JOURNAL OF SEMICONDUCTOR TECHNOLOGY AND SCIENCE, VOL. 25, NO. 2, APRIL, 2025
https://doi.org/10.5573/JSTS.2025.25.2.142
ISSN(Print) 1598-1657
ISSN(Online) 2233-4866

Comparative Study of Thermal Dissipation in Increasing DRAM Layers of HBM Using 3D FEA Simulations

Jeong Hun Song and Sang Won Yoon

***Abstract*:** Large language models (LLM) and generative artificial intelligence (AI) require extensive data processing and fast data transfer between components, increasing interest in high bandwidth memory (HBM). The high-speed data processing capability of HBM drives the need for next-generation HBM with additional dynamic random access memory (DRAM) layers. However, this increased stacking leads to more severe thermal issues, along with higher power consumption, potentially limiting HBM performance. This study explores these thermal challenges through 3D finite element analysis (FEA) simulations of simplified HBM models incorporating non-conductive film (NCF) layers. Three models with 4, 8, and 12 DRAM layers were simulated and compared. The results show that the maximum simulated temperature reaches 80°C, close to the maximum allowable DRAM temperature, and approaches 110°C, exceeding the recommended operational temperature for HBM. Therefore, this study highlights the potential thermal limitations of highly stacked HBM configurations.

***Index terms*:** Heat dissipation, HBM, NCF, simplified model, TSV

I. INTRODUCTION

HBM is a 3D stacked memory device implemented by stacking multiple layers of DRAM using a technology called through-silicon via (TSV). Unlike traditional bonding methods, such as wire bonding, TSV creates shorter connections between chips by drilling tiny holes in the chip and connecting the top and bottom layers. As a result, HBM can utilize power efficiently and achieve higher performance while being more compact than traditional memory. Furthermore, it offers wider bandwidth and faster speeds due to its multi-layer approach and numerous IOs [1].

Due to these characteristics, HBM is primarily used in high-end devices, such as high-performance GPUs. With the expansion of the AI market and the growth of data centers, the demand for high-performance graphics cards for High Performance Computing (HPC) has increased, leading to a significant rise in the demand for HBM, making it one of the main memory products [2].

The fundamental structure of HBM is composed of a base die, known as the logic or buffer die, at the bottom, with multiple core dies stacked on top. The dies are bonded through insulating materials, such as NCF or molded underfill (MUF) [3], along with numerous micro bumps. The dies are electrically connected via TSVs and micro bumps. Micro bumps that are not connected to TSVs are called dummy bumps, which help improve heat dissipation. All of these components are assembled at a microscale level to form a single product.

However, due to its multi-layered structure with small, densely packed contact points, there is a challenging heat dissipation issue [4,5]. For example, the gap between the upper and lower chips is too small to dissipate heat effectively, and the thermal conductivity of the adhesive layer is low. Moreover, as the demand for higher performance continues to grow, HBM technology has advanced toward increasing the number of stacks. While the first HBM product featured 4 layers, subsequent generations, such as HBM2, HBM2E, and HBM3, have introduced products

Manuscript received Sep. 13, 2024; revised Dec. 3, 2024; accepted Dec. 22, 2024
Department of Electrical and Computer Engineering, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea
E-mail : swyoon@snu.ac.kr

with up to 12 layers. Currently, research and development are progressing further, with the next generation of products expected to feature 12 layers or even 16 layers and beyond. As a result, it is anticipated that thermal issues will become more significant.

Therefore, predicting the thermal issues of HBMs through simulation is essential for the future development of high-monolayer HBMs. HBM is categorized as HBM 4Hi, 8Hi, or 12Hi, depending on the number of stacked DRAM dies. We conducted thermal simulations using 3D models of HBMs with different numbers of layers to understand how heat generation and distribution occur in 4Hi, 8Hi, and 12Hi HBM models. In the simulations performed in this study, a TSV-stacked structure model was constructed using non-conductive NCF material for bonding. However, NCF is a complex structure that bonds the top and bottom chips and wraps around small protrusions, making a detailed 3D simulation time-consuming and resource-intensive, necessitating simplification. Therefore, in this study, we aim to simulate a simplified HBM model to analyze its thermal characteristics and predict potential thermal issues that may arise in future designs with increasing stack heights, to find solutions.

