INVITED: Design Automation Needs for Monolithic 3D ICs: Accomplishments and Gaps

Lingjun Zhu and Sung Kyu Lim
School of Electrical and Computer Engineering
Georgia Institute of Technology

2023 60th ACM/IEEE Design Automation Conference (DAC) | 979-8-3503-2348-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/DAC56929.2023.10247666

***Abstract*—In this paper, we provide an overview of design automation tools and methodology for Monolithic 3D ICs, focusing on the accomplishments in recent years and the gaps that remain to be filled. Monolithic 3D integration is an emerging technology with high 3D interconnect density and performance benefits, but it proposes new challenges for computer-aided design tools. In this paper, we first revisit the current status of design automation tools for Monolithic 3D and highlight the recent developments in tier partitioning, 3D placement and routing, inter-tier via controls, and power and thermal integrity analysis. Then, we discuss the gaps to be met for next-generation system-level heterogeneous Monolithic 3D IC design. Finally, we present our vision for the future of design automation developments for Monolithic 3D ICs.**

I. Introduction

In recent years, high-density 3D integration technologies, such as Face-to-Face (F2F) hybrid bonding 3D and Monolithic 3D (M3D) [1], have demonstrated great promises in enabling high-performance and low-power computing systems. These 3D ICs can overcome the critical bottlenecks that limit the area scaling and timing improvement in conventional 2D ICs.

Among them, M3D provides the smallest 3D contact pitch and the highest inter-tier interconnect density [1], representing the future trend for high-density 3D integration. In M3D ICs, multiple device tiers can be fabricated sequentially and integrated in the vertical direction. These tiers are bonded in a face-to-back fashion, which means the Back End Of Line (BEOL) layers of the bottom tier are attached to the substrate of the top tier, and Monolithic Inter-tier Vias (MIVs) are inserted to connect the two tiers. Fig. 1 shows the cross-section view of a two-tier M3D IC.

MIVs are small vertical vias that penetrate through the silicon substrate of the top tier for 3D interconnects. Thanks to the sequential fabrication process, the transistors of the two tiers can achieve an excellent alignment, which leads to extremely small MIV pitches and high inter-tier interconnect density.

The conventional 2D Electrical Design Automation (EDA) tools do not work for M3D ICs since they do not consider multiple device tiers. Several commercial EDA tools have recently started supporting 3D IC designs, including Synopsys' 3DIC Compiler, Cadence's Integrity 3D-IC, and Siemens' Calibre. However, these tools mainly focus on Through Silicon Via (TSV) or micro-bump-based 3D ICs with relatively low 3D interconnect density. They generally perform placement and routing in a tier-by-tier manner, which cannot optimize a large number of MIV locations or consider the tight inter-tier coupling in M3D ICs. Therefore, more EDA solutions special for M3D ICs are needed to address these challenges.

In this paper, we first revisit the current status of M3D EDA development in literature surveys. Then, we summarize the recent accomplishments of EDA to address the M3D design challenges. Furthermore, we look into the future of M3D EDA and discuss the gaps to be met for the next-generation M3D ICs with advanced technology and system-level heterogeneous integration, etc.

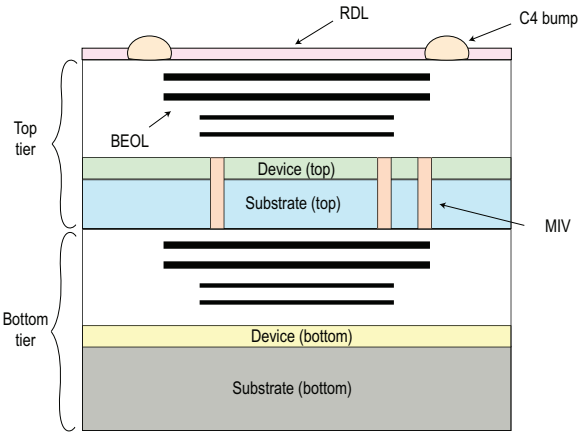

Fig. 1. Cross-section view of a M3D IC.

II. Prior Works

In this section, we first review the existing papers and surveys on M3D design automation tools and Power, Performance, and Area (PPA) studies in the literature.

