SoIC for Low-Temperature, Multi-Layer 3D Memory Integration

M. F. Chen, C. S. Lin, E. B. Liao, W. C. Chiou, C. C. Kuo, C. C. Hu, C. H. Tsai, C. T. Wang and Douglas Yu
Integrated Interconnect & Packaging, R&D, Taiwan Semiconductor Manufacturing Company, Ltd.
166, Park Ave. II, Hsinchu Science Park, Hsinchu 300-75, Taiwan, R.O.C.
Phone: 886-3-5636688 Ext. 722-3896, Email: mfchen@tsmc.com

***Abstract*— A low-temperature System-on-Integrated-Chip (LT-SoIC) technology has been successfully applied to multi-layer 3D memory cube integration, which enables high bandwidth, low power and small footprint memory for future HPC applications. In addition, using the technology, each memory layer can be thinned to required thickness to maintain total height while supporting more layer counts.**

Two LT-SoIC processes were presented. One is through-via reveal last and the other one is through-via reveal first. The through-via revealing of each stacked die is one of the most critical process steps. Various conditions of planarization on chip backside after the through-via revealing were studied to mitigate the corner or edge rounding issue, as it may cause poor bonding. By improving the back-side revealing process, we can achieve good bonding for multi-layer stacking, including DRAM stacking with 4-Hi/ 8-Hi/ 12-Hi, and SRAM stacking with 4-Hi, using the LT-SoIC technology.

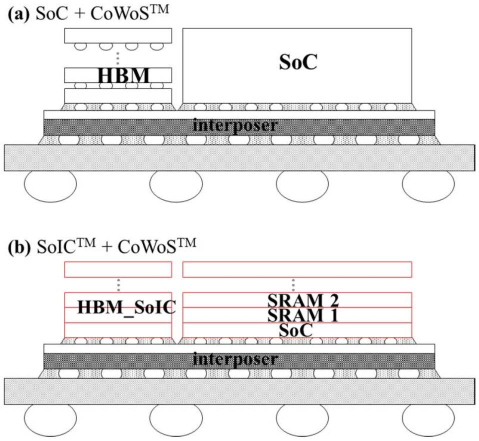
The bonding quality of the LT-SoIC is measured using I-V curve and shear stress equipment. Linear I-V curve was obtained to show it is an Ohmic contact and the bonding strength of >2.5J/m^2 was measured to show good bonding force. The through-via chain resistances for 4-Hi SRAM and 4/8/12-Hi DRAM and the breakdown voltage for 8/12-Hi DRAM were measured, too. Reliable results were obtained to indicate the integrity of the through-via chains. After that, the bandwidth density and power consumption between typical 3D memory and the SoIC memory were compared. High bandwidth density and low power consumption are obtained in the SoIC memory. Clear advantages of the SoIC CoW stacking technology are realized for multi-layer memory stacking at low temperature.

Keywords—3DIC; SoIC; 5G/AI; High-Bandwidth Memory (HBM); Multi-layer stacking; CoW

I. INTRODUCTION

Due to the rapid growth of AI, 5G, and HPC market, the demand to integrate high-bandwidth 3D memory cube with logic chip has been increased significantly, as the integration provides advantages of low latency, low power consumption, low input voltage, high memory bandwidth and small form factor for system applications. To achieve this, advanced wafer-level heterogeneous system integration technology, such as Chip on Wafer on Substrate (CoWoS), has been used [1-4]. For the high-bandwidth 3D memory cube, called high bandwidth memory (HBM) DRAM, it is a multi-layer memory stacking, where μ-bumps along with through-via connection is used for the multi-layer die interconnect performed by thermal compression bonding (TCB). However, the development of the HBM stacking technology is limited by its large μ-bumps, high bonding force, and high bonding temperature for future memory cube requirements, more bandwidth and more memory capacity. Smaller bond pitch and more memory stacking, such as 12-Hi or 16-Hi stacking, will be required for the future memory. Therefore, a new 3D memory system integration technology is desired to resolve the challenges of bond pitch, die thickness, and data retention issues in current HBM integration technology.

