The Present Situation, Challenges and Development of CMOS 3D Stacking Technology

Yijun Hao *

School of Microelectronics, Shanghai University, Shanghai 201800, China

* Corresponding Author Email: haoyijun@shu.edu.cn

Abstract. CMOS 3D stacking technology, as a key path to breaking through the bottleneck of Moore's Law, solves the physical limitation and interconnect delay issues faced by traditional two-dimensional scaling through expansion in the vertical dimension. This article systematically reviews the latest progress of this technology in the three pivotal trends of interconnects, thermal management, and architecture design. In terms of interconnect technology, the emphasis is on bumpless hybrid bonding, which is a process of achieving sub-micron high-density vertical interconnects reliably through the simultaneous bonding of Cu-Cu and SiO_2. In terms of thermal management, this article involves modeling optimization based on equivalent thermal conductivity and embedded microchannel cooling technologies, displaying a paradigm shift from passive conduction to active fluid-based heat radiation. At the architectural level, the paper analyzes strategies for the co-optimization of heterogeneous integration and Electronic Design Automation (EDA), especially the potential of software-defined near-memory computing architectures to break the “memory wall”. Research indicates that the paradox problem between heat and electricity in high-density integration can be effectively solved thanks to the innovation across the "material-process-architecture" nexus, paving the way for a theoretical reference for process route selection in next-generation high-performance chips.

Keywords: CMOS 3D stacking technology, Bumpless hybrid bonding, Thermal management, Heterogeneous integration.

1. Introduction

As the gate length of Metal Oxide Semiconductor Field Effect Transistors (MOSFET) approaches atomic length, the traditional two-dimensional scaling strategies have faced numerous dilemmas, including increasing quantum tunneling leakage currents, higher manufacturing costs, interconnect latency and so on. One of the main challenges of relying only on lithography resolution enhancement to meet Moore's Law is its production and economic feasibility, which motivates the industry to pursue new paradigm innovations in chip design. In this article, 3D integration technology that breaks through the physical boundary through vertical expansion has been considered as a potential direction for improving semiconductor performance [1-4].

Existing work on 3D integrating technology has focused mainly on three major topics: (i) bonding and interconnect technologies such as bumpless hybrid bonding and reliability optimization for Through-Silicon Via (TSV); (ii) thermal management to address heat radiation of multi-layer stacking; and (iii) architecture and craft integration to jointly optimize low-temperature production and interlayer alignment accuracy. While much work in this article has focused on single technical points, there has been no systematic and comprehensive analysis on the electrical performance and thermo-mechanical properties of different stacking schemes while other details including the mechanisms involved such as cumulative thermal budget control in multi-layers stacking, interlayer error propagation, and long-term reliability evolution remain lacking.

In this article, different stacking and interconnection technologies are compared by means of multidimensional study and organization. The technical benefits and application fit of different technical routes are further reviewed. Finally, the key challenges limiting stacking layers and information regarding process route selection and reliability of the increasingly prominent CMOS 3D stacking technology are identified. Therefore, this article provides theoretical and technical context to the next generation architecture design.

2. Theoretical basis analysis

2.1. Critical fabrication processes of CMOS 3D stacking technology

The edge of CMOS 3D Stacking Technology is the vertical integration of multi-layer wafers which is beyond the physical limitation of 2D craftsmanship. The typical technological process can be broken down into four steps: thinning and opening, Firing vertical interconnects (TSV/RDL), chemical mechanical polishing (CMP), and bonding annealing. A simple diagram of the process is shown in Figure 1. The difficulty is to control thermal budget and mechanical stress, instead of degrading devices during multiple stacking processes.