II. HBM MODELING AND FEA SIMULATION

1. HBM Modeling

In this research, we simplified the joint and chip structures to efficiently analyze HBM models. This simplification was necessary because detailed modeling of these structures generates an excessive number of mesh elements, significantly increasing the computational load in FEA simulations to the extent that the simulation itself may become impractical. For example, even with the simplification, our 12Hi model generated approximately 10 million mesh elements in the FEA simulations. As a result, many studies adopt simplified bump structures when performing thermal or structural FEA simulations [4-7]. In this study, we utilized a simplified micro-bump structure based on the Cookie model described in a paper by Samsung Electronics [4].

Fig. 1 depicts our simplified model redrawn from previous works [4,5]. It consists of a logic die and core dies (DRAM), and the stacked dies are connected with through-silicon via (TSV). The size of the HBM is designed as 11 by 11 mm. The gap between chips is 15 μm and thickness of silicon chip is 32 μm. The logic die is 0.4 mm larger than other silicon chips. Additionally, the top die is 156 μm, which is larger compared to the other dies. Fig. 1(b) shows the details of joints and materials that are used in our model. We simplified micro bumps using solder and Nickel. And we chose a NCF composed of Epoxy resin as the adhesive material between the chips.

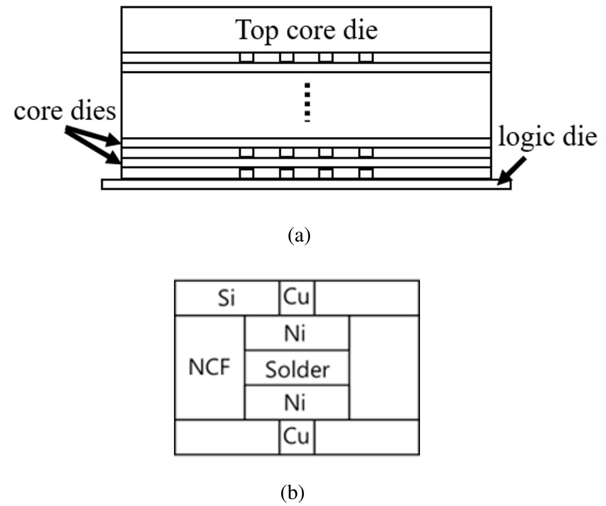


Fig. 1. (a) Cross-sectional view and (b) materials consisting of the contact joint between layers (redrawn from [4,5]).

Table 1. Thermal conductivities of the materials used for the simulation [8,9].

Material	Value
Silicon	148 W/m·K
Copper	400 W/m·K
NCF	0.2 W/m·K
Nickel	90 W/m·K
Solder	48 W/m·K
TIM	2 W/m·K
Heatsink	386 W/m·K
Epoxy molding compound (EMC)	0.9 W/m·K

2. Conditions of Finite Element Analysis (FEA) Simulation

The simulation was simplified by using a quarter model because it has symmetric geometry about the center of the geometry. Also, we simplified simulated model by ignoring Back-End-of-Line (BEOL) of each silicon chips. And we reduced the number of TSVs to 380 and increased the size of TSVs' radius, micro bump's radius, and pitch three times that of the JEDEC standard. Because micro bumps and TSVs are very small elements and there are a large

