# Snap-3D: A Constrained Placement-Driven Physical Design Methodology for High Performance 3-D ICs

Pruek Vanna-Iampikul<sup>®</sup>[,](https://orcid.org/0000-0002-2897-7142) Chengjia Shao, Yi-Chen L[u](https://orcid.org/0000-0003-1481-9167)<sup>®</sup>, Sa[i](https://orcid.org/0000-0003-3966-1749) Pentapati®, Yun Heo, Jae-Seung Choi, and Sung Kyu Lim, *Senior Member, IEEE*

*Abstract***—3-D integration technology is one of the leading options to advance Moore's Law beyond conventional scaling. One of the 3-D integration choice is the heterogeneous integration with the benefits of power saving over the homogeneous integration. With the lack of commercial 3-D tools, existing 3-D physical design flows utilize 2-D commercial tools to perform 3-D integrated circuit (3-D IC) physical synthesis. Specifically, these flows build 2-D designs first and then convert them into 3-D designs. However, several works demonstrate that design qualities degrade during this 2-D–3-D transformation and some of the flows do not support heterogeneous integration. In this article, we propose Snap-3D, a constraint-driven placement approach to build commercial-quality 3-D ICs, which supports both homogeneous and heterogeneous 3-D ICs. Our key idea is based on the observation that if the standard cell height is contracted and partitioned into multiple tiers, any commercial 2-D placer can place them onto the row structure and naturally achieve high-quality 3-D placement. This methodology is shown to optimize power, performance, and area (PPA) metrics across different tiers simultaneously and minimize the aforementioned design quality loss. Experimental results on seven industrial designs demonstrate that Snap-3D achieves up to 10.9% wirelength, 9% power, and 25% performance improvements compared with state-of-the-art 3-D design flows.**

*Index Terms***—3-D integrated circuits (3-D IC), heterogeneous 3-D ICs, partitioning and floorplanning, physical synthesis, placement, pseudo-3D design flow.**

#### I. INTRODUCTION

**3**-D INTEGRATED circuit (3-D IC) design has demon-<br>strated great potential to meet the current and future needs of the semiconductor industry. They significantly improve design quality over traditional 2-D ICs by die stacking methodologies. Based on different die stacking methodologies, 3-D ICs can be categorized into three main categories: 1) throughsilicon via (TSV) based; 2) monolithic; and 3) face-to-face (F2F) bonded. TSV-based 3-D ICs are studied earlier, however, due to the large pitch and high parasitics of TSVs,

Manuscript received 11 March 2022; revised 14 July 2022 and 12 October 2022; accepted 19 October 2022. Date of publication 2 November 2022; date of current version 20 June 2023. This work was supported in part by the DARPA ERI 3DSOC Program under Award HR001118C0096; in part by the Semiconductor Research Corporation under GRC Task 2929; and in part by Samsung Electronics, Inc. This article was recommended by Associate Editor I. H.-R. Jiang. *(Corresponding author: Pruek Vanna-Iampikul.)*

Pruek Vanna-Iampikul, Chengjia Shao, Yi-Chen Lu, Sai Pentapati, and Sung Kyu Lim are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: v.pruek@gatech.edu).

Yun Heo and Jae-Seung Choi are with Samsung Engineering Company Ltd., Seoul, South Korea.

Digital Object Identifier 10.1109/TCAD.2022.3218763

this stacking method is only useful when the connections between dies are relatively few, such as in memory-on-logic designs. The F2F stacking methodology connects two prefabricated dies in an F2F fashion. Since the intertier vias (F2F vias) do not go through the silicon substrate as in the face-to-back stacking of TSV-based and monolithic 3-D ICs, F2F stacking enables much higher 3-D integration density and is more cost-effective from a manufacturing perspective. Recently, an F2F-bonded 3-D chip Lakefield [\[1\]](#page-4-0) developed by Intel has already been utilized in consumer electronics, which demonstrates the potential of this stacking technology. In this article, we present novel physical design (PD) methodologies to build high-quality F2F-bonded 3-D ICs, which support heterogeneous integration.

