114-1125, May, 2008.

[7] T. Brunschwiler, B. Michel, H. Rothuizen, U. Kloter, B. Wunderle, H. Oppermann and H. Reichl, “Interlayer Cooling Potential in Vertically Integrated Packages,” *Microsystem Technologies*, Vol. 15, No. 1, pp. 57-74, Jan., 2009.

[8] X. Y. Chen, K. C. Toh and J. C. Chai, “Direct

Liquid Cooling of a Stacked Multichip Module," *Electronics Packaging Technology Confonference, 2002, 4th*, pp. 380-384, Dec., 2002.

[9] J. Cong, J. Wei and Y. Zhang, "A Thermal-Driven Floorplanning Algorithm for 3D ICs," *Computer-Aided Design 2004, ICCAD 2004, IEEE/ACM International Conference on*, pp. 306-313, Nov., 2004.

[10] J. Cong and Y. Zhang, "Thermal-Driven Multilevel Routing for 3D ICs," *Design Automation Conference, 2005, ASP-DAC 2005, Asia and South Pacific*, pp. 121-126, Jan., 2005.

[11] A. K. Coskun, T. S. Rosing, D. A. Alonso, J. Leblebici, and J. Ayala, "Dynamic Thermal Management in 3D Multicore Architectures," *Design, Automation & Test in Europe Conference & Exhibition, 2009, DATE 2009*, pp. 1410-1415, Apr., 2009.

[12] A. K. Coskun, J. L. Ayala, D. Atienza, and T. Simunic, "Modeling and Dynamic Management of 3D Multicore Systems with Liquid Cooling," *Very Large Scale Integration (VLSI-SoC), 2009, 17th IFIP International Conference on*, pp. 60-65, Oct., 2009.

[13] S. Das, A. Chandrakasan and R. Reif, "Timing, Energy and Thermal Performance of Three-Dimensional Integrated Circuits," *VLSI, 2004, GLSVLSI 2004, 14th ACMGreat Lakes Symposium on*, pp. 338-343, Apr., 2004.

[14] S. Das, A. Fan and K. –N. Chen, "Technology, Performance and Computer-Aided Design of Three-Dimensional Integrated Circuits," *Physical Design, 2004, International Symposium on*, pp. 108-115, Apr., 2004.

[15] B. Goplen and S. Sapatnekar, "Efficient Thermal Placement of Standard Cells in 3D ICs using a Force Directed Approach,"*Computer-Aided Design 2003, ICCAD 2003, IEEE/ACM International Conference on*, pp. 86-89, Nov., 2003.

[16] W.–L. Hung, G. M. Link, Y. Xie, N. Vijaykrishnan, and M. J. Irwin, "Interconnect and Thermal-aware Floorplanning for 3D Microprocessors," *Quality Electronic Design, 2006, ISQED 2006, 7th International Symposium on*, pp. 98-104, Mar., 2006.

[17] S. Im and K. Banerjee, "Full Chip Thermal Analysis of Planar (2-D) and Vertically Integrated (3-D) High Performance ICs," *Electrion Devices Meeting, 2000, IEDM Technical Digest, International*, pp. 727-730, Dec., 2000.

[18] C. Isci and M. Martonosi, "Runtime Power Monitoring in High-End Processors: Methodology and Empirical data", *Microarchitecture, 2003, MICRO 2003, 36th Annual IEEE/ACM International Symposium on*, pp. 93-104, Dec., 2003.

[19] H. B. Jang, E. –Y. Chung, and S. W. Chung, "Adopting the Banked Register File Scheme for Better Performance and Less Leakage," *Electronics and Telecommunications Research Institute Journal*, Vol. 30, No. 4, pp. 624-626, Aug., 2008.

[20] H. B. Jang, I. Yoon, C. H. Kim, S. Shin, and S. W. Chung, "The Impact of Liquid Cooling on 3D Multi-Core Processors," *Computer Design, 2009, ICCD 2009, IEEE International Conference on*, pp. 472-478, Oct., 2009.

[21] J. Kong, J. John, E.-Y. Chung, S. W. Chung, and J. Hu, "On the Thermal Attack in Instruction Caches", *Dependable and Secure Computing, IEEE Transactions on*, Vol. 7, No. 2, pp. 217-223, Apr.-Jun., 2010.

[22] J. Koo, S. Im, L. Jiang and K. Goodson, "Integrated Microchannel Cooling for Three-Dimensional Electronic Circuit Architectures," *Heat Transfer, Journal of*, Vol. 127, No. 1, pp. 49-58, Jan., 2005.