Many works have been done before to enable various physical design and verification procedures in M3D or more generic high-density 3D ICs, including tier partitioning [2], analytical placement [3, 4], heterogeneous integration [5, 6], reliability analysis [7], etc. Previous surveys also summarize the EDA development for generic 3D ICs, including [8, 9]. However, more studies with a focus on EDA for M3D ICs are needed.

The authors of [10] summarize a set of *pseudo-3D* design flows for M3D ICs, including Cascade-2D, Shrunk-2D, and Compact-2D. These pseudo-3D flows use the existing 2D commercial EDA software to design 3D ICs by modifying the technology files and projecting the 3D layout to 2D. The authors point out that the true 3D placer and router are not yet ready for M3D ICs. Also, they provided a quantitative comparison of the pseudo-3D flows, and their results show that Shrunk-2D and Compact-2D can achieve significant wirelength and power reduction. However, the authors have not discussed the performance degradation and drawbacks of the Shrunk-2D method in this paper.

In [11], the authors provide a comprehensive summary of the M3D physical design and testing solutions. They discuss the pros and cons of existing tier partitioning, 3D placement, clock delivery network, and thermal analysis approaches. In addition, they summarized the solutions for M3D testing, including MIV delay models, test pattern generation, and Built-in Self-Test (BIST), etc. However, the authors did not include any new accomplishments after 2020 in this paper.

The authors of [12] compare the micro-bump, hybrid-bonding, and M3D ICs in terms of PPA. They design and implement a RISC-V-

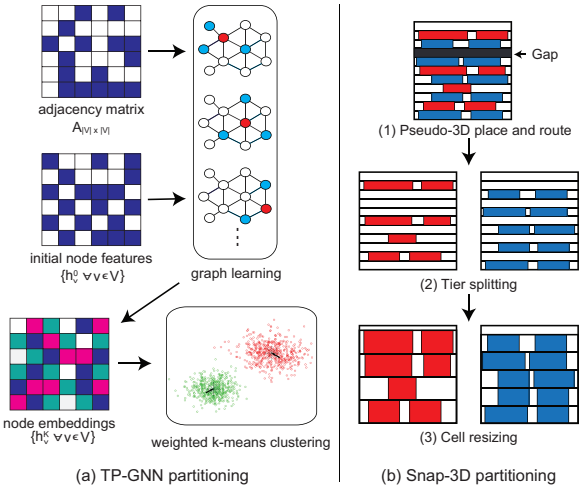

Fig. 2. Partitioning methods: (a) TP-GNN [11], (b) Snap-3D [15].

based CPU core as the design benchmark. Their results demonstrate that the M3D and hybrid-bonding 3D ICs provide more than 74% performance boost and 2.7% energy saving compared with the 2D baseline. These benefits triumph over the micro-bump 3D results thanks to the high interconnect density and low overhead of M3D. On the other hand, this paper focuses on the PPA comparison of M3D and other 3D integration technologies. Still, it does not discuss the unique design automation challenges for M3D ICs.

III. Recent Accomplishments

A. Tier Partitioning

Tier partitioning is a key step for M3D design automation. It determines which tier each instance is placed in and thus significantly impacts the final design quality and PPA.

Previous works mainly use conventional heuristic graph partitioning algorithms for tier partitioning, such as the recursive partitioning using Z-cut [13] and bin-based min-cut partitioning using Fiduccia–Mattheyses (FM) algorithm from Shrunk-2D [14]. These partitioning algorithms consider basic constraints such as area balance and iteratively optimize the cut size (3D inter-tier net number). However, they do not consider other important factors, such as the MIV density, cell timing, design hierarchy, etc., and may lead to sub-optimal partitioning results.

TP-GNN [16] is a Graph Neural Network (GNN)-based tier partitioning framework that considers multiple design and technological factors, including design hierarchy, cell timing, and clock information, etc. These features are encoded into feature vectors and fed into a GNN to represent an initial 2D IC for partitioning. Then, we perform graph contraction and feature aggregation to preserve the hierarchy information and reduce the graph complexity. After that, we perform unsupervised GNN learning to optimize the loss function based on the cross-entropy between neighboring nodes. Finally, we perform the weighted K-means clustering to generate the tier partitioning solution. This process is demonstrated in Fig. 2 (a).

