Snap-3D: A Constrained Placement-Driven Physical Design Methodology for Face-to-Face-Bonded 3D ICs

Pruek Vanna-Iampikul, Chengjia Shao, Yi-Chen Lu, Sai Pentapati, and Sung Kyu Lim
v.pruek@gatech.edu
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

3D integration technology is one of the leading options that can advance Moore's Law beyond conventional scaling. Due to the absence of commercial 3D placers and routers, existing 3D physical design flows rely heavily on 2D commercial tools to handle 3D IC physical synthesis. Specifically, these flows build 2D designs first and then convert them into 3D designs. However, several works demonstrate that design qualities degrade during this 2D-3D transformation. In this paper, we overcome this issue with our Snap-3D, a constraint-driven placement approach to build commercial-quality 3D ICs. Our key idea is based on the observation that if the standard cell height is contracted by one half and partitioned into multiple tiers, any commercial 2D placer can place them onto the row structure and naturally achieve high-quality 3D 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 7 industrial designs demonstrate that Snap-3D achieves up to 5.4% wirelength, 10.1% power, and 92.3% total negative slack improvements compared with state-of-the-art 3D design flows.

CCS CONCEPTS

• **Hardware → 3D integrated circuits**; **Placement**; **Physical synthesis**; **Partitioning and floorplanning**.

ACM Reference Format:
Pruek Vanna-Iampikul, Chengjia Shao, Yi-Chen Lu, Sai Pentapati, and Sung Kyu Lim. 2021. Snap-3D: A Placement-Driven Multi-Die Co-Optimization Physical Design Methodology for Stacking-Based 3D ICs.. In *Proceedings of the 2021 International Symposium on Physical Design (ISPD '21), March 22–24, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3439706.3447049

1 INTRODUCTION

3D Integrated Circuit (3D IC) design has demonstrated great potential to meet the current and future needs of the semiconductor industry. They significantly improve design quality over traditional 2D ICs by die stacking methodologies. Based on different die stacking methodologies, 3D ICs can be categorized into three main categories: (1) through-silicon via (TSV) based, (2) monolithic, and (3) face-to-face (F2F) bonded. TSV-based 3D ICs are studied earlier, however, due to the large pitch and high parasitics of TSVs, 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 pre-fabricated dies in a face-to-face fashion. Since the inter-tier vias (F2F vias) do not go through the silicon substrate as in the face-to-back stacking of TSV-based and monolithic 3D ICs, F2F stacking enables much higher 3D integration density and is more cost-effective from a manufacturing perspective. Recently, a F2F-bonded 3D chip Lakefield [1] developed by Intel has already been utilized in consumer electronics, which demonstrates the potential of this stacking technology. In this paper, we present novel physical design methodologies to build high-quality F2F-bonded 3D ICs.

Due to the absence of 3D commercial physical design (PD) tools, existing commercial 3D PD methodologies such as Shrunk-2D [2], Compact-2D [3], and Cascade-2D [4] leverage 2D commercial tools to build commercial-quality 3D ICs. However, a common drawback in these works is that they all build the full-chip design in a sequential manner, which means they build the 3D chip die-by-die. This sequential PD methodology fails to benefit from the advantages that 3D technology provides, which inevitably results in sub-optimal 3D ICs. In this paper, we aim to overcome this limitation by proposing a novel 3D PD methodology named Snap-3D. We leverage a placement-driven approach to build all the tiers of the 3D full-chip designs together at once, unlike previous works that stack independently placed and routed (P&R) dies together.

In this work, we extend the fundamental idea of using 2D commercial tools to build 3D full-chip designs. However, we substantially improve the final design quality by performing placement in a 3D global manner and by leveraging accurate 3D parasitic models to perform timing and power optimizations. Specifically, the locations of design instances are finalized once we perform placement using the 2D commercial tool, and no partitioning and legalization strategies are required to obtain a legalized 3D design as in previous works [2, 3].