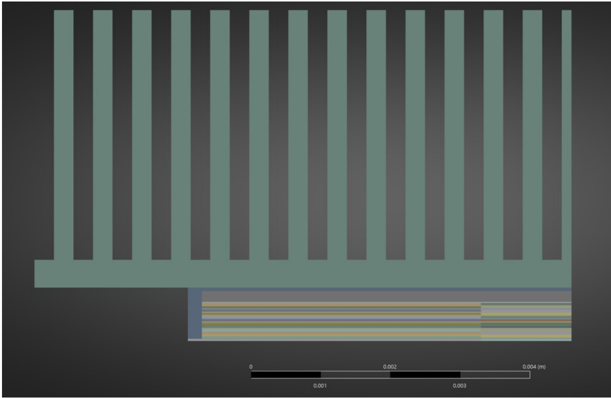
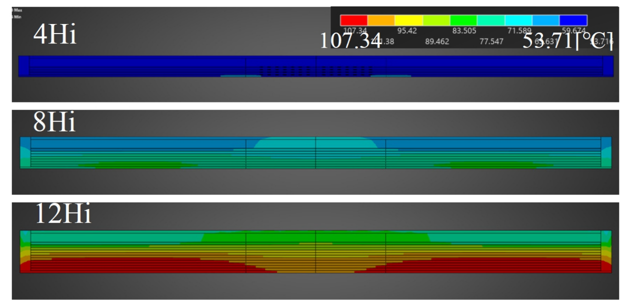
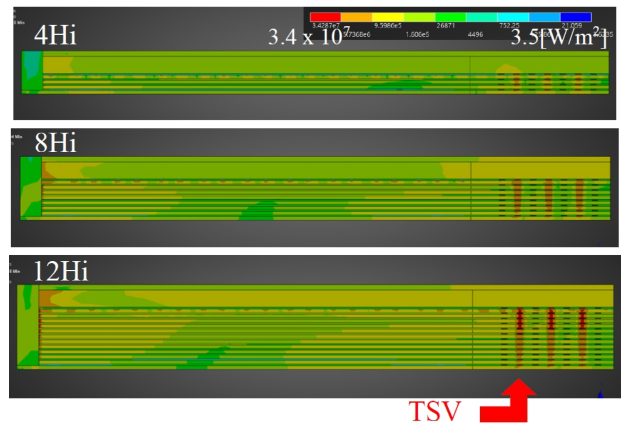


Fig. 2. Cross section view of simulation 12Hi model.



(a)



(b)

Fig. 3. (a) Heat occurring at the cross-section of the 4Hi, 8Hi, and 12Hi models and (b) the heat flux at the TSV cross-section of the quarter models of 4Hi, 8Hi, and 12Hi.

number of them in practice. Excessive number of small mesh elements is one of the major factors hindering 3D FEA analysis.

The sides of the previously described HBM model are molded with EMC, and the top is connected to a heatsink using thermal interface material (TIM) for thermal analysis. The TIM has a thickness of 50 μm. Fig. 2 shows the example of 12Hi HBM simulation model.

The thermal simulation was conducted under steady-state conditions, ambient temperature was set as 45℃ based on JEDEC [5,10]. And it was performed in the forced air-cooling condition using the finite element analysis (FEA) [5]. The thermal conductivities of materials used in simulation are shown in Table 1. We set each core dies consumed 0.5W in TSV area and 1.5W in total bank area. And we set the logic die consumed 2W in TSV area [4].

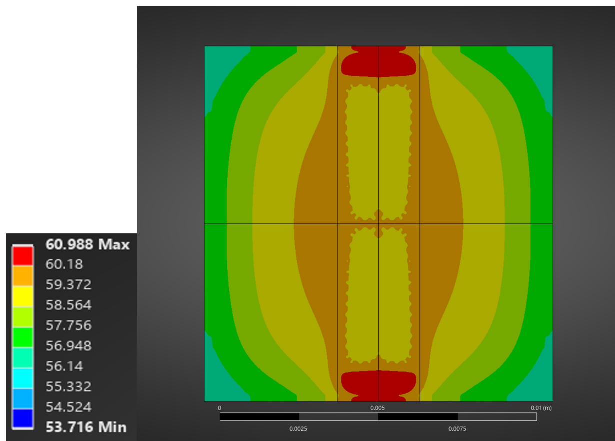
III. Simulation Result and Discussions

Temperatures of the HBM models are depicted in Figs. 3(a) and 4. The 4Hi model recorded a minimum temperature of 53.7°C and a maximum of 60.9°C [11]. The 8Hi model recorded a minimum temperature of 60.2°C and a maximum of 78.2°C. The 12Hi model recorded a minimum temperature of 67.3°C and a maximum of 107.3°C. The 12Hi model shows temperatures similar to the reference, within the margin of error [5].