Experimental results in [16] show that the TP-GNN framework can further reduce the total wirelength in 3D ICs by 7% compared with the bin-based min-cut method, which leads to better performance and power in two RISC-V-based CPUs and five other benchmark designs.

Snap-3D [15] provides another strategy to perform tier partitioning based on the 2D placement solution generated by commercial tools. The main idea is to shrink the height of all standard cells by half and place the cells on a 2D plane first (pseudo-3D stage). Then, we perform tier partitioning by splitting the design by rows (e.g., instances on the odd rows go to the top tier, and even rows go to the bottom tier) and rescale the cells to the original size. Fig. 2 (b) shows the pseudo-3D stage and tier partitioning procedures of Snap-3D.

B. 3D Placement and Routing

After tier partitioning, placement and routing become new challenges for M3D IC design. M3D adds a new dimension and more routing space to the 2D problems, making the development of true 3D placers and routers difficult. Luckily, pseudo-3D design flows provide a reliable solution to build 3D ICs with existing 2D tools. However, researchers have found that the previous pseudo-3D flows, such as Shrunk-2D and Compact-2D, can cause performance degradation, design rule violations, and under buffering [17, 18]. The reason is that they shrink the standard cells or RC parasitics and perform tier-by-tier routing after partitioning, which leads to RC estimation errors and sub-optimal 3D routing. Recent works have proposed new pseudo-3D methods to tackle these issues, as shown in Table I.

Macro-3D is a physical design methodology that overcomes the limitations of Shrunk-2D and Compact-2D. It aims at the 3D ICs with standard cells only on one tier and memory blocks mainly on the other tier (logic-on-memory stacking). The main idea is to create a *double BEOL stack* containing the metal layers from both tiers and then project the pins of the pre-placed memory blocks to the exact locations on the stack based on the floorplan. Then, standard cell placement and routing can be done with 2D tools with regard to the projected pins. The original Macro-3D flow targets F2F-bonded memory-on-logic 3D ICs [17], but have been extended to M3D logic-on-memory 3D ICs with power and thermal integrity analysis. Macro-3D achieves more than 21% performance improvement in CPU designs compared with the 2D IC baseline [19]. M3D-ADTCO is a similar flow enabling M3D logic-on-memory stacking based on a customized M3D Process Design Kit (PDK) [20], which also supports Design for Testing (DfT) architecture and sign-off verifications.

To design more general logic-on-logic stacked M3D ICs (with standard cells on both tiers), Pin-3D is proposed for placement and routing optimization. The flow also adopts a double metal stack. When optimizing one tier, it uses *transparent cells* (cells that occupy no silicon area) to represent the cells from the other tier and projects the transparent cell pins to the metal stack. By doing this, the flow optimizes each tier iteratively with 2D tools until timing closure. Results show that the Pin-3D flow achieves better performance and lower power than the previous Compact-2D flow. However, one limitation of Pin-3D is that it still relies on the tier partitioning results to start the optimization.

Snap-3D [15] is a new flow to build logic-on-logic 3D ICs. It uses half-height cells and can automatically partition the design using row splitting, as discussed in Section III-A. Hier-3D [21] further improves the silicon area utilization based on Macro-3D by allowing standard cells to be placed in the empty space on the memory tier according to the design hierarchy. These two flows are developed for F2F-bonded 3D ICs but can be migrated to M3D ICs with minor modifications.

C. MIV Usage and Metal Layer Sharing

During routing, the extremely high-density 3D interconnects are a unique feature offered by M3D technology. However, it is also challenging to understand and control the MIV usage to achieve the desired PPA benefits. Recent works have provided new insights on this considering *metal layer sharing*.

TABLE I
RECENT WORKS ON M3D PLACEMENT AND ROUTING.

