A Study on the Advanced Chip to Wafer Stack for Better Thermal Dissipation of High Bandwidth Memory

Sangyong Lee
SK hynix Inc.
Icheon-si, Korea
sangyong2.lee@sk.com

Jinwoo Park
SK hynix Inc.
Icheon-si, Korea
jinwoo3.park@sk.com

Jong-Kyu Moon
SK hynix Inc.
Icheon-si, Korea
jongkyu.moon@sk.com

Minsuk Kim
SK hynix Inc.
Icheon-si, Korea
minseok4.kim@sk.com

Gyujei Lee
SK hynix Inc.
Icheon-si, Korea
jerome.lee@sk.com

Kangwook Lee
SK hynix Inc.
Icheon-si, Korea
steward.lee@sk.com

***Abstract*—One of the technical challenges in 2.5D SiP is thermal issue increase with higher performance. In this paper, mass reflow bonding with molded underfll process for chip to wafer stacking of HBM was studied. MUF material and key process parameters were optimized, and 8Hi HBM was successfully demonstrated through MR-MUF method. PKG reliability was verified, and thermal characteristic of HBM was also evaluated. Using MR-MUF method, the maximum junction temperature of memory chips was reduced by 14℃ comparing that of HBM using TC-NCF method at the same DRAM operation condition.**

Keywords—High Bandwidth Memory, TSV, 2.5D SiP, Mass reflow, Molded underfill, and Thermal dissipation

I. INTRODUCTION

High bandwidth memory (HBM) is a prominent solution for high performance computing (HPC) and artificial intelligence (AI) applications. Wide I/O paths between logic and memory devices leads to the reduction of latency and power consumption and to the increase in the bandwidth for data transfer. Chip-stack technologies using through-silicon-via (TSV) and flip chip bonding make it possible to provide high memory capacity. Fig. 1 shows the schematic structure of 2.5D SiP with HBM. However, as the demand for higher speed and capacity continues to increase, thermal issues are expected to be a significant technical barrier to the continued development of HBM products. Currently, in order to enhance the thermal management of 2.5D SiP system consisted of GPU(or CPU), HBM, and Si interposer, many studies are focused on improving the cooling system of SiP or reducing the power of the device itself, but research on how to reduce the thermal resistance of HBM itself from a structural point of view is very limited.

TC-NCF, the thermo-compression bonding with non-conductive film, is commonly used in the chip stack method of HBM packages [2-5]. For better thermal dissipation, more metal bumps are very useful in increasing the thermal dissipation path. However, when the TC-NCF method is applied, there is a limit to increasing the total number of bumps due to the limitation of bonding pressure considering NCF flow-ability and the

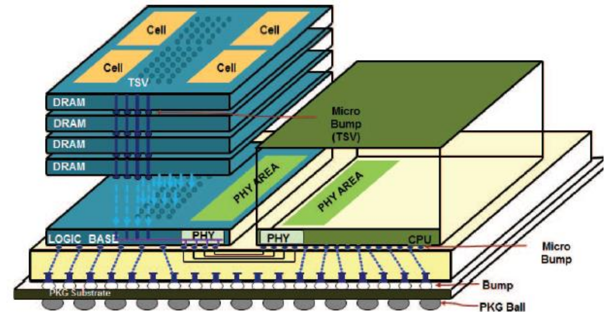

Fig. 1. Schematics of 2.5D SiP with HBM [1]

occurrence of chip cracks in the chip stack process. Therefore, in this study, MR-MUF (Mass reflow bonding with molded underfill) method, which does not use pressure during the chip stack process and thus can use more bumps than the existing TC-NCF, was applied to obtain better thermal dissipation characteristics. MUF material evaluation and MR-MUF process optimization were conducted, and reliability and thermal characteristics were also evaluated.

II. EXPERIMENTAL

A. Materials

A liquid epoxy molding compound (LMC) resin is a key material for the warpage control of the molded wafer, because the molded wafer warpage is mainly caused by the thermal shrinkage difference between Si wafer and LMC [6]. Three different LMC resins were evaluated for minimizing the warpage at room temperature after wafer molding. Table I shows thermo-mechanical properties of LMCs. A blanket test was used to compare their molded wafer warpages by LMC resins as shown in Fig. 2, and LMC C was selected for lower warpage.