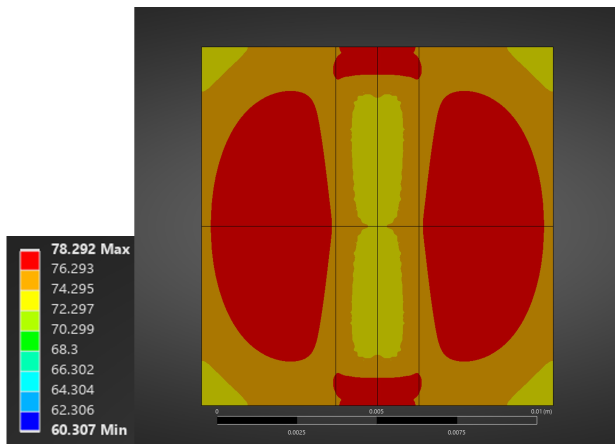
As shown in Fig. 5, the 12Hi model exhibited approximately 25% higher minimum temperature and 76% higher maximum temperature compared to the 4Hi model. It seems that as the number of layers increases, the maximum temperature rises more significantly than the minimum temperature. It was reported that the DRAM temperature should always be maintained below 85°C [12], and the recommended maximum operating temperature for HBM is lower than 95°C [3,13]. In addition, HBM can experience severe issues when its temperature exceeds 120°C [14].

Fig. 3(b) illustrates the heat flux distribution near the TSV areas. In all cases, a significant portion of the heat flux was conducted through the TSV and passed only minimally through the adhesive layer. This behavior is attributed to the adhesive layer's material, NCF, which has a very low thermal conductivity of 0.2 W/m·K. The adhesive layer primarily functions as electrical insulation while simultaneously acting as a thermal barrier, thereby impeding heat transfer.

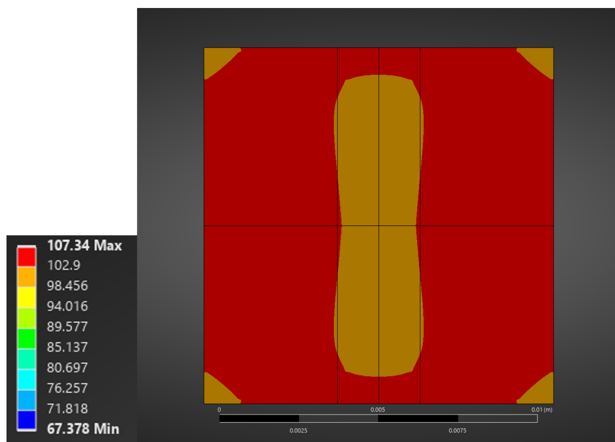
Considering these factors, our simulation results in Fig. 5 show that the maximum temperature already reaches almost 80°C in 8Hi HBM, indicating that it is approaching the allowable limit. Furthermore, in the 12Hi HBM model, the maximum simulated temperature is almost



(a)



(b)



(c)

Fig. 4. Heat occurring at the bottom of the logic die in the (a) 4Hi, (b) 8Hi, and (c)12Hi models.

110°C, which is notably higher than the recommended maximum temperature for HBM. Fig. 5 also indicates that the increase in maximum temperature from 8Hi to 12Hi

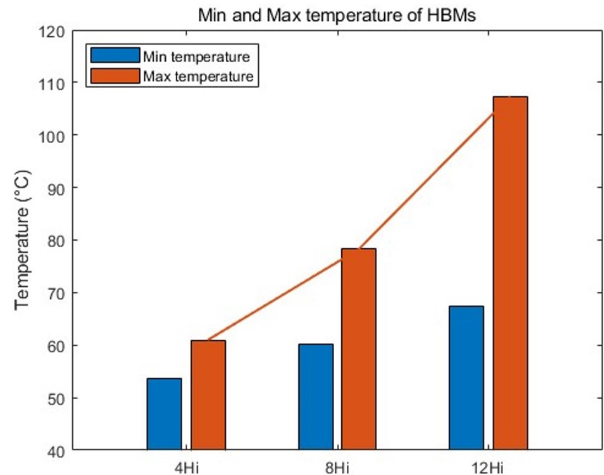


Fig. 5. Minimum and maximum temperature of the simulated HBM models.

is greater than the increase from 4Hi to 8Hi. Thus, critical thermal issues are expected in products under development, such as HBM with more than 12 layers [15], and a technical breakthrough is required.