	3D tech	Stacking	Key idea	Strength
Macro-3D [17, 19]	M3D	Logic-on-memory	Double metal stack and memory pin projection	Power and thermal integrity
M3D-ADTCO [20]	M3D	Logic-on-memory	M3D PDKit generation	M3D DfT and sign-off verification
Pin-3D [18]	M3D	Logic-on-logic	Transparent cells and cell pin projection	True-3D routing and optimization
Snap-3D [15]	F2F 3D	Logic-on-logic	Half-height cell and cell resizing	Automatic tier partitioning
Hier-3D [21]	F2F 3D	Logic-on-logic	Hierarchy scheme for 3D	Silicon area utilization

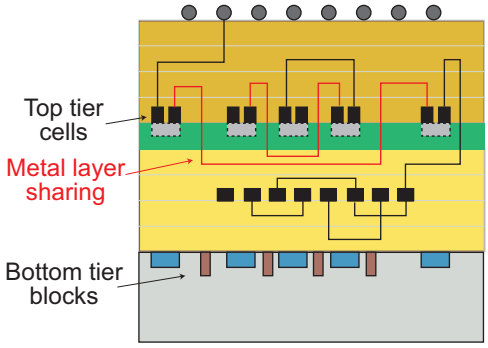

Fig. 3. Metal layer sharing for signal routing in a M3D IC.

In [19], the authors build M3D ICs using the Macro-3D flow and achieve a high MIV usage (more than 500k). They find that the MIVs are used not only for connecting the instances on different tiers but also for optimizing same-tier routing, as shown in Fig. 3. The reason is that the router uses the metal tracks on the relatively empty bottom tier to mitigate the congestion on the top tier. As a result, the router uses many MIVs to share the metal layers between the tiers, leading to improved routing quality and lower wire delays. This observation is called *metal layer sharing*.

The authors of [22] provide a more detailed analysis of metal layer sharing and its impacts. They classify the signal nets into 2D nets (same-tier interconnects) and 3D nets (inter-tier interconnects) and carefully control the amount of metal layer sharing allowed for each class. Their results show that the optimized amount of metal layer sharing can improve the Power Delay Product (PDP) by up to 7% and even allow us to reduce one metal layer in the M3D stack to save 9% of the fabrication cost at the same time.

D. Power and Thermal Integrity Analysis

Power and thermal integrity is another critical issue for 3D integration since 3D ICs tend to suffer from higher power density, longer power delivery path, and lower heat dissipation rate than 2D ICs. Recent works have proposed a new approach to quantify and mitigate the power delivery and thermal issues in M3D ICs.

The authors of [23] present a system-level M3D Power Delivery Network (PDN) model and analysis flow. They first build a M3D IC with PDN using the Shrunk-2D flow and then perform workload-based static and dynamic analysis to quantify the worst-case voltage drop. Their results show that the static IR drop in M3D ICs is 2×higher than 2D due to longer resistive paths, irregular power MIV placement, and fewer power C4 bumps. However, the M3D ICs show lower dynamic voltage drop because they can use decaps from both tiers and thus have improved resiliency against the transient noise.

Based on that, the authors of [19] further improve the PDN for logic-on-memory stacked M3D ICs and perform thermal analysis. They leverage the power density imbalance between the logic tier and memory tier, put the logic tier closer to the power source, and use denser power rails on the logic tier to reduce the IR drop. This structure overcomes the bottlenecks identified in [23] and leads to a minimal IR drop increase compared with 2D ICs. The authors also propose a complete 3D thermal model to perform thermal analysis for M3D ICs. Their results show that the maximum junction temperature of the proposed M3D ICs with higher frequency is only 13°C higher than the 2D ICs. These results demonstrate the power delivery and thermal benefits of logic-on-memory stacking in M3D ICs.

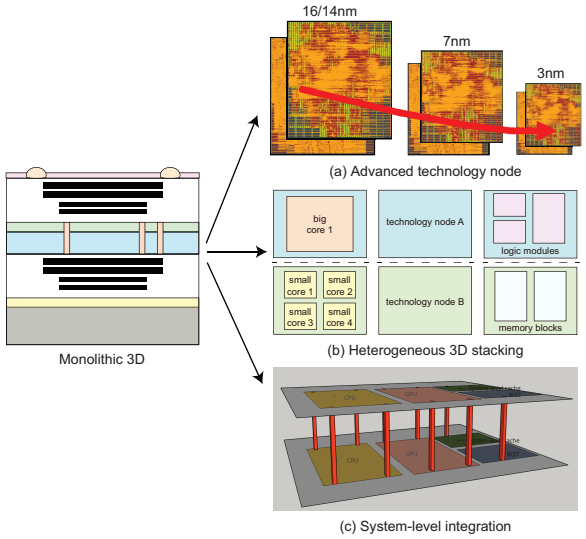

Fig. 4. Gaps for M3D EDA development.