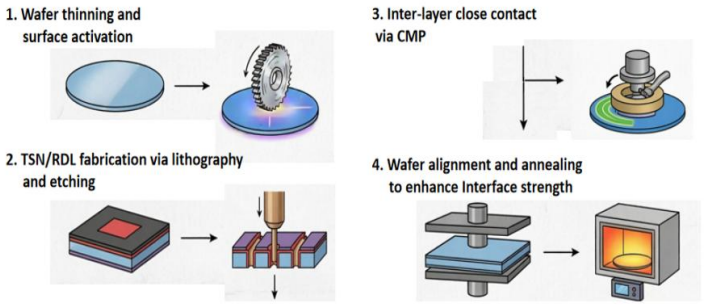

Fig 1. CMOS 3D stacking typical technological process (Picture credit: Original).

2.2. Mechanism of the critical technology

2.2.1. Bumpless hybrid bonding

Combining Cu-Cu metal bond and SiO_2-SiO_2 dielectric bond,bumpless hybrid bonding provides interlayer electrical and mechanical connections with no micro-bumps. The mechanism is that silanol groups on SiO_2 surface are created by plasma at normal temperatures to make physical adsorption. Cu atoms merge during annealing as grain boundary merging and SiO_2 condenses to form covalent bonds [5,6]. This technique strongly controls Cu pads surface composition by chemical-mechanical polishing so that they are slightly raised rather than recessed, allowing atomic contact at the bonding interface and finally sub-micron interconnect pitch and ultra-high connection density.

As is shown in Figure 2, the top image follows the traditional TSV scheme where the through-silicon via needs to vertically pass through the KOZ across the top and bottom silicon substrates, occupying chip area and limiting routing density; the bottom image uses bump-free Cu-co hybrid bonding, achieving interlayer connections by face to face direct bonding which can effectively eliminate the area overhead caused by the KOZ, and can bring higher interconnect density and integration efficiency.

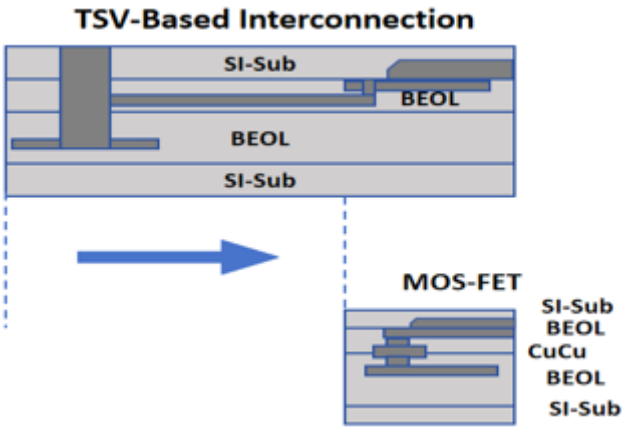

Fig 2. Comparison of structures of TSV and hybrid bonding interconnect schemes (Picture credit: Original).

2.2.2. Sub heading

Three-dimensional stacking causes exponential power densities and thermal management is a limiting factor. By means of a good medium theory and finite element analysis, complex heterogeneous structures containing TSV can be viewed as materials with anisotropic thermal conductivity that leads to much reduced computations [2][7]. The thermal conductivity of TSVs is strongly directional: the in-plane thermal conductive is driven by the thickness of SiO_2 insulation, and the in-plane conductivity is driven mainly by Cu fill ratio and TSV pitch [1][4]. In order to effectively decrease junction temperature and prevent interfacial delamination due to thermal stress, thermal management also involves micro channel forced convection cooling and high-thermal-conductivity underfill.

An equivalent thermal conductivity model, as shown in Figure 3, is constructed based on the finite element analysis method and includes two orthogonal sectional views: Section 1 is a longitudinal section along the axial direction of the vertical via, and Section 2 is a transverse section perpendicular to the via axis. In the model, the yellow cylinders represent copper-filled vertical interconnect vias, and the gray areas represent the silicon dioxide insulating substrate. Thermal boundary conditions are established with isothermal boundaries and uniform heat flux applied to the top and bottom surfaces, while adiabatic conditions are imposed on the lateral and front/back surfaces to ensure unidirectional heat conduction. The same thermal conductivity in the in-plane and cross-plane directions are calculated, providing essential equivalent material parameters for thermal simulation.