We have recently reported the industry-first SoIC technology, called TSMC-SoIC^{TM} [5, 6], which enables sub-10 μm bonding pitch, thin-die stacking, high flexibility for heterogeneous integration. By application of the SoIC technology, it can directly resolve the challenges in current HBM integration technology and facilitate HBM performance development. In this paper, we develop a new SoIC technology, called low-temperature SoIC (LT-SoIC) process to realize multi-layer 3D memory integration. It is successfully applied to stack both of 3D SRAM and 3D DRAM memory cube with good electrical and early reliability performance. Fig. 1 illustrates one potential embodiment of an advanced system integration [7, 8], where a SoC in Fig. 1 (a) can be thinned and attached with SRAM cube, and the conventional μ-bump HBM can be replaced by the SoIC_HBM, in Fig. 1(b). The integration of the SoC with SRAM memory cubes using the 3D SoIC technology can enable higher system performance, as it is boosted by providing extra memory cache. On the other hand, the SoIC integrated DRAM memory cubes can offer higher memory density, bandwidth, and power efficiency.


(a) SoC + CoWoS^{TM}

(b) SoIC^{TM} + CoWoS^{TM}

2377-5726/20/$31.00 ©2020 IEEE
DOI 10.1109/ECTC32862.2020.00139

Figure 1. System integration of 3D memory (a) conventional HBM integrated with SoC, and (b) SoIC HBM integrated with SoIC built SRAM/SoC cubes

II. PROCESS CHALLENGES AND OPTIMIZATION

To achieve multi-layer die stacking with more stacking numbers, such as 8-Hi or 12-Hi, to have lower profile or to maintain the profile at a certain total stacking height, thinner die stacking is required. In thinner die stacking process, finer through-via size and depth can be obtained easily, which would offer higher bonding density and higher memory bandwidth for applications (Detailed calculation and comparison will be discussed in section IV). However, wafer thinning process for very thin dies can easily induce mechanical damage, which may result in drift of memory retention characteristics. Also the copper bonding quality in the thin die stacking process will affect the bond electrical performance and the reliability. Two different process approaches for the SoIC technology for multi-layer 3D memory die stacking will be presented and discussed. Optimization of the process to obtain good integration quality will be discussed, too.

Two SoIC process flows, through-via reveal last and through-via reveal first, have been developed. In the through-via reveal last flow, memory die with a modest thickness is bonded first, next thinned down to a target thickness, processed by through-via reveal and passivation disposition and then, bonded by another memory die to repeat the process. The process is simple, but it easily created die corner/edge rounding problem, particularly for those dies at the wafer edge. On the contrary, the through-via reveal first process starts from the memory wafer temporarily bonded to a glass carrier and then the bonded wafer is thinned down to required die thickness to reveal the through-via and do passivation deposition. After that, the thin wafer was de-bonded from the carrier and diced to discrete dies. The dies were bonded one on one to form a stacking memory cube. The process was free from rounding issue, but it was challenged by surface contamination problem. Both of these two SoIC process flows have been optimized to offer more flexibility in integration.

- *Through-via reveal last: challenge of die corner/edge rounding*

For process control, the most challenge in the through-via reveal last flow (through-via reveal after CoW bonding) is the die corner/edge rounding issue, particularly for the dies in the wafer edge area. After grinding and CoW bonding processes, CMP process is applied to remove the grinding mark to produce a surface with suitable roughness and Cu pad profile for SoIC bonding. If there is no filling material in the gap between each die after CoW bonding, the die corner and edge will get rounding due to polishing stress concentration in the CMP process. The rounding effect was getting worst for the dice on the wafer edge because of larger gaps between the peripheral dies and CMP retaining ring. To mitigate the issue, the distribution of CMP down force was adjusted to alleviate the rounding behavior at wafer edge successfully. The rounding effect can be suppressed by manipulating CMP selectivity. Furthermore, by the application of double-layer bonding-film scheme (including a stop layer and a sacrificial layer), the rounding depth was improved by 78% compared to the single-layer bonding-film scheme, as shown in Fig. 2. To completely mitigate the rounding effect at wafer edge, dummy die stack at wafer edge is also proposed to protect inner functional die stacking