IV. REMAINING GAPS

Based on the existing studies, we identify a few gaps to be filled in the design automation for M3D ICs, as summarized in Fig. 4.

A. Advanced Technology Nodes

Previous studies on M3D EDA are mainly done based on 28nm, 16nm, or 14nm technology nodes. However, more challenges have been introduced as technology scaling continues to 3nm and beyond.

Power scaling is slowing down in the advanced technology nodes, which leads to higher power density and even more challenges for power saving, power delivery, and thermal management in M3D ICs. One previous study [24] has quantified the potential power benefits of M3D ICs in 28nm and 16/14nm and has tried to project the results based on a predictive 7nm PDK. But it has not discussed the power and thermal integrity issues. Therefore, we need to investigate further the power and thermal problems and solutions for M3D ICs in advanced technology nodes.

In addition, M3D is known to suffer from tier degradation issues due to the sequential fabrication process. The authors of [25] investigate the 45nm, 22nm, and 10nm technology nodes. They point out that the transistor degradation is mitigated at the advanced nodes, but

the interconnect degradation worsens. 3nm and beyond will further exacerbate the interconnect degradation issue and thus require more advanced EDA solutions.

B. Heterogeneous 3D Stacking

M3D technology has enabled heterogeneous 3D stacking at different levels, including architectural, technology, and block levels. Previous studies have exploited block-level heterogeneity by partitioning logic modules and memory blocks into different tiers. Some existing EDA flows, such as Pin-3D and Snap-3D, can also be applied to heterogeneous 3D ICs with varying technology nodes. However, heterogeneous stacking has proposed new problems for design automation.

For example, tier partitioning becomes more complicated with heterogeneous technology on different tiers, as partitioning results also determine the technology node of each instance. Therefore, more technological features, such as leakage power, load capacitance, etc., should be incorporated into the tier partitioning framework to optimize the results. Similarly, 3D routing becomes more challenging with heterogeneous technology since the router should consider the significant variance in the interconnect properties at different tiers. Therefore, more design automation algorithms and solutions targeting heterogeneous 3D ICs are needed.

C. System-Level Integration

To design next-generation system-level M3D ICs, we also need more collaborations and interdisciplinary studies between computer architecture, physical design, process engineering, and DfT. For example, novel micro-architectural are required to fully utilize the high-density inter-tier communication and 3D stacking of logic, memory, and other components. Advanced DfT methodology, such as BIST for MIVs, are also critical to ensure the reliability of M3D ICs. The future EDA tools should be able to incorporate these system-level designs and DfT solutions into the physical design flow.

V. Conclusion

In this paper, we summarize the recent development in M3D EDA solutions, with a focus on 3D tier partitioning, placement and routing, MIV usage control, in addition to the power and thermal integrity analysis. We highlight the accomplishments of GNN, pseudo-3D flows, and 3D thermal models in these fields. We envision that more studies need to be done in the future to fill the gaps in M3D EDA, especially for advanced technology nodes, heterogeneous 3D stacking, and system-level integration.

References

[1] P. Batude et al. "3D Sequential Integration Opportunities and Technology Optimization". In: *Int. Interconnect Technol. Conf.* 2014.

[2] Kwangsoo Han et al. "Improved Performance of 3DIC Implementations through Inherent Awareness of Mix-and-Match Die Stacking". In: *Des. Autom. Test Eur. Conf. Exhib. DATE*. 2016.

[3] Chau-Chin Huang et al. "NTUplace4dr: A Detailed-Routing-Driven Placer for Mixed-Size Circuit Designs With Technology and Region Constraints". In: *Trans. Comput.-Aided Des. Integr. Circuits Syst.* (2018).

[4] Jingwei Lu et al. "ePlace-3D: Electrostatics Based Placement for 3D-ICs". In: *Proc. Int. Symp. Phys. Des.* 2016.

[5] Montserrat Fernandez-Bolaños et al. "Advanced Sensor Systems by Low-Temperature Heterogeneous 3D Integration Processes". In: *Symp. Des. Test Integr. Packag. MEMS MOEMS DTIP*. 2018.