<span id="page-0-3"></span><span id="page-0-2"></span><span id="page-0-1"></span><span id="page-0-0"></span>Due to the absence of 3-D commercial PD tools, existing commercial 3-D PD methodologies, such as Shrunk-2D [\[2\]](#page-4-1), Compact-2D [\[3\]](#page-4-2), and Cascade-2D [\[4\]](#page-4-3) leverage 2-D commercial tools to build commercial-quality 3-D ICs. However, a common drawback in these works is that they all build the fullchip design in a sequential manner, which means they build the 3-D chip die-by-die. This sequential PD methodology fails to benefit from the advantages that 3-D technology provides, which inevitably results in suboptimal 3-D ICs. To overcome this limitation, The Snap-3D [\[5\]](#page-4-4) leverages a placement-driven approach to build all the tiers of the 3-D full-chip designs together at once, unlike previous works that stack independently placed and routed (P&R) dies together. However, it does not support the heterogeneity and the final routing are performed in die-by-die.

<span id="page-0-4"></span>This article extends the Snap-3D [\[5\]](#page-4-4) to layouts such as heterogeneous 3-D IC designs, where the technology nodes are different between tiers. In addition, we modified the design methodology to optimize the design performance due to the fact that [\[5\]](#page-4-4) finalize the 3-D routing in die-to-die manner which slightly degrade the performance. Moreover, we added a detailed analysis among 2-D and 3-D designs, including homogeneous and heterogeneous 3-D ICs.

# II. RECENT WORK ON HETEROGENEOUS 3-D ICS

The heterogeneous 3-D IC design is a type of 3-D IC setting where different 3-D dies are fabricated using different technology nodes. The key benefit of heterogeneity is the usage of a high-performance technology node along with a low-power technology node within a single design which leads to designs with the same performance but have a smaller area and lower

1937-4151 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.



<span id="page-1-0"></span>Fig. 1. Snap-3D flow for heterogeneous 3-D IC.



<span id="page-1-1"></span>Fig. 2. Handling partial blockage area in a heterogeneous design. (a) Top die. (b) Bottom die. (c) Flatten die. (d) Final FP.

<span id="page-1-4"></span>power consumption. The related work with heterogeneous 3-D ICs [\[6\]](#page-4-5), which extends the Pin-3D flow [\[7\]](#page-4-6), provides the comparison between heterogeneous and homogeneous 3-D ICs by utilizing the two cell variations in the 28-nm technology node. The experimental results also prove that the power and footprint benefits are better than homogeneous 3-D.

## <span id="page-1-5"></span>III. METHODOLOGY

#### *A. Overview*

We extend [\[5\]](#page-4-4) to support heterogeneous 3-D ICs and enhance the design performance by performing post-route optimization. The Snap-3D flow is shown in Fig. [1.](#page-1-0) Since the heterogeneous 3-D ICs contain different technology nodes between top and bottom dies, we modify our the Pseudo-3D stage to handle different cell sizes, including row planning, technology scaling, and handling memory marcos in Figs. [2](#page-1-1) and [3,](#page-1-2) respectively. Fig. [4](#page-2-0) illustrates the steps prior to F2F design optimization. After the Pseudo-3D stage, we perform post-route optimization to finalize the design.

#### *B. Tier Partitioning*

In the first two steps in Fig. [1,](#page-1-0) we perform 2-D placement and partition the gate-level netlist by applying the placement-driven bin-based FM min-cut algorithm [\[8\]](#page-4-7). For the heterogeneous 3-D IC, we partition the cell such that there are more cells in the die with smaller technology nodes to balance the area.