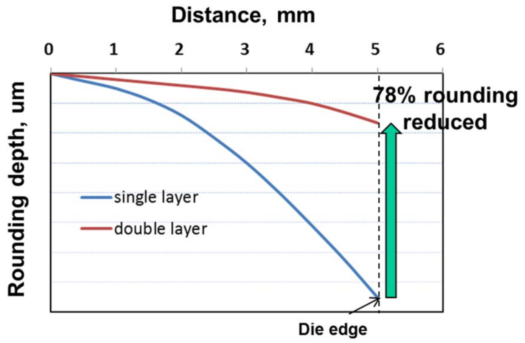


Figure 2. CMP edge rounding improvement by double-layer bonding film scheme

- *Through-via reveal first: challenge of surface cleanness control*

The SoIC process imposes stringent requirements on the cleanness of copper pad and surrounding dielectric surface for the bumpless joint technology. In the through-via reveal first flow, both glue layers and sacrificial organic material are used in thermal processes, such as carrier bonding and backside passivation film deposition, after through-via reveal. All of these materials must be completely removed to ensure the surface cleanness for bonding.

Generally, glue residue was easy to be found on the SoIC bonding pad, as shown in Fig. 3(a), if the integration process is not well controlled. Before carrier bonding, the glue on carrier and the sacrificial layer on memory wafer are baked, respectively, to drive out solvent. To ensure we can remove the glue material completely, the baking condition was optimized to achieve bubble-free carrier bonding. Furthermore, extra heating locally may enhance the reaction between glue and sacrificial layer, which had to be taken into consideration. Additionally, during the backside passivation film deposition, wafer surface was heated up by plasma. The Cu-filled through-via provided a high thermal conduction path compared to other regions, which resulted hot spot on the SoIC bonding pad and the glue on the top. It caused the glue to be hardened and to be removed hardly. To achieve glue residue-free SoIC bonding pad, as shown in Fig. 3(b), the optimization of sacrificial layer baking condition and low thermal-budget passivation film deposition process were developed.

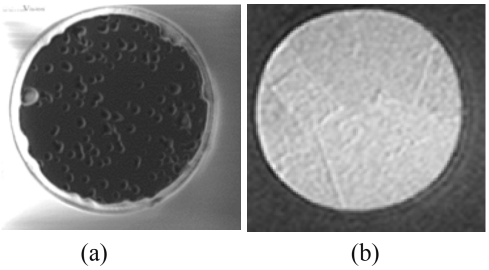


Figure 3. SoIC bonding pad with (a) Glue residue; (b) clean surface

The silicon debris adhered to the wafer surface from the sacrificial layer delamination, generated in grinding or dicing, is shown in Fig. 4(a), which resulted in non-bond happened at die corner or edges. In the through-via reveal first flow, sacrificial layers were applied on both front-side and backside of memory wafer to prevent surface contamination from bonding glue or dicing tape adhesive. A strong bonding between the wafer and the sacrificial layer is demanded as the wafer experienced high mechanical impact or vibration force during backside grinding or mechanical dicing process. To avoid the sacrificial layer delamination to occur, the backside sacrificial layer was particularly baked in a vacuum environment with optimized temperature to maintain its integrity. Therefore, the surface cleanness can be achieved for SoIC bonding, as shown in Fig. 4(b). Both through-via reveal first and through-via reveal last process were used for SRAM and DRAM cube making from 4-Hi, 8-Hi to 12-Hi. Their performances were compared.