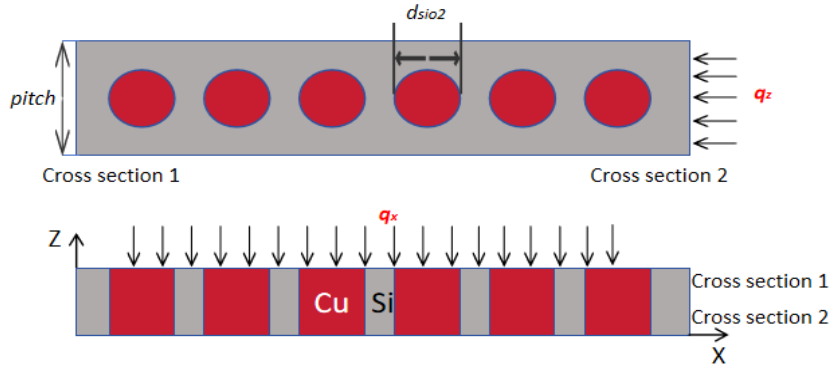

Fig 3. Geometry for equivalent thermal conductivity extraction (Picture credit: Original).

2.2.3. Architecture and EDA design

Heterogeneous monolithic 3D integration achieves an optimal trade-off among power, performance, area, and cost by employing different process nodes across different layers[3].Its key designing methods include: a timing-aware partitioning algorithm that allocates critical path units to high-speed layers to optimize delay; three-dimensional clock tree synthesis that reduces clock skew through cross-layer clock buffers; and voltage domain co-management that avoids the overhead of level shifters by limiting inter-layer voltage differences. The EDA process usually has pseudo 3D approach: mapping in the inter-layer by partitioning and resartition after the optimization is completed in 2D, as the integration size increases the complexity of interconnection networks. To solve the convergence problems of traditional algorithms in thermo-electric design, this direction also proposes machine learning in order to utilize its powerful data processing power.

3. Key technology case analysis

3.1. Bumpless hybrid bonding

Based on the advanced Cu-Cu hybrid bonding technology reported by the Kagawa research team, Sony Corporation has successfully achieved mass production of back-illuminated CMOS image sensors. This technology established a standard process of trench preparation, barrier layer PVD, Cu

electroplating, deposition annealing, CMP planarization, wafer bonding, and bonding annealing in order. Among these, the controlled CMP process is a key technology; by optimizing the polishing slurry and pressure parameters, the Cu pads maintain nanometre level protrusions, ensuring Cu-Cu preferential contact during bonding and eliminating interface voids [5].

However, simply increasing interconnect density can no longer address power integrity issues in high-density stacking. In 2022, L. Zhu's team at Georgia Tech revealed the challenges and opportunities of hybrid bonding technology in power delivery networks. By comparing micro-bump and hybrid bonding, the research observed that in addition to better bandwidth, hybrid bonding will also cause IR drop risks. Zhu's team considered a co-optimization strategy, taking advantage of the ultra-high density vertical interconnect resources offered by hybrid bonding to physical partition power and signal layers. The research results indicate that by decoupling the power delivery paths from the high-density logic layers, not only can IR drop be effectively reduced, but the chip's PPA metrics can also be significantly optimized [8].

This process represents the evolution of hybrid bonds from simple high density interconnection to joint design. The ultimate value of this process is that it applies process physical features to solve system-level electrical problems and gets functional integration from the connections to power supply.

3.2. Thermal management

Xiao's research team proposed an efficient numerical method for thermal management based on finite element analysis, targeting 3D stacked devices such as high-bandwidth memory, and systematically analyzed the influence of TSV geometric parameters on thermal transport characteristics. The study established an anisotropic thermal conductivity model, simplifying the complex three-dimensional structure containing periodic TSV arrays into a homogeneous equivalent material, with computational efficiency increasing threefold compared to detailed 3D models. On this basis, regarding material-level optimization, experiments showed that improving the thermal conductivity of the bottom filling material—for example, increasing it from 0.2 W/(m·K) to 4 W/(m·K) in experiments—can reduce the junction temperature of logic chips by more than 2°C, though the marginal benefit diminishes when thermal conductivity exceeds 2 W/(m·K) [4].

