The Impact of Liquid Cooling on 3D Multi-Core Processors

Hyung Beom Jang[1], Ikroh Yoon[2], Cheol Hong Kim[3], Seungwon Shin[4], and Sung Woo Chung[1]
[1]*Division of Computer and Communication Engineering, Korea University, Seoul, Korea*
[2 4]*Department of Mechanical Engineering, Hongik University, Seoul, Korea*
[3]*School of Electronics and Computer Engineering, Chonnam National University, Gwangu, Korea*
[1]*{kuphy01,swchung}@korea.ac.kr,* [2]*yoonikroh@gmail.com,* [3]*chkim22@chonnam.ac.kr*
[4]*sshin@wow.hongik.ac.kr*

Abstract— **Recently, 3D integration has been regarded as one of the most promising techniques due to its abilities of reducing global wire lengths and lowering power consumption. However, 3D integrated processors inevitably cause higher power density and lower thermal conductivity, since the closer proximity of heat generating dies makes existing thermal hotspots more severe. Without an efficient cooling method inside the package, 3D integrated processors should suffer severe performance degradation by dynamic thermal management as well as reliability problems. In this paper, we analyze the impact of the liquid cooling on a 3D multi-core processor compared to the conventional air cooling. We also evaluate the leakage power consumption and the lifetime reliability depending on the temperature of each functional unit in the 3D multi-core processor. The simulation results show that the liquid cooling reduces the temperature of the L1 instruction cache (the hottest block in this evaluation) by as much as 45 degrees, resulting in 12.8% leakage reduction, on average, compared to the conventional air cooling. Moreover, the reduced temperature of the L1 instruction cache also improves the reliability of electromigration, stress migration, time-dependent dielectric breakdown, thermal cycling, and negative bias temperature instability significantly.**

I. INTRODUCTION

As technology scales down and integration densities continue to increase, interconnect delay has become the dominant factor for microprocessor performance. To reduce the interconnect delay, three-dimensional (3D) integration technology has drawn quite attentions, since the vertically integrated 3D processor enables faster on-chip communication and lower power consumption by reducing global wire lengths drastically [21][22][23][25][32]. Increasing the number of stacked dies makes the 3D integration technology more advantageous due to the reduced wire lengths.

Currently, even in 2D processors, thermal problem is serious [13][14], which causes unexpected functional error or permanent damage. Thus, most high-performance 2D processors have thermal management schemes. The 3D processors have more severe thermal problems than the planar (2D) processors, since they have higher power density and lower heat dissipation capability. As more dies are stacked, temperature of the dies increases more significantly due to the higher heat density of the power dissipating components. Longer heat dissipation path from the die to the heat sink is another factor worsening thermal dissipation. For these reasons, 3D integration technology deteriorates existing thermal hotspots and incurs new thermal hotspots [32]. Therefore, without an efficient cooling method inside the package, the 3D integrated processor should suffer severe performance degradation by Dynamic Thermal Management (DTM) [27].

According to several previous works on 3D integration technologies [7][11], thermal-aware techniques can be considered to sustain the temperature in the 3D integrated processor below thermal emergency. Many studies have focused on thermal optimization of the 3D integrated processor. Cong et al. proposed the thermal-driven floorplanning method [5] and the wire routing technique [6] to improve the heat dissipation capability of the 3D integrated processor. Shiu et al. also proposed the thermal-driven floorplanning method to mitigate the thermal problems for the 3D integrated processor [10]. Furthermore, thermal-aware placement methods have been proposed for lowering the temperature in the 3D integrated processor [8][9]. Unfortunately, the above-mentioned approaches are insufficient to solve the thermal problem in the 3D integrated processors. As an alternative, inserting thermal Through Silicon Via (TSV) into the vertically integrated processor is regarded as an efficient method to reduce on-chip temperature [31]. In this method, more thermal TSVs are required for the hottest region. However, thermal TSVs cannot be inserted directly into hot regions of the 3D integrated processor, since most of the hot areas are occupied by macro blocks or functional units. As a result, thermal dissipation through thermal TSVs is not enough to reduce the peak temperature of the 3D integrated processor below thermal emergency.

For more aggressive cooling for the 3D integrated processors, interlayer liquid cooling methods with water as a coolant also have been investigated by several researchers [3][4][15]. Chen et al. demonstrated that interlayer liquid cooling and a die spacing for the heat transfer configuration reduced the heat densities of 25W/cm^2 ~ 50 W/cm^2 [4]. Koo et al. also showed that a layer of integrated microchannel cooling with the water as a coolant reduced the heat densities up to 135W/cm^2 within a 3D integrated processor [15]. Brunschwiler et al. examined the heat dissipation capability of the direct liquid cooling at practical operation conditions [3]. Additionally, scalable interlayer cooling concept using the correlation-based prediction method was proposed [3]. In this method, the numerical modeling method was used to identify the optimal liquid cooling structure.