MUF material certainly needs to low shrinkage and excellent gap-fill characteristics. After LMC resin selection, the composition of MUF was optimized in terms of filler size, filler content, and SRA (stress release agent) content. Minor warpage

2377-5726/23/$31.00 ©2023 IEEE
DOI 10.1109/ECTC51909.2023.00151

TABLE I. THERMOMECHANICAL PROPERTIES OF LMCS

	LMC A	LMC B	LMC C
T_g (°C)	186	182	152
CTE α1 (ppm/°C)	5	5	8
CTE α2 (ppm/°C)	18	16	21
Modulus@R.T (GPa)	15	25	12.5

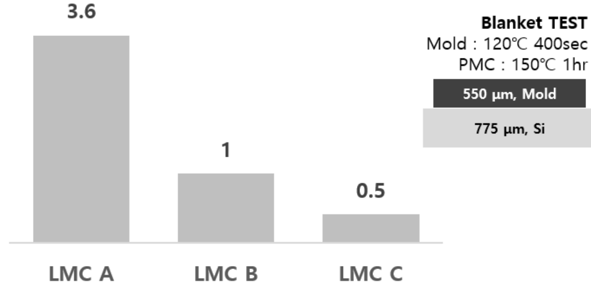

Fig. 2. Blanket test results

control is possible with SRA, but there is a reliability problem such as delamination between chip and MUF due to adhesion degradation [7, 8]. Reliability and low warpage can be achieved at the same time with the optimal composition, such as 77% filler content, 3 μm maximum size, 20% reduction in SRA content, and 20% increase in LMC resin. The thermomechanical properties of the optimized MUF are a CTE (α1) of 18 ppm/°C and a modulus of 11 GPa.

B. Process optimization and Characterization

MR-MUF evaluation was conducted through 8Hi HBM product that has eight memory chips on one logic chip (Fig.3). The Logic chip size is 10.8 mm x 9.8 mm, and memory chip is 10.5 mm x 9.5 mm, respectively. PKG height of HBM is 720 μm, and the chip to chip gap height is 15 μm.

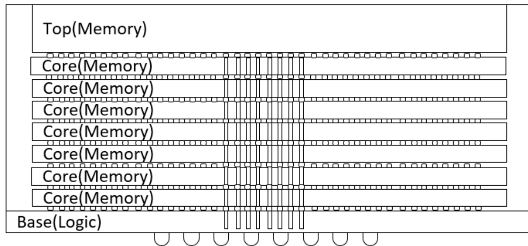

Fig. 3. HBM3 8Hi structure

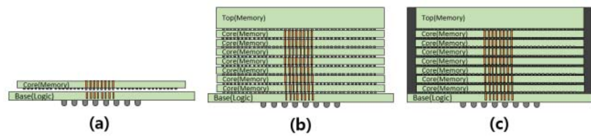

Fig. 4. MR-MUF process flow: (a) chip mount, (b) mass reflow bond, (c) wafer mold for gap-fill

The process flow, as shown in Fig. 4, described steps: (a) core chip mount after flux dipping, (b) multiple chip bond by mass reflow, (c) gap fill and wafer mold with MUF. In the MR process, the chip warpage range was defined to prevent non-wet defects caused by chip warpage and experiments were carried out to expand chip warpage range for non-wet free. In terms of MUF, process optimization was conducted to minimize the gap fill void, especially focusing on the impact of the dispense patterns.

The chip warpage was measured by shadow moire and the joint quality was analyzed through optical microscope (OM) and secondary electron macroscopy (SEM). The Gap fill quality was evaluated not only through scanning acoustic microscopy (SAM), a non-destructive analysis, but also through destructive analysis. Finally, the package reliability test and thermal characteristics were evaluated using proxy packages consisted of 8Hi HBM, Si interposer, and organic substrates.

III. RESULTS AND DISCUSSION

A. Chip to Wafer Stack with MR bonding

One of the most risky item of the multi-chip stack with MR is the non-wet defect caused by chip warpage miss-match. In particular, non-wet defects were mainly found at slice0 where base logic wafer and 1st memory chip are bonded and slice7 where thick top memory chip are bonded (Fig. 5). First of all, the chip warpage range (@220°C) that can be stably joined all chips was defined. As a result, the joint shape was stably formed without a non-wet defect within ±20 μm range of the memory chip warpage (Fig. 6).