# *C. Technology Scaling*

We discuss the scaling of the cells and rows for heterogeneous 3-D IC. This work utilizes the two cell variations from a single technology node where the cell's power, performance, and height are different. The two different cell variations create two different SITE rows in the design with unequal row



<span id="page-1-2"></span>Fig. 3. Technology scaling options for heterogeneous 3-D design. (a) Placement row alignment with original size rows. (b) Placement row alignment with 0.6-um height. (c) Row alignment with 0.45-um height with gap rows.

heights. In this study, we utilize the equal footprint for both top and bottom dies, so the die with taller cells would contain fewer rows, while the die with shorter cells has more cell rows. We choose two variations of the 28-nm technology node: 12 and 9 track cells to avoid adding level shifters, which adds additional power. However, our flow supports any variations or even different technology nodes. Compared to 9-track cells, 12-track cells are larger and faster with more power consumption. The height of the 12-track cell is 1.2 um, while the 9-track cell height is 0.9 um. With a given 3-D footprint of 5 um  $\times$  5 um, there are five 9-track rows for the bottom die and four 12-track rows for top die as shown in Fig. [4\(](#page-2-0)a). Since the PnR tool (Cadence Innovus) used in this work does not allow placing rows with different heights in the same core area. Therefore, each technology node is scaled to the same height. In order to fit all the placement rows within the footprint, the scaled height is required to be half of the shorter cell as shown in Fig.  $3(c)$  $3(c)$ . Otherwise, the rows would exceed the footprint as in Fig.  $3(b)$  $3(b)$ .

#### <span id="page-1-3"></span>*D. Placement Row Alignment*

After the target cell height is determined for both rows, we first determine the row pattern for the 12-track and 9-track cells. In the given 3-D footprint (here, 5 um $\times$ 5 um), we generate the rows with the original row height for both cell variations one at a time. Next, we count the number of rows in the footprint and repeat the step for another cell variation. Once we obtain the number of rows for each technology node, we calculate the ratio of cell height between two cell variations which is 4 to 3 for taller cell to shorter cell. Therefore, we set up the row pattern of 7 which contains three 12-track rows and four 9-track rows and align them as in Fig. [4\(](#page-2-0)c). Since we scale 12-track cell height more than half, so we add a gap row to compensate for the over-scaling. Each row of the 12-track cells is over-scaled by 0.15 um, so for every pattern containing three 12-track rows, we insert one empty row. The orientation of the row is alternating between flip and nonflip among the same technology node such that they will be recomposed in abutted row fashion later in the steps shown in Fig.  $4(f)$  $4(f)$ –(h). Next, we repeat the row pattern until it fill the footprint.

#### <span id="page-1-6"></span>*E. Cell Resizing Constraints*

As we have an additional technology node with a slower cell in heterogeneous design, to maintain the cell area balancing Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 14:10:15 UTC from IEEE Xplore. Restrictions apply.

VANNA-IAMPIKUL et al.: SNAP-3D: A CONSTRAINED PLACEMENT-DRIVEN PD METHODOLOGY 2333



<span id="page-2-0"></span>Fig. 4. Different steps of the Pseudo-3D stage for heterogeneous 3-D IC. (a) Determine the number of rows. (b) Cell/row scaling. (c) Row alignment. (d) Pseudo-3D P&R. (e) Tier splitting. (f) Cell shifting. (f) Cell shifting. (g) Cell resizing.

and avoid the tool from adding only the faster technology node, we add another set of cell groups for the slower technology node in addition to Homogeneous 3-D. With the new resizing constraint, it is guaranteed that the cells can only be resized with cells from the same group, and so a cell in slow die cannot be replaced with a cell of faster technology node.

# *F. Handling Memory Macro*

The memory macros are manually preplaced beforehand. Since the original row height between the top and bottom die is different, the placement blockage area needs to be considered carefully to avoid the cell displacement. We propose the rowbased placement blockage projection approach. With a given memory placement for top and bottom die in Fig. [2\(](#page-1-1)a) and (b), memory M1 and M3 are placed on the top die while memory M2 is placed on the bottom die. Next, we identify the available placement area by flattening both die into one footprint as in Fig. [2\(](#page-1-1)c). The yellow area illustrates the full blockage area where both dies have memory macros. The green area denotes the partial blockage area where only one die occupies the area, which allow cell placement on another die. After specifying the blockages area, we mark the row for both top and bottom dies from normal size footprint in Fig. [2\(](#page-1-1)a) and (b). Next, we generate the placement rows as in Section [III-D](#page-1-3) and create the placement blockage for the marked row. In the top die, we create the placement blockages on all rows (1–4), while in the bottom die we create the placement blockages on the first three rows as in Fig.  $2(d)$  $2(d)$ . With this approach, it ensures that the standard cells are placed in the correct row and no displacement occurs.