978-1-4244-5028-2/09/$25.00 ©2009 IEEE

As far as we know, there has not been any study about the architectural effects of the liquid cooling scheme on the microprocessors. In this paper, we analyze the impact of the liquid cooling on temperature, leakage, and reliability in a 3D multi-core processor, based on the most recent liquid cooling scheme by IBM corporation [3].

The rest of this paper is organized as follows. Section 2 explains related work on 3D integration and liquid cooling. Section 3 describes the 3D integrated processor incorporated with the liquid cooling scheme. Section 4 provides the details about our evaluation environments and modeling. Section 5 analyzes our evaluation results on temperature, leakage, and reliability. Section 6 concludes this paper with future work.

II. Related Work

In this section, we present related work on the 3D integration techniques and the liquid cooling techniques.

A. 3D Integration Techniques

1) Structure of 3D Integration: Research for 3D integration techniques can be categorized into two major groups; (1) those dealing with die-bonding techniques and (2) those dealing with Multi-layer Buried Structures (MLBS) [30].

The die-bonding 3D integration techniques use the conventional 2D manufacturing processes and insert metal vias to bond the two planar dies [26]. Various bonding materials have been used for bonding the dies. The die-bonding 3D integration techniques can have three different topologies for interfacing between multiple planar dies; face-to-face (F2F), face-to-back (F2B), and back-to-back (B2B). In MLBS, it is possible to stack many heterogeneous dies [17][22]. By using the MLBS, the 3D integration techniques enable mixing dissimilar process technologies such as high-speed CMOS with high-density DRAM. In this paper, we just consider a face-to-face bonding technique because it provides a very 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 used for the electrical interconnection between multiple layers passing vertically through a silicon die. Before using TSVs, multiple layers have been connected by wiring together at their edges. However, TSVs replace the edge-wiring interconnection by the vertical interconnection, leading to significant reduction of total wire length. It has been reported that the vertical latency for traversing the height of a 20-layer stack is just only 12 ps [20]. Therefore, interconnection using TSVs through the body of the die can enable fast on-chip communication for the 3D integrated processors.

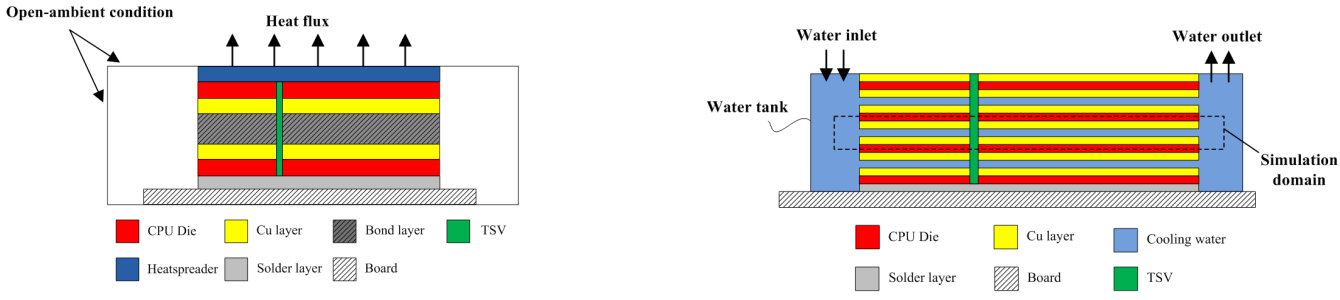
TSVs are also different from the typical metal wires primarily in their size. Even though TSVs are wider than the typical metal wires, TSVs have very short height, since each wafer is thinned to only tens of microns. After fabrication, each wafer is thinned to only 10-100μm in thickness [1][8], and then TSVs which have pitches of only 4-10 μm [17] are etched through the bulk silicon. Lastly, thermo-compression is used to bond the individual layers together to form the 3D integration structure [19].

B. Liquid Cooling Techniques

Liquid cooling using water as a coolant is one of the most efficient cooling methods for the 3D integrated processors due to its superior heat dissipation efficiency of water. The liquid cooling techniques can be categorized into two major types; one is the indirect cooling and the other is the direct cooling (also called immersion).