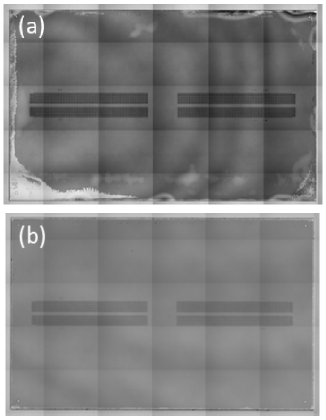


Figure 4. OM images when sacrificial layer bonded to wafer surface (a) sacrificial layer peeling and die corner/edge contaminated; (b) free of peeling and contamination

III. MULTI-LAYER 3D MEMORY BY LT-SoIC

To obtain multi-layer die stacking, high-temperature bonding process is normally applied to ensure good joint quality and low contact resistance. However, the high-temperature process conflicts with the process temperature restriction of memory chips, which hinders the development of multi-layer 3D memory, especially for higher layer memory, such as 8-Hi or 12-Hi.

A low temperature (LT) SoIC bonding process, with thermal budget comparable to the lead-free solder reflow, has been successfully developed, which provides a promising solution to mitigate the challenge from high temperature process and also show merit of cost reduction, resulting from the lower process temperature. In the early stage of the LT-SoIC technology development, bad joint was observed due to insufficient metal diffusion. I-V curve measurements of the LT-SoIC bonds were used to show the bonding quality. Fig. 5(a) shows a Schottky behavior on the interconnect chains when the LT-SoIC bonds were in bad joint quality, and Fig. 5(b) shows the achievement of linear behavior in Ohmic contact on those interconnect chains with good joint quality. The physical images of the LT-SoIC multi-layer 3D DRAM samples with 4, 8 and 12-Hi are shown in Fig. 6 and the 3D-Xray image, taken from 12-Hi DRAM stacking, is shown, too.

Based on the LT-SoIC bonding process, a bonding strength of >2.5J/m^2 was obtained, which indicates the process is able to provide high bonding strength at low thermal budget which is required for 3D memory integration. The developed LT-SoIC process is ready for multi-layer 3D memory integration application. In the following sections, electrical and reliability data of the LT-SoIC technology for 3D memory, including SRAM and DRAM, will be studied and presented.

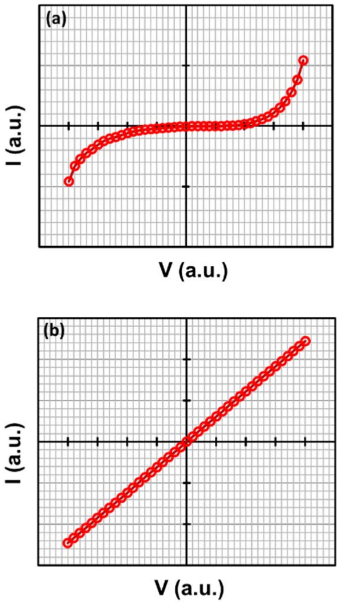


Figure 5. IV curve of through-via chain in 4-Hi DRAM memory showing LT-SoIC developing CIP, from (a) Schottky contact, to (b) Ohmic contact

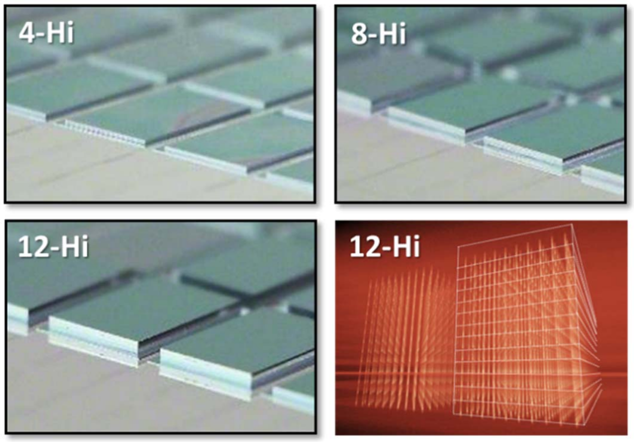


Figure 6. Demonstration of physical images taken from LT-SoIC multi-layer DRAM samples, 4-Hi/ 8-Hi/ 12-Hi, including a 3D-Xray image of 12-Hi DRAM