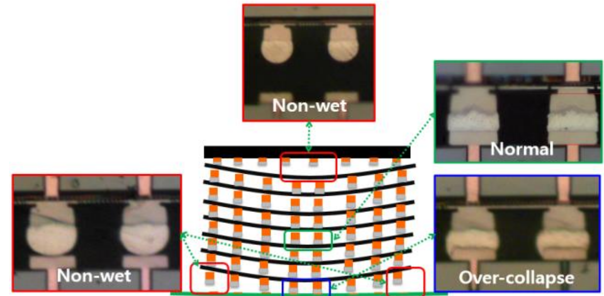

Fig. 5. Schematic and joint shapes due to the chip warpage miss-match

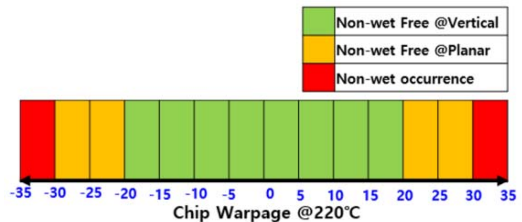

Fig. 6. Non-wet free range according to the chip warpage level

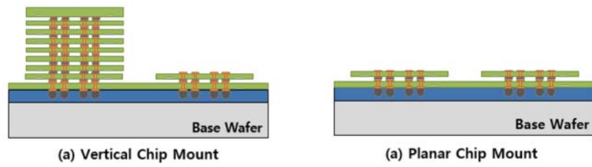

Fig. 7. Comparison of core chip mount sequence

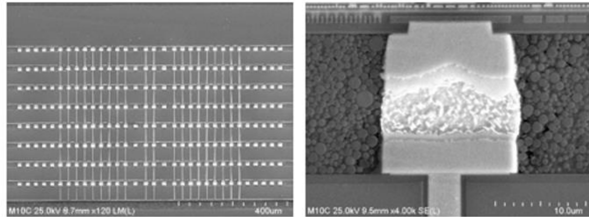

Fig. 8. SEM images of HBM 8Hi joint after MR-MUF process

Furthermore, to expand the chip warpage range for non-wet free, the chip mount order was changed from the vertical method to the planar method. As a result, the chip warpage at solder liquidus temperature was enlarged up to ±30um without non-wet defects. In the wafer, adjacent chips have similar levels of chip warpage, and if the same shape of chips are accumulated, it is easy to manifest the warpage shape. As a result, a non-wet defect occurs if chips at outside of ±20 µm are agglomerated. On the other hand, if dispersed chips in wafer are randomly stacked, the shape of multiple chips is balanced, and the non-wet defects could be suppressed up to the level of chip warpage ±30 µm (Fig. 7, 8).

B. Wafer Molding

The biggest obstacle to applying MUF to the wafer mold process is gap-fill voids that occur intensively in the chip center area. In the MUF process, gap-fill voids are strongly influenced by entrapped air during EMC dispensing (Fig. 9), initial vacuum before gap-fill, and mold pressure [9]. Air traps are known to vary depending on whether the EMC is dispensed onto a wafer or film [10]. It was also confirmed that the face-down method had a greater effect in terms of void suppression than the face-up method in our study. Through SAM analysis, in the case of face-up molded wafers, it was confirmed that a void of 200µm size existed in the center of most chips, and in the case of face-down molded wafers, it was confirmed that the gap fill quality was improved as shown in Figure 10. However, even when the face-down mold was applied, voids were still observed in the chip near the center of the wafer, which is the EMC dispensing area. The mechanism of the void formation is shown in Fig. 11. During the molding process, voids caused by air traps are collected in the center of the chip by EMC flow, and it is assumed that voids remain in the center of the chip when it is balanced with the mold pressure.

In this study, the maximum mold pressure and the minimum vacuum level were applied to minimize the void, and various EMC dispense patterns were evaluated as shown in Fig. 12. Using center dispensing resulted in gap-fill voids due to initial air trapped in the central area of the wafer as similar to face-up molded wafer. In addition, a spiral pattern and several serpentine patterns were used, but voids were still found in certain areas as a result of SAM analysis. To improve this phenomenon, we tried to minimize the initial air trap of the chip by aligning the EMC dispensing line to the center of each chip as shown in Figure 13, and the void rate could be achieved up to 0%.