[6] Wilfred Gomes et al. "8.1 Lakefield and Mobility Compute: A 3D Stacked 10nm and 22FFL Hybrid Processor System in 12×12mm2, 1mm Package-on-Package". In: *Int. Solid- State Circuits Conf. - ISSCC*. 2020.

[7] K-W. Lee et al. "Characterization and Reliability of 3D LSI and SiP". In: *Int. Electron Devices Meet.* 2013.

[8] Paul D. Franzon et al. *Handbook of 3D Integration, Volume 4: Design, Test, and Thermal Management*. 2019. 488 pp.

[9] G. Van der Plas et al. "Design and Technology Solutions for 3D Integrated High Performance Systems". In: *Symp. VLSI Circuits*. 2021.

[10] Heechun Park et al. "Pseudo-3D Approaches for Commercial-Grade RTL-to-GDS Tool Flow Targeting Monolithic 3D ICs". In: *Proc. Int. Symp. Phys. Des.* 2020.

[11] Lingjun Zhu et al. "Design Automation and Test Solutions for Monolithic 3D ICs". In: *J. Emerg. Technol. Comput. Syst.* (2022).

[12] Jinwoo Kim et al. "Micro-Bumping, Hybrid Bonding, or Monolithic? A PPA Study for Heterogeneous 3D IC Options". In: *Des. Autom. Conf. DAC*. 2021.

[13] Mohit Pathak et al. "Through-Silicon-via Management during 3D Physical Design: When to Add and How Many?" In: *Int. Conf. Comput.-Aided Des. ICCAD*. 2010.

[14] Shreepad Panth et al. "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs". In: *Trans. Comput.-Aided Des. Integr. Circuits Syst.* (2015).

[15] Pruek Vanna-Iampikul et al. "Snap-3D: A Constrained Placement-Driven Physical Design Methodology for High Performance 3D ICs". In: *Trans. Comput.-Aided Des. Integr. Circuits Syst.* (2022).

[16] Yi-Chen Lu et al. "TP-GNN: A Graph Neural Network Framework for Tier Partitioning in Monolithic 3D ICs". In: *Des. Autom. Conf. DAC*. 2020.

[17] Lennart Bamberg et al. "Macro-3D: A Physical Design Methodology for Face-to-Face-Stacked Heterogeneous 3D ICs". In: *Des. Autom. Test Eur. Conf. Exhib. DATE*. 2020.

[18] Sai Surya Kiran Pentapati et al. "Pin-3D: A Physical Synthesis and Post-Layout Optimization Flow for Heterogeneous Monolithic 3D ICs". In: *Proc. Int. Conf. Comput.-Aided Des.* 2020.

[19] Lingjun Zhu et al. "High-Performance Logic-on-Memory Monolithic 3-D IC Designs for Arm Cortex-A Processors". In: *Trans. VLSI Syst.* (2021).

[20] Sebastien Thuries et al. "M3D-ADTCO: Monolithic 3D Architecture, Design and Technology Co-Optimization for High Energy Efficient 3D IC". In: *Des. Autom. Test Eur. Conf. Exhib. DATE*. 2020.

[21] Anthony Agnesina et al. "Hier-3D: A Hierarchical Physical Design Methodology for Face-to-Face-Bonded 3D ICs". In: *Int. Symp. Low Power Electron. Des.* 2022.

[22] Sai Pentapati et al. "Metal Layer Sharing: A Routing Optimization Technique for Monolithic 3D ICs". In: *Trans. Very Large Scale Integr. VLSI Syst.* (2022).

[23] Kyungwook Chang et al. "System-Level Power Delivery Network Analysis and Optimization for Monolithic 3-D ICs". In: *Trans. VLSI Syst.* (2019).

[24] Kyungwook Chang et al. "Match-Making for Monolithic 3D IC: Finding the Right Technology Node". In: *Des. Autom. Conf. DAC*. 2016.

[25] Shreepad Panth et al. "Tier Degradation of Monolithic 3-D ICs: A Power Performance Study at Different Technology Nodes". In: *Trans. Comput.-Aided Des. Integr. Circuits Syst.* (2017).