IV. COMPARISON BETWEEN TYPICAL 3D MEMORY and LT-SoIC 3D MEMORY

High bandwidth, high capacity, and low power consumption memory is desired for AI, 5G, and HPC applications. To achieve this, JEDEC, a global leader in developing open standards and publications for the microelectronics industry, had proposed a high bandwidth memory (HBM) DRAM standard for the applications. The HBM DRAM is a 3D memory stacking structure using μ-bump for chip to chip interconnect. However, there are challenges in bandwidth, capacity, thermal dissipation and stacking height for the 3D μ-bump stacking memory when more memory chip stacking is required for new HBM DRAM. To solve the challenges, different system integration technology has to be proposed. We have shown the SoIC technology has stronger process advantages against other conventional packaging technology [1]. The new technology, LT-SoIC, is a good solution to solve the challenges as it has the characteristics of SoIC, allowing higher-tier stacking using thinner dies and smaller bond-line thickness. And at the same time, it uses lower temperature bonding, not to hurt memory performance. Table I presents a system performance comparison between a typical μ-bump 3DIC and a LT-SoIC with 9-tier multiple stacking, including a logic base die and 8 DRAM dies. It shows the LT-SoIC outperforms the μ-bump 3DIC by 17%, 25%, and 36% in bandwidth density where the LT-SoIC die thickness is 44, 35, and 25 μm, respectively, compared with the 50 μm thickness for the μ-bump 3DIC memory. The Z-form factor (height) of the LT-SoIC is 64%, 50%, and 36% of that of the μ-bump 3DIC, respectively. Furthermore, the power consumption can also be reduced by 9%, 14%, and 19% respectively, associating with the decrease of die thickness.

Table I. Comparison of bandwidth density and power consumption between conventional (μ-bump) 8-Hi DRAM and LT-SoIC 8-Hi DRAM

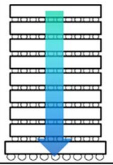
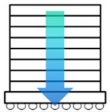
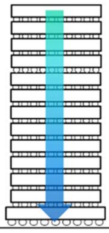
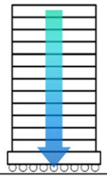
Package Type (Controller + **8*DRAM**)		**Typical 3D** 	**LT-SoIC 3D** 		
Structure	Bond Technique	μ-bump	LT SoIC-bond		
	Z-form factor (die thickness)	1X (50um)	0.76X (45um)	0.59X (35um)	0.42X (25um)
Electrical Performance	Bandwidth density (Bandwidth / area)	1X	1.17X	1.25X	1.36X
	Power consumption (Energy / bit)	1X	0.91X	0.86X	0.81X

Table II. Comparison of bandwidth density and power consumption between conventional 12-Hi DRAM and LT-SoIC 12-Hi DRAM

Package Type (Controller + **12*DRAM**)		**Typical 3D** 	**LT-SoIC 3D** 		
Structure	Bond Technique	μ-bump	LT SoIC-bond		
	Z-form factor (die thickness)	1X (50um)	0.64X (45um)	0.50X (35um)	0.36X (25um)
Electrical Performance	Bandwidth density (Bandwidth / area)	1X	1.18X	1.27X	1.28X
	Power consumption (Energy / bit)	1X	0.92X	0.86X	0.81X

The LT-SoIC applied to a 12-Hi DRAM stacking for next generation of 3D memory had been studied, too. The bandwidth density and power consumption of the 12-Hi DRAM using μ-bump and LT-SoIC bond, respectively, were analyzed and compared in Table II. The LT-SoIC bond 3D memory outperforms the μ-bump 3D memory by 18%, 27% and 28% in bandwidth density where the LT-SoIC die thickness is 45, 35, and 25 μm respectively, compared with the 50 μm thickness for the μ-bump 3DIC memory. The Z-form factor (height) of the LT-SoIC is 64%, 50%, and 36% of that of the μ-bump 3DIC, respectively. For the power consumption, the LT-SoIC is lower than the μ-bump 3DIC by 8% to 19%, according to the die thickness.