[23] J. S. Lee, K. Skadron and S. W. Chung, "Predicting Future Temperature for Temperature-Aware DVFS," *Computers, IEEE Transactions on*, Vol. 59, No. 1, pp. 127-133, Jan., 2010.

[24] W. Liao, L. He and K. M. Lepak, "Temperature and Supply Voltage Aware Performance and Power Modeling at Microarchitecture," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, Vol. 24, No. 7, Jul., 2005.

[25] G. H. Loh, "3D-Stacked Memory Architectures for Multi-Core Processors," *Computer Architecture, 2008, ISCA 2008, International Symposium on*, pp. 453-464, Jun., 2008.

[26] G. H. Loh, "A modular 3d processor for flexible product design and technology migration," *Computing Frontiers, 2008, CF 2008, 5th conference on*, pp. 159-170, May, 2008.

[27] G. H. Loh, Y. Xie and B. Black, "Processor Design in 3D Die-Stacking Technologies," *Micro, IEEE*, Vol. 27, No. 3, May-Jun., 2007.

[28] G. L. Loi, B. Agarwal, N. Srivastava, S. –C. Lin and T. Sherwood, "A Thermally-Aware Performance Analysis of Vertically Integrated (3-D) Processor-Memory Hierarchy," *Design Automation Conference, 2006, DAC 2006, 43rd ACM/IEEE*, pp. 991-996, Jul., 2006.

[29] H. Mizunuma, C. Yang, and Y. Lu, "Thermal Modeling for 3D-ICs with Integrated Microchannel Cooling," *Computer-Aided Design, 2009, ICCAD 2009, IEEE/ACM International Conference on*, pp. 256-263, Nov., 2009.

[30] L. D. Paulson, "Onboard Cooler Keeps Chips Comfortable," *Computer*, Vol. 43, No. 2, pp. 15-18, Feb., 2010.

[31] K. Puttaswamy and G. H. Loh, "Dynamic Instruction Schedulers in a 3-Dimensional Integration Technology," *VLSI, 2006, GLSVLSI 2006, 16th ACM Great Lakes Symposium on*, pp. 153-158, May, 2006.

[32] K. Puttaswamy and G. H. Loh, "Implemeting Caches in a 3D Technology for High Performance Processors," *Computer Design, 2005, ICCD 2005, IEEE International Conference on*, pp. 525-532, Oct., 2005.

[33] K. Puttaswamy and G. H. Loh, "The Impact of 3-Dimensional Integration on the Design of Arithmetic Units," *Circuits and Systems, 2006, ISCAS 2006, IEEE International Symposium on*, pp. 4951-4954, May, 2006.

[34] K. Puttaswamy and G. H. Loh, "Thermal Herding: Microarchitecture Techniques for Controlling Hotspots in High-Performance 3D-Integrated Processors," *High Performance Computer Architecture, 2007, HPCA 2007, 13thIEEE International Symposium on*, pp. 193-204, Feb., 2007.

[35] P. Reed, G. Yeung, and B. Black, "Design Aspects of a Microprocessor Data Cache using 3D Die Interconnect Technology," *Integrated Circuit Design and Technology, 2005, ICICDT 2005, International Conference on*, pp. 15-18, May, 2005.

[36] R. Reif, A. Fan, K. –N. Chen and S. Das, "Fabrication Technologies for Three-Dimensional Integrated Circuits," *Quality Electronic Design, 2002, International Symposium on*, pp. 33-37, Mar., 2002.

[37] P. H. Shiu, R. Ravichandran, S. Easwar and S. K. Lim, "Multi-layer Floorplanning for Reliable System-on-Package," *Circuits and Systems, 2004, ISCAS 2004, IEEE International Symposium on*, pp V69-V72, May, 2004.

[38] K. Skadron, K. Sankaranarayanan, S. Veluasmy, D. Tarjan, M. R. Stan, and W. Huang, "Temperature-Aware Microarchitecture: Modeling and Implementation," *Architecture and Code Optimization, ACM Transactions on*, Vol. 1, No. 1, pp. 94-125, Mar., 2004.

[39] J. Srinivasan, S. V. Adve, P. Bose, J. A. Rivers, "Exploiting Structural Duplication for Lifetime Reliability Enhancement," *Computer Architecture, 2005, ISCA 2005, International Symposium on*, pp. 520-531, Jun., 2005.