In the indirect cooling techniques, the coolant does not contact with the electronic components directly. Instead, microchannels are commonly used. Microchannels can be integrated into a substrate or a heat sink and then inserted between each layer [15]. Different from the indirect cooling techniques, the dielectric coolant is allowed to go through between the each layer for the direct cooling techniques. Therefore, the coolant can absorb the heat flux from the each layer of the 3D integrated processor, as shown in Fig. 1(b).

On the contrary, in the conventional air cooling scheme as depicted in Fig. 1(a), the heat spreader is used for dissipating heat flux instead of the coolant, which is widely used for recent high performance processors. In this paper, we com-



(a) Air cooling

(b) Direct liquid cooling using water as a coolant

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

pare architectural effects of the direct interlayer liquid cooling scheme to those of the conventional air cooling scheme in the 3D integrated processor.

III. 3D Integrated Processor with The Liquid Cooling Scheme

Fig. 2(a) shows the conventional planar processor with an L2 cache and Fig. 2(b) illustrates the corresponding 3D integrated processor. For the planar processor, we model a processor based on the Intel Core 2 microarchitecture according to the [24]. For the 3D integrated processor, we partition and stack one core and half of the L2 cache on each of the die [22][30]. Through this model, we can reduce the total area by 50% approximately and reduce the wire delay. The thickness of each layer in our 3D integrated processor is modeled based on [1].

We adopt the liquid cooling structures proposed by IBM corporation [3]. As explained earlier, the direct liquid cooling technique is depicted in Fig. 1(b), where four dies are stacked symmetrically with TSVs and the coolant is pumped in-between the individual dies. However, we considered the two-die stacked 3D processor in our evaluation. The heat dissipated from each die is transferred to the coolant through the die silicon slab or the wiring layers. A hermetic sealing of the electrical TSVs is necessary to use water as a coolant [3].

IV. Evaluation Methodology

A. Evaluation Environments

To evaluate power consumption of each functional unit, we used Wattch [2] and perfmon2 [33]. We first modified Wattch based on the empirical power model proposed by Isci et al. [12] to read the performance counts for obtaining the activity factor of each functional unit. However, this modified Wattch was originally implemented to operate with the Pentium 4. Therefore, we also modified that Wattch to configure and to read performance counters of the Core 2 processor by using the perfmeon2. We obtained power values needed for the each functional unit from [18].

The Intel Core 2 processor provides three fixed-function performance counters and two general-purpose performance counters for counting events. Each counter is associated with the configuration register, which indicates the corresponding performance counter. 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. Fig. 3 shows the floorplan of one die for the 3D integrated processor.

We also assume that the coolant goes through in-between the individual dies with the volumetric flow rate (Q) of $6.55\times10^{-3}m^3$/hr. The power required for the flowing coolant in-between individual dies is calculated as follows:

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

where $\Delta Pressure$ is the pressure drop calculated from numerical simulation. In our evaluation, power consumption for the forced flowing coolant is only about 0.823W due to the low volumetric flow rate, which is enough to absorb the heat flux of the two-die 3D integrated processor. Compared to the 1~2W power consumption of the conventional air cooling scheme, the forced convective driving power used in our evaluation is reasonable.

From 26 SPEC2000 benchmarks [34], we select two applications (gcc, gzip) that show the most significant temporal variations in the perspective of thermal behavior [27], since long simulation time is required to evaluate all benchmarks. Note all the benchmarks do not cause enough heat to be interesting thermally.

B. Thermal Modeling

To investigate the heat dissipation capability of the 3D integrated processor incorporated with the liquid cooling scheme, we used Fluent Package (ICEPAK) simulation engine [35] that is used to analyze heat flow. Each of the die composed of one core and half of the L2 cache is mounted at the board by Flip-chip Ball Grid Array (FBGA) type and the components of the package (498-solderball, Flip-chip underfill, and Flip-chip substrate) are modeled as a single combined layer which has equivalent thermal resistance. The other thermal parameters for the 3D integrated processor are obtained from [1]. 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 model the liquid cooling scheme more accurately considering the flow of coolant, we need to increase the number of grid elements. Since we wanted to

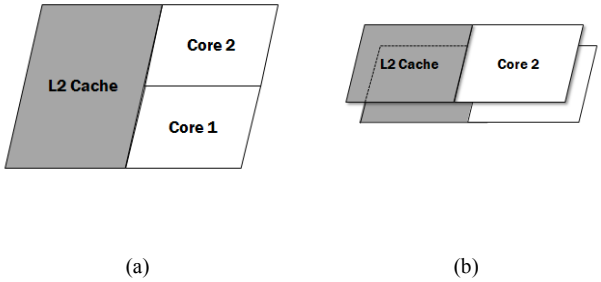


Fig. 2. (a) Planar layout of cores and an L2 cache, (b) 3D implementation of cores and an L2 cache.