In summary, the LT-SoIC technology can be applied to the formation of next generation of 3D DRAM memory as it enables bumpless integration process at lower temperature with much smaller thickness bondline and thinner dies stacking. It helps the gain of bandwidth density and the reduction of power consumption. Moreover, the LT-SoIC technology can be applied to much higher-tier stacking, such as 16-Hi, and 24-Hi, of 3D memory. The more advanced process is under development.

V. ELECTRICAL PERFORMANCE

To check integration integrity of the LT-SoIC 3D memory, the interconnect chain resistance and its stability were measured. The LT-SoIC SRAM and DRAM samples with various stacking numbers were prepared and electrically tested. The chain resistance data for 4-Hi SRAM cube are shown in Fig. 7 and Fig. 8, respectively. Fig. 7 shows a very steep of chain resistance accumulation line, which indicates the chain resistance across different 4-Hi SRAM cubes, tested on the same chain, is the same. Fig. 8 shows the linearity of resistance between different lengths of chains in the same 4-Hi SRAM cube. The results of Fig. 7 and Fig. 8 tell us the uniformity of the LT-SoIC bonding quality across the wafer and the robustness of the bonding quality in long chain interconnection. Besides the test of 3D SRAM memory cube, the chain resistance of the 3D DRAM memory was also tested. The chain resistance of 4-Hi/8-Hi/12-Hi 3D DRAM cube, respectively, is shown in Fig. 9. A linear resistance behavior was obtained between the different stacking number DRAMs. This demonstrated the stability and capability of the LT-SoIC bond technology for multi-layer stacking.

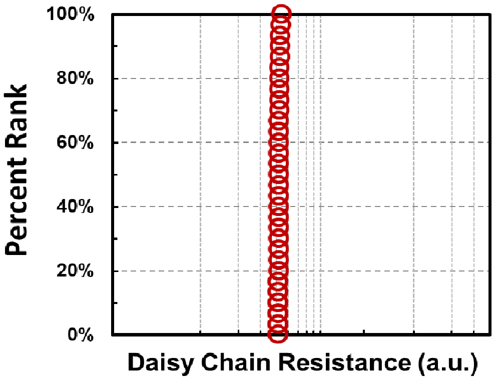


Figure 7. Through-via Chain resistance of LT-SoIC 4-Hi SRAM, and data is consistent values across different samples

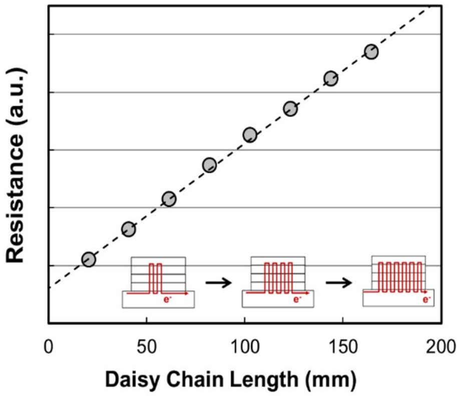


Figure 8. Through-via Chain resistance is linear between different length of chains within the same LT-SoIC 4-Hi SRAM