[40] J. Srinivasan, S. V. Adve, P. Bose, J. A. Rivers, "The Case for Lifetime Reliability-Aware Micro-processors," *Computer Architecture, 2004, ISCA 2004, International Symposium on*, pp. 276-287, Jun., 2004.

[41] C. Sun, L. Shang, and R. P. Dick, "Three-dimensional multiprocessor system-on-chip thermal optimization," *Hardware/software Codesign and System Synthesis, 2007, CODES+ISSS 2007, 5th International Conference on*, pp. 117-122, Sep., 2007.

[42] Y.–F. Tsai, Y. Xie, N. Vijaykrishnan, and M. J. Irwin, "Three-Dimensional Cache Design Exploration Using 3D Cacti," *Computer Design, 2005, ICCD 2005, IEEE International Conference on*, pp. 519-524, Oct., 2005.

[43] F. M. White, Fluid Mechanics, McGraw-Hill, pp. 184, 2002.

[44] E. Wong and S. Lim, "3D Floorplanning with Thermal Vias," *Design, Automation & Test in Europe, 2006, DATE 2006*, pp. 1-6, Mar., 2006.

[45] Y. Xie, G. H. Loh, B. Black and K. Bernstein, "Design Space Exploration for 3D Architecture," *Emerging Technologies in Computing Systems, ACM Journal on*, Vol. 2, No. 2, pp. 65-103, Apr., 2006.

[46] Perfmon2 patch. Available: http://perfmon2.source-forge. net. 2009.

[47] Intel Core 2 Duo Technical Documents. Available in http://www.intel.com/design/core2duo/documen-tation. html

[48] Linux Kernel CPUfreq subsystem. Available in http://www.kernel.org/pub/linux/utils/kernel/cpufre

q/cpufreq.html.

[49] Notebook Hardware Control, personal edition. Available in http://www.pbus-167.com/.

[50] SPEC, Standard Performance Evaluation Corporation. Available: http://www.spec.org/cpu2000/. 2009.

[51] User's guide, Icepak, 4.4.6. ANSYS/Fluent Inc., Lebanon, NH, 2007.

[52] UC Berkeley Device Group, "Berkeley Predictive Technology Model (BPTM)," University California, Berkeley, CA, Jul., 2002.

Hyung Beom Jang received the B.S. and M.S. degrees in Computer Science and Engineering from Korea University, Seoul, Korea, in 2007 and 2009, respectively. He is currently working toward the Ph.D. degree in the Department of Computer and Radio Communication Engineering at Korea University. His research interests include low-power design, temperature-aware design, and 3D integration design.

Ikroh Yoon received the B.S. and M.S. degrees in the Department of Mechanical Engineering from Hongik University, Seoul, Korea, in 2009 and 2011, respectively. He is currently working at the Korea Institute of Marine Science and Technology Promotion. His research interests include computational analysis of fluid and heat transfer.

Cheol Hong Kimreceived the B.S., M.S., and Ph.D. degreesin Computer Engineering from Seoul National University, Seoul, Korea,in 1998, 2000, and 2006, respectively. He worked as a senior engineer in Samsung Electronics, Korea from 2005 to 2007. In 2007, he joined the faculty of the School of Electronics and Computer Engineering, Chonnam National University, Gwangju, Korea. His research interests include high-performance multicore architecture, power-aware processor architecture, embedded systems design, and GPU architecture design.

Seungwon Shin received the B.S. and M.S. degrees in Mechanical Engineering from Seoul National University, Seoul, Korea, in 1995 and 1998, respectively. He then received his Ph.D. from Georgia Tech, in 2002. He is currently a professor in the School of Mechanical and System Design Engineering at Hongik University, Seoul, Korea. His research interests include computational fluid dynamics, multiphase flow, surface tension effect, phase change process.

Sung Woo Chung received the B.S., M.S., and Ph.D. degrees in Electrical Engineering and Computer Science from Seoul National University, Seoul, Korea, in 1996, 1998, and 2003, respectively. He is currently an associate professor in the Department of Computer and Radio Communication Engineering at Korea University. His research interests include low-power design, temperature-aware design, and user-aware design. He is a member of the IEEE and the IEEE Computer Society. He serves (and served) on the technical program committees in many conferences, including International Conference on Computer Design, International Symposium on Quality Electronic Design, and Asia and South Pacific Design Automation Conference. He is currently an Associate Editor of IEEE Transactions on Computers.