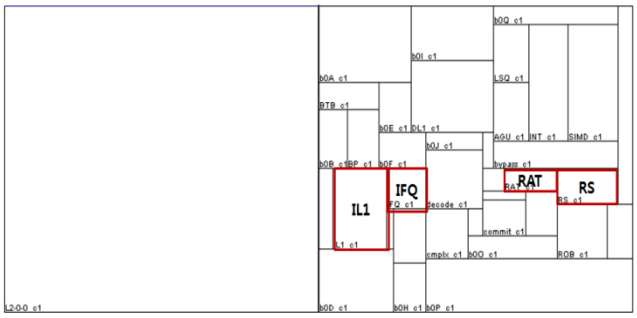


Fig. 3. Floorplan of a die in the 3D integrated processor.

reduce simulation time without hurting accuracy, we did not increase the number of grid elements for air cooling modeling.

In the air cooling scheme, the coolant material is the ambient air of 293.15K temperature passing through heat spreader and finned heat sink. The constant convective heat transfer coefficient of 15000W/m^{2}K, used in this paper, is sufficiently high enough to represent a well-designed air cooling scheme at the upper side of the heat spreader.

In the evaluated liquid cooling scheme, we assume that the coolant (293.15K water) flows in a one-dimension through in-between the individual dies with a constant mass flow rate. Additionally, the heat flux of all functional units on each die is considered individually, since the heat flux of each functional unit is different depending on its corresponding power consumption.

C. Temperature-Dependent Leakage Modeling

We also examine the leakage gain from the liquid cooling compared to the air cooling. Leakage power mainly consists of subthreshold leakage power and gate leakage power [16]. On estimating the subthreshold leakage power, we consider two types of leakage; one is leakage in the logic circuits such as functional units and the other is leakage in SRAM-based units such as caches and register files.

The leakage power of the logic circuits is calculated as the product of the number of gates and the average subthreshold leakage current per gate. On the other hand, the leakage power of the SRAM-based units is the sum of SRAM memory cells' leakage power and their peripheral circuits' leakage power. Moreover, the gate leakage power is calculated by gate direct tunneling current – including tunneling current between gate and substrate and current between gates and channels.

In this paper, we take into account both the subthreshold leakage power for logic circuits and the gate leakage power for SRAM-based units, reflecting thermal effects on leakage as well. We choose the power values of 65nm technology corresponding to the Intel Core 2 processor from which we have traces [36].

D. Reliability Modeling

For evaluating lifetime reliability of our processor model, we used application-aware architecture-level methodology called as Reliability Aware Microprocessor (RAMP) [29]. RAMP can dynamically track the lifetime reliability depending on the application behavior. This methodology represents the processor lifetime reliability in terms of Mean Time to Failure (MTTF) or the expected lifetime of the processor and calculates an instantaneous MTTF based on the current temperature and the utilization of each functional unit.

The standard method for representing constant failure rates is Failures in Time (FIT), which means the number of failures per 10^9 device operating hours. The MTTF models used in RAMP provide the estimated FIT values based on fixed operating parameters, such as temperature, voltage, and frequency. In this paper, we evaluate the lifetime reliability of the 3D integrated processor with liquid cooling compared to that with air cooling in the perspective of FIT.

There are four models for evaluating reliability of the processor in RAMP [29]. First model is electromigration (EM) which occurs in aluminum and copper interconnects due to the mass transport of conductor metal atoms in interconnects. The electromigration is exponentially dependent on temperature. The effects of electromigration modeled in RAMP are as follows [29]:

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

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

Second model is stress migration (SM) that occurs when the metal atoms in interconnects migrate. It is caused by different thermal expansion rates of different materials in the device. This mechanical stress is proportional to the temperature changes. The stress migration models used in RAMP are as follows [29]:

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

where T_0 and T are absolute temperature in Kelvin, N and E_{aSM} are constant values depending on materials.