Furthermore, it was observed that the highest heat concentration consistently occurred at the bottom die across all 4Hi, 8Hi, and 12Hi models, as shown in Fig. 3(a). It seems that the heat around the logic die is not effectively dissipated and appears to be trapped around the bottom in all models. Additionally, in the 4Hi model, higher heat generation is observed around the outer regions of the TSV area in the logic die, whereas in the 8Hi and 12Hi models, heat concentration is observed in the central area of the logic die, which corresponds to the bank area in the core die. This also suggests that the heat generated in the bank regions of the core die in the 8Hi and 12Hi models is not effectively dissipated, potentially affecting both the layers above and below. However, both models exhibited a tendency for the temperature to decrease as it moved upward toward the heatsink. And it is found that the temperature decreases as one moves away from the center. As expected, the greatest heat generation occurred in the 12Hi model due to the additional stacking of core dies, which are the primary heat sources. However, those models exhibited different heat generation patterns due to the differences in stacking.

And we clearly found that heat does not flow through the middle of the chips in the 8Hi and 12Hi models compared to the 4Hi model. This is likely due to the relatively lower height of the 4Hi HBM, where the TSVs are closer to the heatsink, allowing for more efficient heat transfer.

For this reason, significant heat is estimated to have been generated in the bank area of the 8Hi and 12Hi models. It appears that the heat generated in the center of the bank area is trapped between the upper and lower layers, hindering effective heat dissipation. However, because of the absence of precise dimensions, we have only presented the possibility and tendencies.

IV. Conclusions

In this study, we analyze the thermal distribution and heat flux of the 4Hi, 8Hi, and 12Hi HBM models through simulation. We observed that most of the heat is trapped within the lower layers and is transferred through the TSVs and micro bumps. Additionally, we noticed that the adhesive layer does not effectively disperse heat. Based on this, we propose that this may hinder effective heat dissipation. Furthermore, as the stack height increases, it appears that the heat generated in the bank area does not disperse well to the surroundings and seems to be trapped by materials such as NCF. Based on the simulation results, the rate of increase in maximum temperature becomes steeper as the number of layers increases, suggesting that this issue will worsen heat dissipation as more layers are added.

It is therefore estimated that developing HBM with more than 12 layers using similar technology will result in severe thermal issues, and managing these issues will become increasingly challenging. Thus, with the next generation of HBM expected to include higher-stacked models, such as 16 layers, there is a growing need for research into more effective heat dissipation strategies.

Considering these findings, we propose that research into new adhesive materials or bonding methods between the chips is necessary. Such improvements are anticipated to enhance heat transfer from the lower silicon chips to the heatsink, thereby improving overall thermal management. We believe that this research will become increasingly important as the number of layers in HBM continues to rise.

Acknowledgments

This work was supported by the New Faculty Startup Fund from Seoul National University and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) under Grant 2023R1A2C2006661. The authors appreciate the support from Inter-university Semiconductor Research Center, Seoul National University, Seoul, Korea.

References

[1] J. Kim and Y. Kim, "HBM: Memory solution for bandwidth-hungry processors," *Proc. of 2014 IEEE Hot Chips 26 Symposium (HCS)*, IEEE, 2014.

[2] S. S. N. Larimi, B. Salami, O. S. Unsal, A. C. Kestelman, H. Sarbazi-Azad, and O. Mutlu, "Understanding power consumption and reliability of high-bandwidth memory with voltage underscaling," *Proc. of 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2021.

[3] T. Kim, J. Lee, Y. Kim, H. Park, H. Hwang, and J. Kim, "Thermal improvement of HBM with joint thermal resistance reduction for scaling 12 stacks and beyond," *Proc. of 2023 IEEE 73rd Electronic Components and Technology Conference (ECTC)*, IEEE, 2023.

[4] T. Kim, J. Lee, J. Kim, E.-C. Lee, H. Hwang, and Y. Kim, "Thermal modeling and analysis of high bandwidth memory in 2.5 D Si-interposer systems," *Proc. of 21st IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*, 2022.