#### *G. Tier Splitting and Cell Shifting*

After the Pseudo-3D stage, we split the combined placement into two as in Fig. [4\(](#page-2-0)e). Next, we relocate the cells and rows based on their row order. With the row order, we calculate and move the location of the cell based on the original height as in Fig. [4\(](#page-2-0)f). Once cells and rows are in place, we resize the height back to its original height completing the pseudo-3D stage as in Fig.  $4(g)$  $4(g)$ .



<span id="page-2-1"></span>Fig. 5. Post-route optimization steps.

# *H. Full-3D Routing Optimization*

In Snap-3D flow [\[5\]](#page-4-4), we insert F2F pads and perform the routing die-by-die. However, the heterogeneous design causes lots of cell shifting to align them in place before resizing them to their original height. The increased cell shifting could lead to timing degradation due to additional wire delay from routing changes. We propose full-3D routing approach by extending [\[7\]](#page-4-6). First, we change all the 9-track cell's properties into cover cells such that the tool would consider them as hard macro. One of the main reasons we choose a 9-track cell to be the cover cell is because it is more efficient to insert a faster cell to optimize the power when closing the timing. The memory macros are scaled down to the SITE size to allow standard cells to be placed in the partial blockage area. The cell pins are kept to their related die location: 12-track cell has pins on metal layer M12 while 9-track cell has pins on metal layer M1. Next, we parse the 3-D placement and perform global and detail routing with the timing closure stage. Fig. [5](#page-2-1) shows the initial 3-D placement after tier splitting. Next, we parse placement location to a single design. Final designs are shown in Fig. [6.](#page-3-0)

# IV. EXPERIMENTAL RESULTS

# *A. Experimental Setup*

We analyze the impact of our Snp-3D flow by comparing the essential metrics with state-of-the-art designs for both homogeneous and heterogeneous designs. We utilize a commercial 28-nm PDK with two variations: 1) 12-track

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 14:10:15 UTC from IEEE Xplore. Restrictions apply.





<span id="page-3-1"></span>TABLE II

ISO-PERFORMANCE PPA COMPARISON AMONG 12-TRACK HOMOGENEOUS 3-D (12T), 9-TRACK HOMOGENEOUS 3-D (9T), AND HETEROGENEOUS 3-D WITH 12-TRACK AS A TOP DIE AND 9-TRACK AS A BOTTOM DIE (12 + 9T). WNS AND TNS, RESPECTIVELY, DENOTE THE WORST AND TOTAL NEGATIVE SLACK

<span id="page-3-2"></span>



<span id="page-3-0"></span>Fig. 6. GDS layout of TATE circuit with three different designs. (a) Homogeneous 3-D IC with 12-track cells. (b) Homogeneous 3-D IC with 9-track cells. (c) Heterogeneous 3-D IC.

cell and 2) 9-track cell. The 2-D design uses six metal layers while the 3-D ICs contain two dies of six metal layers with hybrid bonding with F2F pad as interdie connection. Therefore, there are two main experiments: 1) homogeneous and 2) heterogeneous analysis.

In the homogeneous experiment, we compare the powerperformance-area (PPA) benefits among 2-D, Shrunk-2D [\[2\]](#page-4-1),

Compact-2D [\[3\]](#page-4-2), Pin-3D [\[7\]](#page-4-6), and our Snap-3D flow. The clock target frequency is set to provide the worst negative slack of 2-D design within 5% for optimal results.

In the heterogeneous experiment, we compare PPA, clock metrics, and full-chip timing among 12-track 3-D IC, 9-track 3-D IC, and heterogeneous 3-D IC with both 12 and 9-track cells. All the 3-D design options are implemented with our Snap-3D flow. The target frequency is set to the maximum clock frequency where the 9-track 3-D IC obtains the WNS since at 2-D maximum clock frequency the 3-D designs implemented with Snap-3D flow obtain zero negative slack.