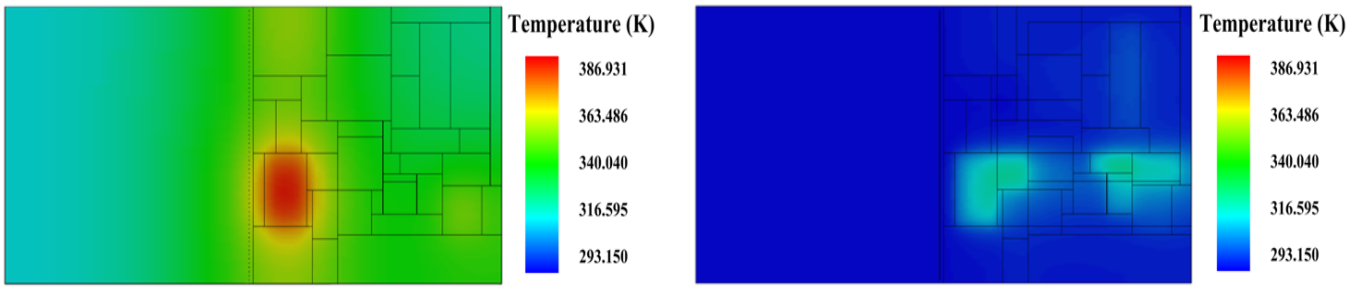
Third one is Time-Dependent Dielectric Breakdown (TDDB), or gate oxide breakdown. The gate dielectric wears down with time and fails when a conductive path is formed in the dielectric. The advent of thin and ultra-thin gate oxides coupled with aggressive scaling of supply voltage accelerates TDDB failure rates. The TDDB model highly relies on voltage and temperature. The mean time to failure from the time-dependent dielectric breakdown can be given as follows [29]:

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

where a, b, X, Y, and Z are constant parameters based on data in [29].

Last model in RAMP is thermal cycling (TC). All parts of the device experience with fatigue damages when there is 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 mean time to failure due to the thermal cycling is based on [29] and is given by:

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



(a) air cooling (b) liquid cooling

Fig. 4. Thermal profile of the 3D integrated processor (air cooling vs. liquid cooling)

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 Negative Bias Temperature Instability (NBTI) for the emerging critical failure model. The NBTI takes place when the gate is biased negative with respect to the source and drain. The equation used to determine the mean time to failure from the negative bias temperature instability is [28]:

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

where A, B, C, D, and β are constant values based on [28], and k is the Boltzmann's constant value.

We evaluate the lifetime reliability of the 3D integrated processor corresponding to the cooling scheme by using the above five models.

V. Evaluation Results and Analysis

In this section, we present our evaluation results to analyze the impact of liquid cooling scheme on the 3D integrated processor in the perspective of temperature, leakage, and lifetime reliability.

A. Temperature

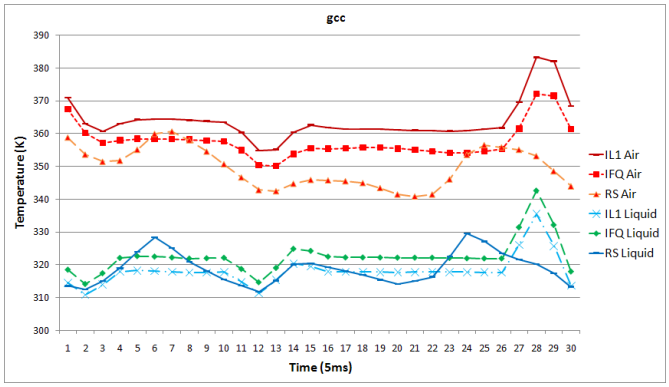
Fig. 4(a) and Fig. 4(b) show steady state thermal profiles of the 3D integrated processor with the conventional air cooling scheme and the liquid cooling scheme, respectively, in case of the gcc application. As shown in the figures, the L1 instruction cache (IL1) is turned out to be a hotspot with the air cooling scheme. With the liquid cooling scheme, however, the temperature of the IL1 is reduced drastically. The coolant flows from left (IL1) to right (rename table (RAT), instruction fetch queue (IFQ), and reservation station (RS)) in Fig. 3 absorbing the heat flux of the IL1. Thus, RAT, IFQ, and RS are cooled down with relatively warm coolant that already went through the IL1, resulting in less thermal reduction of RAT, IFQ, and RS. If they were placed at upper side or lower side of the L1 instruction cache in Fig. 3, the temperature of them would be more reduced.

Among the functional units depicted in Fig. 3, we concentrate on the three hottest functional units for our temporal thermal evaluation. Fig. 5 shows the temperatures of these functional units depending on the cooling schemes. As shown in Fig. 5(a), when we adopted the liquid cooling scheme with the gcc application, temperature of IL1, IFQ, and RS is reduced by as much as 45 degrees, 35 degrees, and 31 degrees, respectively, on average. In case of the gzip application (Fig. 5(b)), as much as 42 degrees, 31 degrees and 26 degrees are reduced by the liquid cooling scheme, respectively, on average. In addition, we notice that the volumetric flow rate used in our liquid cooling scheme, which is $6.55\times10^{-3}m^3$/hr, is enough to reduce the temperature of the hottest functional units below the thermal emergency.