Although the optimization of TSV can improve heat radiation, solely relying on material thermal conductivity is approaching the physical limit at ultra-high power densities. The “Embedded Reconfigurable Cooling Method” proposed by the teams of N. Zhang and Y. Ye at the Institute of Microelectronics, Chinese Academy of Sciences (IMECAS), marks a significant advancement in thermal management technology. This technology directly embeds micro-channel structures into the chip stacking layers, using fluid convection to replace traditional heat conduction. Unlike conventional approaches, this method is configurable and replaceable, allowing the cooling channel layout to be flexibly adjusted according to the dynamic distribution of hot spots within the chip. This embedded cooling solution not only significantly enhances heat transfer efficiency but also addresses the thermal accumulation bottleneck of high-power logic chips in three-dimensional stacking [9].

Moving from passive heat conduction to active heat dissipation is an extremely successful and remarkable achievement. What is unique is that this approach solves package thermal problems at chip level by fluid dynamics. This leads to a greater complexity of process, but it is an important choice that goes over the barriers to high-power-density 3D integration.

3.3. Architecture and EDA design

Pentapati and Lim proposed a complete design flow for heterogeneous monolithic 3D IC, achieving comprehensive optimization in power, performance, area, and cost through process heterogeneity. The study is based on the Pin-3D enhanced flow and develops an EDA solution that supports heterogeneous integration. The core innovation lies in the timing-aware partitioning algorithm: after initial optimization is completed using a single-process node during the pseudo-3D stage, intelligent partitioning is performed based on the timing criticality of the cells, assigning critical path cells to the high-speed layer and non-critical cells to the low-speed, low-power layer. This

strategy reduces the length of critical paths, with 75% of the cells placed in the high-speed layer, thereby lowering power consumption while maintaining performance [10].

But with the increasing need for computing power by AI, traditional EDA partitioning has struggled to overcome the limitations of the 'memory wall'. The team of Xu A. and Deng C.at Tsinghua University proposed in 2025 the 'Software-defined process-near-memory architecture using 3D hybrid bonding integration' which brings system-level innovation to architecture design. This study makes use of 3D hybrid bonding to achieve the ultra-high-density vertical stacking of logic processors and memory, thereby establishing a software-defined Processing-Near-Memory (PNM) architecture. Distinct from conventional heterogeneous EDA optimization, this architecture exploits software-defined flexibility to overcome the I/O bandwidth bottlenecks inherent in traditional packaging, facilitating ultra-low latency data transfer between memory and processors [11]. An advantage of this design is that it integrates the physical limits (such as the high density links enabled by 3D hybrid bonding) with software, addressing the storage wall problem at the physical level and also being a useful practical example for guiding the future chip models from purely hardware based to hardware software co-optimisation.

3.4. Comparison of current technology

Table 1. Comparison of the merits and demerits of Three Technical Directions.

Comparison object	Bumpless Hybrid Bonding	Thermal management	Architecture and EDA Design
Core objective	Ultra-fine pitch high-density vertical interconnect	Optimize TSV array thermal transfer characteristics	Optimize system-level PPAC to achieve cross-layer timing convergence
Key technology	Cu-Cu metal bonding SiO_2 dielectric bonding	Equivalent Anisotropic Thermal Conductivity Model TSV Geometric Parameter Optimization	Timing-Aware Partitioning Algorithm Voltage Domain Collaborative Management
Performance indicators	Contact resistance ~0.5 Ω TDDB lifetime >10 years	Equivalent thermal conductivity error <2.5% Junction temperature reduced by 2-5°C	PPC increased by 20-30% PDP decreased by about 16%
Main challenge	Interface voids Submicron-level alignment difficulty	Thin chip hot-spot effect Micro-channel leakage risk	Cross-layer clock skew control ML model training is complex
Typical application	Back-illuminated CMOS image sensor	High-power logic chip	Mobile processor