We leverage six design benchmarks containing three purelogic designs, and three processor designs. The Cortex-A7 and Cortex-A53 numbers are normalized according to our NDA with Arm.

# *B. Homogeneous: PPA Comparison*

From Table [I,](#page-3-1) we analyze the impact of our Snap-3D flow with 2-D IC and State-of-the-art 3-D ICs. We observe that our Snap-3D flow obtains the best power-delay-product in all Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 14:10:15 UTC from IEEE Xplore. Restrictions apply.

TABLE III

<span id="page-4-8"></span>CLOCK METRICS COMPARISON AMONG 12-TRACK HOMOGENEOUS 3-D (12T), 9-TRACK HOMOGENEOUS 3-D(9T), AND HETEROGENEOUS 3-D  $(12 + 9T)$ . NOTE THAT THE NUMBER FOR CORTEX- A7 AND CORTEX-A53 DESIGNS ARE NORMALIZED TO 12T VALUE TO PROTECT THE SENSITIVE INFORMATION DUE TO NDA

		Cortex- $A53*$		
<b>Metrics</b>		12T	9Т	$12 + 9T$
<b>Clock WL</b>	(mm)		0.83	0.92
<b>Clock Net Pwr</b>	(mW)		0.82	0.92
<b>FF</b> Power	(mW)		0.85	0.92
<b>Clock Latency</b>	(p <sub>S</sub> )		1.39	0.98
<b>Clock skew</b>	(ps)		1.98	0.82
<b>Slack</b>	`ps)		36x	

designs except in the TATE circuit where Pin-3D flow obtains the best power-delay product (PDP). The S2D [\[2\]](#page-4-1) and C2D [\[3\]](#page-4-2) provide the wirelength and power saving from the 2-D design but the timing is degraded. Both Pin-3D and our Snap-3D flows improve power and performance from 2-D design. The main differences between Pin-3D and Snap-3D is the Pin-3D design has slightly more power and better wirelength. However, the improvement in wirelength increases the cell count, which contributes to more power consumption since the placement is not performed in both die simultaneously. The performance between Snap-3D and Pin-3D is comparable.

#### *C. Heterogeneous 3-D: PPA Comparison*

From Table [II,](#page-3-2) we analyze the impact of heterogeneous 3-D IC by comparing essential metrics over 12-track homogeneous 3-D and 9-track homogeneous 3-D. We observe that the heterogeneous 3-D obtains the comparable performance as 12-track homogeneous 3-D in all design. Moreover, it also provides better wirelength and power saving over 12-track 3-D due to the combination of fast and slow cell with a smaller footprint. While the 9-track 3-D obtains the best power and clock wirelength, it fails to close the timing, which results in the worst PDP. In summary, the heterogeneous 3-D provide the best performance from the best PDP among all benchmarks.

#### *D. Clock Metric Comparison Among 3-D Designs*

From Table [III,](#page-4-8) we observe the 9-track homogeneous 3-D has the shortest clock WL from the smallest footprint, which results in the lowest in clock and flip-flop power. The heterogeneous 3-D has a medium clock wirelength and clock power from a smaller footprint than 12-track 3-D. For the clock performance, we observe that the 9-track 3-D has the worst clock latency and skew despite the smaller footprint because the commercial tool try to close the timing by inserting buffers. On the other hand, the heterogeneous 3-D obtains better clock latency and clock skew, which implies that the combination of faster and slower cell help close the timing without using only faster cells. In summary, the heterogeneous 3-D obtain the best clock metrics in average.

# *E. Full-Chip Timing Comparison*

This section evaluates the full-chip timing among 2-D and 3-D designs on Cortex-A53 in Table [IV.](#page-4-9) The critical path in 2-D design is from register to memory, while the critical path

<span id="page-4-9"></span>TABLE IV FULL-CHIP TIMING SUMMARY COMPARISON OF CORTEX-A53 CIRCUIT BETWEEN 2-D DESIGN AND 12-TRACK HOMOGENEOUS 3-D (12T), 9-TRACK HOMOGENEOUS 3-D (9T), AND HETEROGENEOUS  $3-D(12+9T)$ 