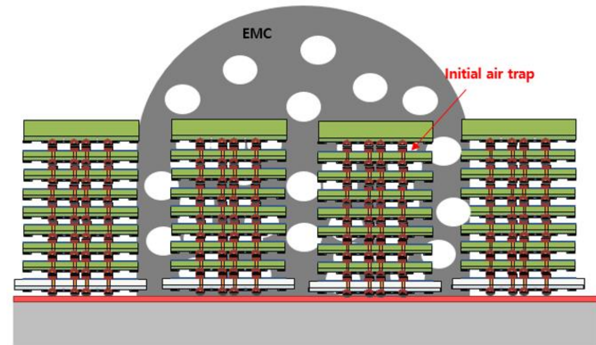

Fig. 9. Schematic diagram of initial air trap at face up type mold

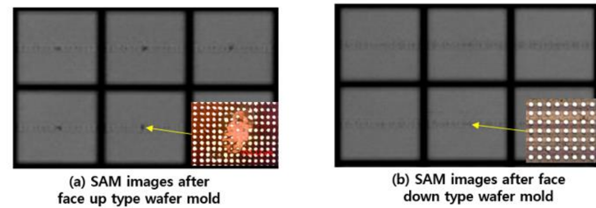

Fig. 10. SAM and flat grinding images after wafer mold

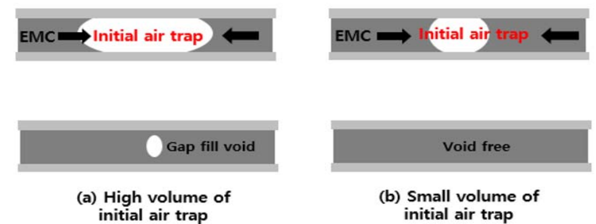

Fig. 11. The gap fill void formation for different initial air trap

Center Dispense	Spiral Dispense	Serpentine	Serpentine Imp.1	Serpentine Imp.2
Loss: 21.3%	Loss: 11.7%	Loss: 7.4%	Loss: 0.8%	Loss: ~0%

Fig. 12. EMC dispense pattern vs. SAM defect loss trend

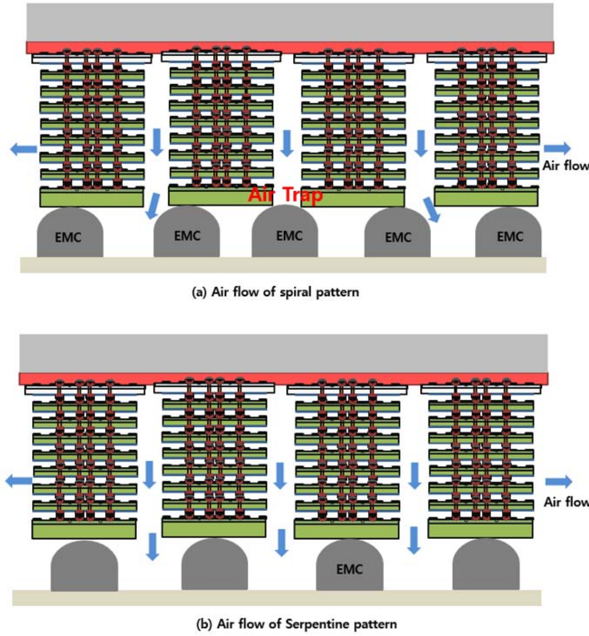

Fig. 13. Air flow comparison of the spiral and serpentine pattern, respectively

C. *Reliability and thermal characteristic*

The 8Hi HBM was assembled with Si interposer and organic substrate to evaluate the thermo-mechanical reliability in 2.5D SiP-like environment. The reliability tests were done in accordance with the JEDEC standard conditions, and the MR-MUF test vehicle passed all test without any issues as shown in Table I.

Thermal characteristic of MR-MUF 8Hi HBM was compared to that of TC-NCF 8Hi HBM at the same DRAM operation condition, IDD4. The junction temperature(T_j) of DRAM was measured at the pin speed 2.0Gbps and VDD 1.26V. As a result, average T_j of MR-MUF 8Hi HBM was 14℃ lower than that of TC-NCF 8Hi HBM as shown in Fig. 14. It can be explained as improved thermal dissipation characteristic by lowering the thermal resistance of HBM itself by connecting more metal bumps than the TC-NCF method through the MR-MUF method. Therefore, in terms of thermal throttling, better performance of SiP systems can be expected using MR-MUF HBM.