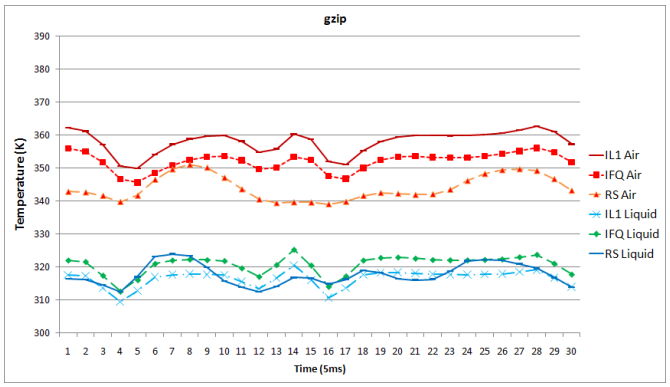
Unfortunately, we could not investigate the performance impact of the air cooling scheme and the liquid cooling scheme, since it takes 10~20 hours to evaluate performance of each application with Fluent Package (ICEPAK) engine while considering the DTM triggers. However, thermal simulation results show that the 3D integrated processor incorporated with the liquid cooling scheme does not invoke any DTM trigger, which will surely improve performance significantly.

B. Leakage

We also investigate the leakage consumption of three hottest functional units (IL1, IFQ, and RS). In the gcc application, shown in Fig. 6, the liquid cooling scheme reduces the leakage consumption of IL1, IFQ, and RS by as much as 12.8%, 12.7%, and 11.1%, respectively, on average. In case of the gzip application, the liquid cooling scheme shows leakage reduction as much as 11.5% (IL1), 11.3% (IFQ), and 9.4% (RS), respectively, on average. Since leakage power consumption is dependent on temperature, the decreased temperature by the liquid cooling scheme leads to the leakage reduction of these hottest functional units. Total processor leakage is also reduced by 11.5% and 10.1%, in gcc and gzip, respectively.



(a) gcc



(b) gzip

Fig. 5. Temperature comparison of three hottest functional units between the conventional air cooling scheme and the liquid cooling scheme.

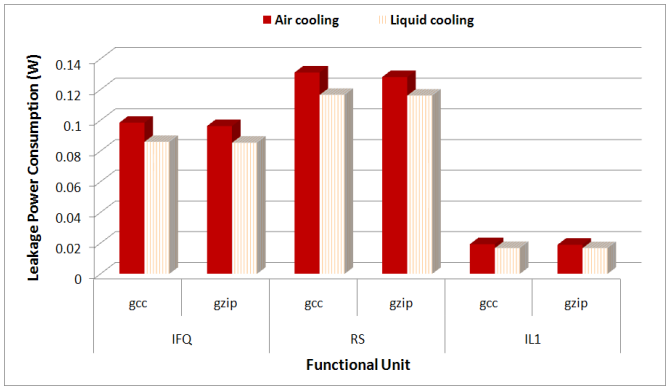


Fig. 6. Leakage consumption of the three hottest functional units.

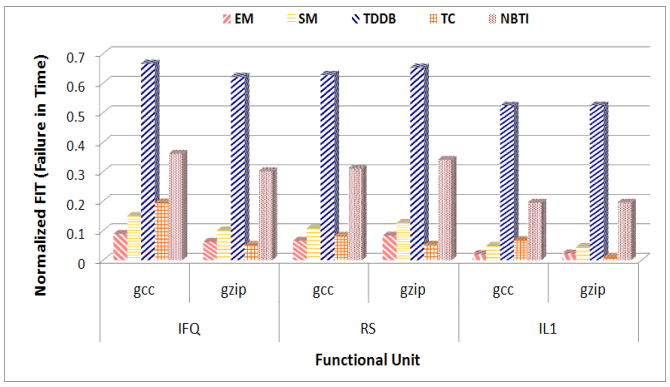


Fig. 7 Normalized FIT values for five different lifetime reliability models.

C. *Lifetime Reliability*

We evaluate the processor lifetime reliability in terms of its FIT values. Fig. 7 shows the FIT values with the liquid cooling scheme that is normalized to those with the air cooling scheme. In Fig. 7, the normalized FIT values for five different lifetime reliability models are presented. In case of the gcc application, the liquid cooling scheme improves the lifetime reliability of the L1 instruction cache by as much as 97.9%, 95.1%, 47.6%, 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 the instruction fetch queue and the reservation station.