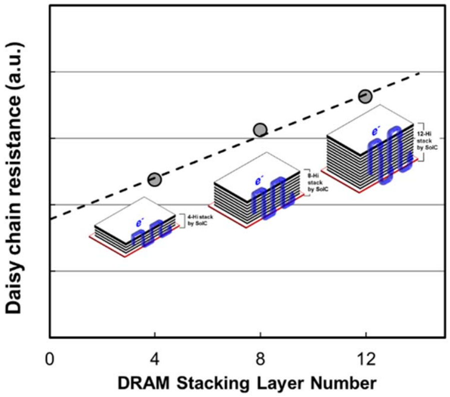
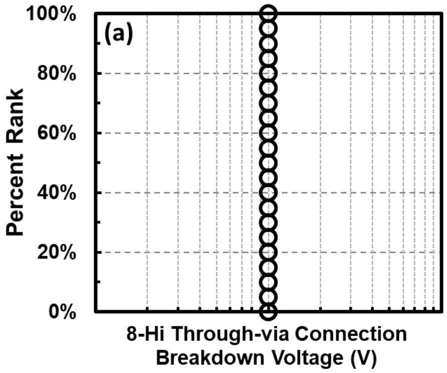


Figure 9. Through-via daisy chain resistance comparison between different DRAM stacking layer number, 4-Hi/ 8-Hi/ 12-Hi

VI. RELIABILITY PERFROMANCE

Thermal compression bond (TCB) and non-contact film (NCF) bond are the main applied bonding techniques for the μ-bump based 3D memory cubes. However, the process in TCB and the interlayer epoxy material in NCF bond brings concerns of solder bridging and filler entrapment in the bump joints, respectively, consequently resulting interconnect reliability issues. The LT-SoIC bond technology provides a bump-less solution without using epoxy material, which can directly solve those issues to mitigate the reliability concerns in the μ-bump 3D memory. To evaluate the reliability quality of LT-SoIC bond, breakdown voltages (Vbd) for the through-via connection in the 8-Hi and 12-Hi built DRAM memory samples were measured, as shown in Fig. 10. Fig. 10(a) is the accumulated breakdown voltage data for the 8-Hi through-via connections from different samples and Fig. 10(b) is the data for the 12-Hi through-via connections. The samples had similar breakdown voltage. The result demonstrated the structural integrity of long through-via chains built by LT-SoIC bond, no matter they were 8-Hi or 12-Hi 3D memory cubes.

The chain resistance for the through-via connection in the 8-Hi and 12-Hi samples were measured, too. The resistance at time zero (T_0) and after Thermal-Cycling Test B (TC-B) 200 cycles, respectively, was measured and the results were shown in Fig. 11a for 8-Hi and Fig. 11b for 12-Hi DRAM cubes. The result at TC-B 200 cycles matched to the result at T_0, which indicated the LT-SoIC bonded through-via chain is very reliable.



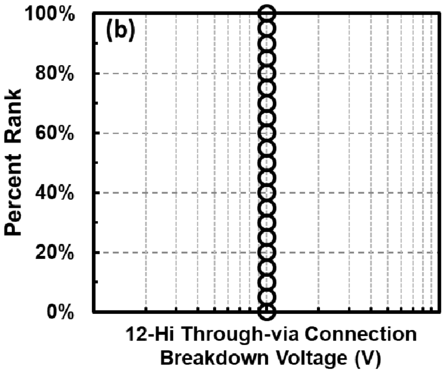


Figure 10. Breakdown voltage of LT-SoIC 8-Hi DRAM (in Figure 10a) and 12-Hi DRAM (in Figure 10b), where consistent Vbd data was presented across different samples