	2D	12T	9Т	$12 + 9T$
Path type	Memory	Register	Register	Register
Cell delay	High	Low	<b>Best</b>	Slightly low
Wire delay	High	Low	<b>Best</b>	Medium
<b>Clock Latency</b>	High	Low	Medium	Slightly low
<b>Clock Skew</b>	Medium	Low	Slightly low	High
<b>Slack</b>	High	Low	Medium	<b>Best</b>

in 3-D designs is between registers except in heterogeneous 3-D. We observe that the launch and capture clock latencies are higher in 2-D compared to the corresponding 3-D designs due to the larger footprint. Also, the total delay in 3-D designs is better than in 2-D because of the shorter wirelength in data path. Among 3-D designs, 12T 3-D has the best total delay because of faster cells, while the heterogeneous 3-D and 9T 3-D have a slightly higher delay. Next, the critical path slack in 3-D designs is within 5% of the clock period, while the slack in 2-D design is much larger due to the higher total delay and clock latency.

#### V. CONCLUSION

In this article, we presented our PD flow named Snap-3D, which provides a commercial-quality 3-D designs. The Snap-3D flow overcomes the timing degradation and displacement issue in Pseudo-3D stages. The experimental results with industrial designs have shown that the Snap-3D flow significantly improves the design performance with a higher maximum clock frequency. Moreover, the Snap-3D flow supports heterogeneous 3-D and provides comparable performance to homogeneous 3-D with faster technology nodes but obtains power and footprint saving.

#### **REFERENCES**

- <span id="page-4-0"></span>[\[1\]](#page-0-0) S. Khushu and W. Gomes, "Lakefield: Hybrid cores in 3D package," in *Proc. Hot Chips Symp.*, 2019, pp. 1–20.
- <span id="page-4-1"></span>[\[2\]](#page-0-1) S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Shrunk-2D: A physical design methodology to build commercial-quality monolithic 3-D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 36, no. 10, pp. 1716–1724, Oct. 2017.
- <span id="page-4-2"></span>[\[3\]](#page-0-2) B. W. Ku, K. Chang, and S. K. Lim, "Compact-2D: A physical design methodology to build two-tier gate-level 3-D ICs," *IEEE Trans. Comput.- Aided Design Integr. Circuits Syst.*, vol. 39, no. 6, pp. 1151–1164, Jun. 2020.
- <span id="page-4-3"></span>[\[4\]](#page-0-3) K. Chang et al., "Cascade2D: A design-aware partitioning approach to monolithic 3D IC with 2D commercial tools," in *Proc. Int. Conf. Comput.-Aided Design*, 2016, pp. 1–8.
- <span id="page-4-4"></span>[\[5\]](#page-0-4) P. Vanna-Iampikul, C. Shao, Y.-C. Lu, S. Pentapati, and S. K. Lim, "Snap-3D: A constrained placement-driven physical design methodology for face-to-face-bonded 3D ICs," in *Proc. Int. Symp. Phys. Design*, 2021, pp. 39–46. [Online]. Available: https://doi.org/10.1145/3439706.3447049
- <span id="page-4-5"></span>[\[6\]](#page-1-4) S. S. K. Pentapati and S. K. Lim, "Heterogeneous monolithic 3D ICs: EDA solutions, and power, performance, cost tradeoffs," in *Proc. 58th ACM/IEEE Design Autom. Conf. (DAC)*, 2021, pp. 925–930.
- <span id="page-4-6"></span>[\[7\]](#page-1-5) S. S. K. Pentapati, K. Chang, V. Gerousis, R. Sengupta, and S. K. Lim, "Pin-3D: A physical synthesis and post-layout optimization flow for heterogeneous monolithic 3D ICs," in *Proc. 39th Int. Conf. Comput.-Aided Design*, 2020, pp. 1–9.
- <span id="page-4-7"></span>[\[8\]](#page-1-6) S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Placement-driven partitioning for congestion mitigation in monolithic 3D IC designs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 4, pp. 540–553, Apr. 2015.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 10,2024 at 14:10:15 UTC from IEEE Xplore. Restrictions apply.