As shown in Fig. 7, the liquid cooling scheme shows substantial lifetime reliability improvements in the EM and SM model, 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 not prominent compared to the other models, since thermal effect is limited when voltage is not changed [29]. In the TC model, there is the largest enhancement of the lifetime reliability, since reliability is highly dependent on the difference between the temperature of each functional unit and the ambient temperature of the 3D integrated processor. Additionally, the lifetime reliability with the liquid cooling scheme is also greatly improved in the NBTI model due to its strong dependence on temperature.

VI. Conclusion and Future Work

The 3D integration technique provides significant benefits in terms of area, wire length, and power consumption. However, higher heat density of the 3D integration requires more efficient cooling methods. In this paper, we evaluate the architectural effects (temperature, leakage, and reliability) of the direct interlayer cooling method [3] for the 3D integrated processor, where the dielectric coolant flows in-between individual dies. The evaluation results show that this liquid cooling scheme significantly reduces on-chip temperature under 350K, which completely eliminates thermal emergency. The temperature reduction also leads to more than 10% leakage reduction of the 3D integrated processor. In addition, the lowered temperature also improves the lifetime reliability significantly. From our evaluation results, we found that the liquid cooling scheme is very efficient on the 3D integrated

processor in the perspective of temperature, leakage power consumption, and lifetime reliability.

In practice, thermal hotspots of the 3D integrated processor are changed depending on floorplan and die-stacking method. Thus, we will evaluate the efficiency of the liquid cooling scheme considering the various 3D integration organizations. Additionally, the impact of DTM triggers on performance will be evaluated to see how much benefit we get in performance by using the liquid cooling scheme.

ACKNOWLEDGMENT

This work was supported by the Second Brain Korea 21 Project and Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. R01-2007-000-20750-0).

REFERENCES

[1] 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," in *Proc. of the 39th Int. Symp. on Microarchitecture*, pp. 469-479, Dec. 2006.

[2] D. Brooks, V. Tiwari, and M. Martonoshi, "Wattch: A framework for architectural-level power analysis and optimizations", in *Proc. of the 27th Int. Symp. on Computer Architecture*, pp.83-94, Jun. 2000.

[3] T. Brunschwiler, B. Michel, H. Rothuizen, U. Kloter, B. Wunderle, H. Oppermann and H. Reichl, "Forced Convective Interlayer Cooling in Vertically Integrated Packages," in *Proc. of the 11th Intersociety Conf. on Thermal and Thermomechanical Phenomena in Electronic Systems*, pp. 1114-1125, May 2008.

[4] X. Y. Chen, K. C. Toh and J. C. Chai, "Direct Liquid Cooling of a Stacked Multichip Module," in *Proc. of the 4th Electronics Packaging Technology Conf*., pp. 380-384, Dec. 2002.

[5] J. Cong, J. Wei and Y. Zhang, "A Thermal-Driven Floorplanning Algorithm for 3D ICs," in Proc *of Int. Conf. on Computer Aided Design*, pp. 306-313, Nov. 2004.

[6] J. Cong and Y. Zhang, "Thermal-Driven Multilevel Routing for 3D ICs," in *Proc. of the Asia and South Pacific – Design Automation Conf*., pp. 121-126, Jan. 2005.

[7] S. Das, A. Chandrakasan and R. Reif, "Timing, Energy and Thermal Performance of Three-Dimensional Integrated Circuits," in *Proc. of the 14th ACM Great Lakes Symp. on VLSI*, pp.338-343, Apr. 2004.

[8] S. Das, A. Fan and K. –N. Chen, "Technology, Performance and Computer Aided Design of Three Dimensional Integrated Circuits," in *Proc. of the Int. Symp. on Physical Design*, pp. 108-115, Apr. 2004.

[9] B. Goplen and S. Sapatnekar, "Efficient Thermal Placement of Standard Cells in 3D ICs using a Force Directed Approach", in *Proc. of Int. Conf. on Computer Aided Design*, pp. 86-89, Nov. 2003.

[10] P. H. Shiu, R. Ravichandran, S. Easwar and S. K. Lim, "Multi-layer Floorplanning for Reliable System-on-Package," in *Proc. of Int. Symp. on Circuits and Systems*, Vol. 5, pp V69-V72, May 2004.

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

[12] C. Isci and M. Martonosi, "Runtime Power Monitoring in High-End Processors: Methodology and Empirical data," in *Proc. of Int. Symp. on Microarchitecture,* pp. 93-104, Dec. 2003.