It is likely that future CMOS 3-D stacking will follow three clear technologies (one-to-one as previously discussed in table 1). First of all, submicron or even nano-level hybrid bonding technology will push density toward monolithic integration; secondly, embedded micro-channels and phase change materials will become active-passive collaborative thermal management solution; thirdly, AI-driven intelligent partitioning algorithms and 3D physical design technology will be employed to automatically realize heterogeneous integration.

4. Conclusion

In this article, it reviews three fundamental aspects of CMOS 3D stacking: bumpless hybrid bonding, thermal management and architectural design optimization. Research founds that the accurate control of the sub-micron process window helped us overcome the electromigration and interfacial void problem, and it provides a solid process foundation for high-density vertical interconnects. In terms of thermal management, active cooling techniques such as micro-channel cooling can transition from passive material-based heat dissipation to structural active cooling in

order to overcome thermal resistance bottlenecks in high-power-density chips. In the case of architecture and EDA, machine learning-based thermal-aware placement methods overcome the convergence problem encountered by traditional algorithms for large-scale thermal-electrical co-design problems, thus initiating heterogeneous integration into intelligent design. Nonetheless, the current technology has multiple challenges. In the future, with the development of atomic layer deposition and deep integration of machine learning algorithms, 3D integration will move towards a system-level architecture that incorporates thermal, electrical and mechanical multi-physical fields together, providing some key momentum for integrated circuits after Moore era.

References

[1] L. Ma, M. Xiang, T. Wu, Development and Challenges of Three-Dimensional Heterogeneous Integration. Electron. Packag. 24(06), 125-132 (2024)

[2] S. Zhang, Z. Li, P. He, Research Progress on 3D Heterogeneous and Heterostructured Integration of Microsystems. Electron. Packag. 21(10), 78-88 (2021)

[3] Y. Yu, Research Progress on Silicon-Based Heterogeneous Three-Dimensional Integration Technology. Res. Prog. Solid State Electron. 41(01), 1-9 (2021)

[4] W.-W. Shen, K.-N. Chen, Three-Dimensional Integrated Circuit (3D IC) Key Technology: Through-Silicon Via (TSV). Nanoscale Res. Lett. 12, 56 (2017)

[5] Y. Kagawa, N. Fujii, K. Aoyagi, et al., Novel stacked CMOS image sensor with advanced Cu-Cu hybrid bonding. Proc. IEEE IEDM, 8.4.1-8.4.4 (2016)

[6] Y. Chen, K. Yue, F. Lü, et al., Progress in Heterogeneous Integration Packaging Technology of Integrated Circuits. Electron. Packag. 24(09), 51-62 (2024)

[7] C. Xiao, H. He, J. Li, et al., An effective and efficient numerical method for thermal management in 3D stacked integrated circuits. Appl. Therm. Eng. 121, 200-209 (2017)

[8] L. Zhu, C. Jo, S. K. Lim, Power Delivery Solutions and PPA Impacts in Micro-Bump and Hybrid-Bonding 3D ICs. IEEE Trans. Compon. Packag. Manuf. Technol. 12(12), 1969-1982 (2022)

[9] N. Zhang, B. Jiao, Y. Ye, et al., Embedded cooling method with configurability and replaceability for multi-chip electronic devices. Energy Convers. Manag. 253, 115160 (2022)

[10] S. Pentapati, S. K. Lim, Heterogeneous monolithic 3-D IC designs: Challenges, EDA solutions, and power, performance, cost tradeoffs. IEEE Trans. VLSI Syst. 32(3), 413-421 (2024)

[11] A. Xu, C. Deng, J. Zhu, et al., Software-defined process-near-memory architecture using 3D hybrid bonding integration. Sci. China Inf. Sci. 68, 112402 (2025)