IV. Conclusion

In this study, the 8Hi HBM through mass reflow bonding and molded underfill was successfully demonstrated. Excellent bonding quality was achieved with high bond accuracy using mass reflow bonding, and MUF voids could be eliminated after the wafer mold process by optimizing process parameters and dispensing patterns. Their PKG reliability was verified by temperature cycling (TC), unbiased high accelerated stress test (uHAST), high temperature storage (HTS), and temperature humidity bias life test (THB) after preconditioning test. Enhanced thermal dissipation characteristic was also achieved.

TABLE II. Reliability test

Test item	Condition	Results
Precon.	L2a & Reflow 6 cycles	Pass
TC (B)	-55/125℃, 1000 cycles	Pass
uHAST	130℃/85%RH, 96hrs	Pass
HTS	150℃, 1008hrs	Pass
THB	85℃/85%RH w/ Vmax, 1008hrs	Pass

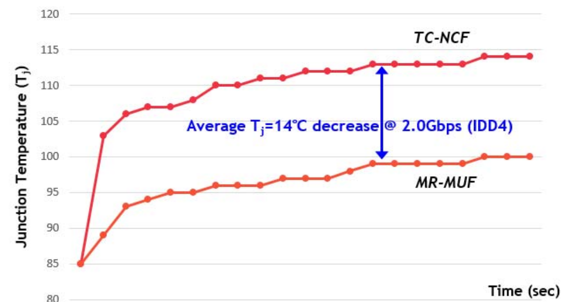

Fig. 14. DRAM Tj comparison: TC-NCF vs. MR-MUF

References

[1] D. U. Lee et al., "Design considerations of HBM stacked DRAM and the memory architecture extension," 2015 IEEE Custom Integrated Circuits Conference (CICC), San Jose, CA, USA, 2015, pp. 1-8.

[2] H. Clauberg, A. Rezvani, V. Venkatesan, G. Frick, B. Chylak and T. Strothmann, "Chip-to-Chip and Chip-to-Wafer Thermocompression Flip Chip Bonding," 2016 IEEE 66th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, USA, 2016, pp. 600-605.

[3] J. -S. Jung et al., "A Study of 3D Packaging Interconnection Performance Affected by Thermal Diffusivity and Pressure Transmission," 2019 IEEE 69th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, USA, 2019, pp. 204-209.

[4] S. J. Kim, H. Kim, J. Hong, O. Kwon and H. Lee, "Process optimization of micro bump pitch design in 3-dimensional package structure," 2021 IEEE 71st Electronic Components and Technology Conference (ECTC), San Diego, CA, USA, 2021, pp. 1870-1875.

[5] J. P. Hong et al., "Novel method for NCF flow simulation in HBM thermal compression bonding process to optimize the NCF shape," 2022 IEEE 72nd Electronic Components and Technology Conference (ECTC), San Diego, CA, USA, 2022, pp. 519-523.

[6] J. Kim et al., "Large Area Encapsulation: Solid Type Epoxy Molding Compound," 2018 International Wafer Level Packaging Conference (IWLPC), San Jose, CA, USA, 2018, pp. 1-5

[7] T. Kamimura et al., "Development of Liquid Compression Molding (LCM) Material for Low Warpage", Proceeding of 50th International Microelectronics Assembly and Packaging Society (IMAPS), pp. 25-28, 2017

[8] J. Chao et al., "Low Warpage Liquid Compression Molding (LCM) Material for High Density Fan-out and Wafer Level Packaging Applications," 2020 IEEE 70th Electronic Components and Technology Conference (ECTC), Orlando, FL, USA, 2020, pp. 924-930.

[9] L. Bu, F. X. Che, M. Z. Ding, S. C. Chong and X. W. Zhang, "Mechanism of Moldable Underfill(MUF) process for fan-out wafer level packaging," 2015 IEEE 17th Electronics Packaging and Technology Conference (EPTC), Singapore, 2015, pp. 1-7.

[10] Y. Kajikawa, "Fan-Out Wafer-Level Packaging Advanced Manufacturing Solution for Fan-Out WLP/PLP by DFD (Die Face Down) Compression Mold," 2020 International Wafer Level Packaging Conference (IWLPC), San Jose, CA, USA, 2020, pp. 1-8