[13] H. B. Jang, E. –Y. Chung, and S. W. Chung, "Adopting the Banked Register File Scheme for Better Performance and Less Leakage," *ETRI Journal*, Vol. 30(4), pp.624-626, Aug. 2008.

[14] J. Kong, J. John, E.-Y. Chung, S. W. Chung, and J. Hu, "On the Thermal Attack in Instruction Caches", *IEEE Trans. on Dependable and Secure Computing*, accepted.

[15] J. Koo, S. Im, L. Jiang and K. Goodson, "Integrated Microchannel Cooling for Three-Dimensional Electronic Circuit Architectures," in *Journal of Heat Transfer*, Vol. 127, pp. 49-58, Jan. 2005.

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

[17] G. H. Loh, "3D-Stacked Memory Architectures for Multi-Core Processors," in *Proc. of the 35th Int. Symp. on Computer Architecture*, pp. 453-464, Jun. 2008.

[18] G. H. Loh, "A modular 3d processor for flexible product design and technology migration," in *Proc. of the 2008 Conf. on Computing Frontiers*, pp. 159-170, May 2008.

[19] G. H. Loh, Y. Xie and B. Black, "Processor Design in 3D Die-Stacking Technologies," *IEEE Micro Magazine*, 27(3), May-Jun. 2007.

[20] 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," in *Proc. of the 43rd ACM/IEEE Design Automation Conf*., pp. 991-996, Jul. 2006.

[21] K. Puttaswamy and G. H. Loh, "Dynamic Instruction Schedulers in a 3-Dimensional Integration Technology," in *Proc. of the ACM Great Lakes Symp. On VLSI*, pp. 153-158, May 2006.

[22] K. Puttaswamy and G. H. Loh, " Implemeting Caches in a 3D Technology for High Performance Processors," in *Proc. of the Int. Conf. on Comp. Design*, pp. 525-532, October 2005.

[23] K. Puttaswamy and G. H. Loh, "The Impact of 3-Dimensional Integration on the Design of Arithmetic Units," in *Proc. of the Int. Symp. on Circuits and Systems*, pp. 4951-4954. ,May 2006.

[24] K. Puttaswamy and G. H. Loh, "Thermal Herding: Microarchitecture Techniques for Controlling Hotspots in High-Performance 3D-Integrated Processors," in *Proc. of the 13th Int. Symp. on High Performance Computer Architecture*," pp. 193-204, Feb. 2007.

[25] P. Reed, G. Yeung, and B. Black, "Design Aspects of a Microprocessor Data Cache using 3D Die Interconnect Technology," in *Proc. of the Int. Conf. on Integrated Circuit Design and Tech*., pp. 15-18, May 2005.

[26] R. Reif, A. Fan, K. –N. Chen and S. Das, "Fabrication Technologies for Three-Dimensional Integrated Circuits," in *Proc. of the 3rd Int. Symp. on Quality Electronic Design*, pp. 33-37, Mar. 2002.

[27] K. Skadron, K. Sankaranarayanan, S. Veluasmy, D. Tarjan, M.R. Stan, and W. Huang. "Temperature-Aware Microarchitecture: Modeling and Implementation." *ACM Transactions on Architecture and Code Optimization*, Vol. 1(1), pp. 94-125, Mar. 2004.

[28] J. Srinivasan, S. V. Adve, P. Bose, J. A. Rivers, "Exploiting Structural Duplication for Lifetime Reliability Enhancement," in *Proc. of the 32nd Annu. Int. Symp. on Comp. Architecture*, pp. 520-531, Jun. 2005.

[29] J. Srinivasan, S. V. Adve, P. Bose, J. A. Rivers, "The Case for Lifetime Reliability-Aware Microprocessors," in *Proc. of the 31st Annu. Int. Symp. on Computer Architecture*, pp. 276-287, Jun. 2004.

[30] Y. –F. Tsai, Y. Xie, N. Vijaykrishnan, and M. J. Irwin, "Three-Dimensional Cache Design Exploration Using 3D Cacti," in *Proc. of IEEE Int. Conf. on Computer Design*, pp. 519-524, Oct. 2005.

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

[32] Y. Xie, G. H. Loh, B. Black and K. Bernstein, "Design Space Exploration for 3D Architecture," in *ACM Journal on Emerging Tech. in Computing Systems*, Vol. 2, pp. 65-103, Jul. 2006.

[33] Perfmon2 patch. Available in http://perfmon2.sourceforge.net

[34] SPEC, Standard Performance Evaluation Corporation, available in http://www.spec.org/.

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

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