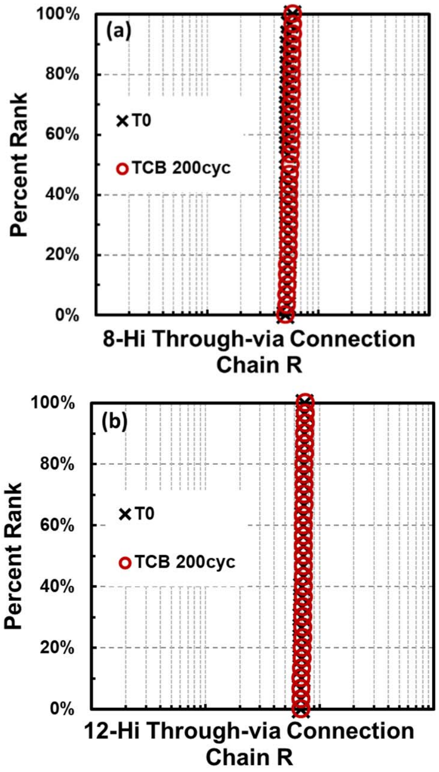


Figure 11. Comparison of through-via chain resistance under T_0 and TCB-200 for the (a) LT-SoIC 8-Hi DRAM samples, and (b) LT-SoIC 12-Hi DRAM samples

VII. CONCLUSION

In this paper, the LT-SoIC process had been successfully developed and the electrical and reliability performance were measured. Two process flows were applied to achieve the LT-SoIC technology. One is through-via reveal-last and the other one is through-via reveal-first process. For LT-SoIC copper bond, the linear I-V curve and >2.5J/m^2 shear strength were obtained. For through-via chain, the resistance and breakdown voltage (Vbd) were measured over tens of thousands through-via connected interconnection. Promising Vbd and TC-B data supported the integrity of the through-via interconnections. The LT-SoIC technology can provide thin die multi-layer stacking to short the interconnect length between dies for better bandwidth density and power consumption, compared to conventional μ-bump 3DIC stacking technology. Currently, we have achieved 12-Hi DRAM stacking by the LT-SoIC technology, and furthermore, higher-tier stacking, such as 16-Hi stacking, will be developed. In conclusion, the robust and reliable LT-SoIC bonding technology provides a potential solution to replace conventional μ-bump multi-layer memory stacking for the AI/5G applications.

ACKNOWLEDGEMENT

Authors would like to acknowledge TSMC front-end and back-end Fabs, PE, and Quality & Reliability Assurance colleagues for their good supports and contributions to this paper.

REFERENCES

[1] Doug C.H. Yu, "Advanced System Integration Technology Trend", SEMICON Taiwan SiP Global Summit, Taipei, Taiwan, 2018

[2] Doug C.H. Yu, "WLSI and Wafer Foundry Growth with Moore's Law and More-than-Moore, and Vice Versa", IWLPC Keynote speech, San Jose, CA, 2018

[3] An-Jhih Su, Terry Ku, Chung-Hao Tsai, Kuo-Chung Yee, Douglas Yu, "3D-MiM (MUST-in-MUST) Technology for Advanced System Integration", in 2019 IEEE 69th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, 2019

[4] Chien-Fu Tseng, Chung-Shi Liu, Chi-Hsi Wu, Douglas Yu, "InFO (Wafer Level Integrated Fan-Out) Technology", in 2016 IEEE 66th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, 2016

[5] C.C. Hu, M.F. Chen, W.C. Chiou and Doug C.H. Yu, "3D Multi-chip Integration with System on Integrated Chips (SoIC™)", 2019 Symposium on VLSI Technology, Kyoto, Japan, 2019

[6] Ming-Fa Chen, Fang-Cheng Chen, Wen-Chih Chiou, Doug C.H. Yu, "System on Integrated Chips (SoIC) for 3D Heterogeneous Integration", 2019 IEEE 69th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, 2019

[7] S. Y. Hou et al., "Wafer-Level Integration of an Advanced Logic-Memory System Through the Second-Generation CoWoS Technology", IEEE Transactions on Electron Devices, vol. 64, no. 10, pp. 4071-4077, 2017

[8] W. Chris Chen, Clark Hu, K. C. Ting, Vincent Wei, T. H. Yu, S. Y. Huang, V. C. Y. Chang, C. T. Wang, S. Y. Hou, C. H. Wu and Doug Yu, "Wafer Level Integration of an Advanced Logic-Memory System Through 2nd Generation CoWoS® Technology", 2017 Symposium on VLSI Technology, Kyoto, Japan, 2019