[5] K. Son, J. Park, S. Kim, B. Sim, K. Kim, and S. Choi, "Thermal analysis of high bandwidth memory (HBM)-GPU module considering power consumption," *IEEE Electrical Design of Advanced Packaging and Systems (EDAPS)*, IEEE, 2023.

[6] J.-Y. Zhou, S.-B. Liang, C. Wei, W.-K. Le, C.-B. Ke, and M.-B. Zhou, "Three-dimensional simulation of the thermo-mechanical interaction between the micro-bump joints and Cu protrusion in Cu-filled TSVs of the high bandwidth memory (HBM) structure," *Proc. of IEEE 69th Electronic Components and Technology Conference (ECTC)*, IEEE, 2019.

[7] S. O, J. Hong, S. Lee, S. Kyung, J. Lee, and K. Kim, "Predicting reliability behavior in HBM packages through numerical simulation," *Proc. of IEEE 73rd Electronic Components and Technology Conference (ECTC)*, IEEE, 2023.

[8] C. Feng, F. Wei, K.-Y. Sun, Y. Wang, H.-B. Lan, H.-J. Shang, F.-Z. Ding, L. Bai, J. Yang, and W. Yang, "Emerging flexible thermally conductive films: mechanism, fabrication, application," *Nano-Micro Letters*, vol. 14, no. 1, 127, 2022.

[9] K. Chatterjee, Y. Li, H. Chang, M. Damadam, P. Asrar, and J. Kim, "Thermal and mechanical simulations of 3D packages with custom high bandwidth memory (HBM)," *Proc. of 2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*, IEEE, 2024.

[10] K. Son, S Kim, H. Park, S. Kim, K. Kim, and S. Park, "A novel through mold plate (TMP) for signal and thermal integrity improvement of high bandwidth memory (HBM)," *Proc. of 2020 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO)*, IEEE, 2020.

[11] A. Agrawal, S. Huang, G. Gao, L. Wang, J. DeLaCruz, and L. Mirkarimi, "Thermal and electrical performance of direct bond interconnect technology for 2.5 D and 3D integrated circuits," *Proc. of 2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, IEEE, 2017.

[12] D. M. Mathew, H. Kattan, C. Weis, J. Henkel, N. Wehn, and H. Amrouch, "Longevity of commodity DRAMs in harsh environments through thermoelectric cooling," *IEEE Access*, vol. 9, pp. 83950-83962, 2021.

[13] J. Hong, S. Cho, G. Park, W. Yang, Y.-H. Gong, and G. Kim, "Bandwidth-effective DRAM cache for GPU s with storage-class memory," *Proc. of 2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, IEEE, 2024.

[14] AMD, "Versal adaptive SoC programmable network on chip and integrated memory controller v1.1," LogiCORE IP Product Guide, v1.1, 91, 9 Aug. 2024.

[15] Y. Chen, D. Zhao, F. Liu, J. Gao, and H. Zhu, "Thermal layout optimization for 3D stacked multichip modules," *Microelectronics Journal*, vol. 139, 105882, 2023.

Jeong Hun Song received his B.S. degree in automotive engineering from the Hanyang University, South Korea in 2023. He is currently working toward the unified master's and doctor's degrees with the Department of Electrical and Computer Engineering, Seoul national University, Seoul. His current research interests include advanced packaging and power module packaging.

Sang Won Yoon received his B.S. degree in electrical engineering from Seoul National University, Seoul, Korea, in 2000 and his M.S. and Ph.D. degrees in electric engineering and computer science from University of Michigan, Ann Arbor, MI, USA, in 2003 and 2009, respectively. From 2009 to 2013, he was a Senior Scientist and a Staff Researcher at the Toyota Research Institute of North America, Ann Arbor, MI, USA, where he conducted research in power electronics and sensor systems for automobiles. From 2013 to 2023, he was Assistant Professor, Associate Professor, and Professor in the Department of Automotive Engineering, Hanyang University, Seoul, Korea. Since 2023, he has been with the Department of Electrical and Computer Engineering at Seoul National University, Seoul, Korea. His research interests include packaging and reliability of semiconductors, electronics for mobility, and their applications.