2 RELATED WORKS AND MOTIVATIONS

2.1 TSV-Based 3D Placers

With attempts to build 3D EDA tools, many studies have focused on each step of PD flow. One of the crucial stages which determine the quality of a physical design is placement. In the early works

ISPD '21, March 22–24, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8300-4/21/03…$15.00
https://doi.org/10.1145/3439706.3447049

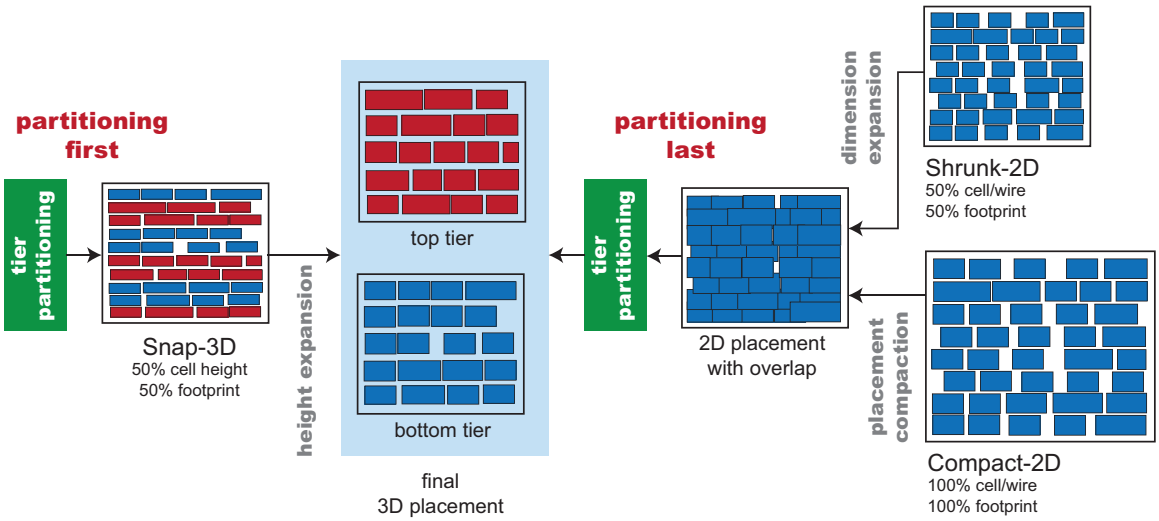


Figure 1: Comparing 2D-3D Transformations of Snap-3D (this work), Shrunk2D [2] and Compact2D [3].

on 3D, the focus of placement was on TSV-based 3D designs [5–9]. Due to the high coupling effect and area cost of TSVs, their goal is to minimize the number of TSVs while maximizing the 3D interconnections. So they do not fully utilize high bandwidth and the smaller 3D pitch possible with a denser 3D bonding. Besides, they lack timing closure capabilities and are not complete RTL-to-GDS flows.

2.2 Pseudo-3D Placers

In recent years, utilizing commercial 2D engines to build 3D ICs (Pseudo-3D flow) with minor configuration tweaks became an appealing approach as the commercial tools provide optimization and timing closure capabilities. However, existing works like Shrunk2D and Compact2D flows [2, 3] fail to account for inter-tier routing and cell displacement during 2D-to-3D transformation, resulting in timing degradation and sub-optimal 3D designs.

Partitioning-last Pseudo-3D flows. Shrunk2D and Compact2D partition the cells onto different tiers after the timing closure stage. The difference between these flows is the Pseudo-3D representation of the final-3D layout, which is illustrated in Figure 1. In Shrunk2D flow, standard cells and Back-End-Of-Line (BEOL) are shrunk by the 3D:2D footprint ratio to fit the cells in a smaller 3D footprint. Compact2D flow, on the other hand, does not shrink standard cells but scales the RC parasitics. As both of these flows do not assign cells to a specific 3D tier in the Pseudo-3D stage, the BEOL at this stage is a 2D metal stack and cannot account for 3D routing and inter-tier effects. After the Pseudo-3D stage, cells are resized back to their original size in Shrunk2D and linearly projected to a 3D footprint in Compact2D. As a result of these transformations, the cells are not contained within the rows and also overlap with other cells. Next, tier partitioning is performed to bisect cells into two tiers with bin-based Fiduccia-Mattheyses (FM) min-cut partitioning approach. After cells are partitioned, they are legalized to fit into rows and remove overlaps. This brings up two issues: cell displacements after legalization, and in-accurate parasitic estimation due to neglecting the inter-tier routing.

Partitioning-first Pseudo-3D flow. The Cascade-2D flow [4] implements pre-partitioned 3D ICs by appending the 3D dies next to each other and converting them into a 2D footprint. Each inter-tier connection is modeled by a pair of dummy anchor cells (a receiver, and a driver) with a zero-parasitic wire connecting them. A sequential die-by-die dummy initial placement of the top and bottom dies determines the location of receiver anchor cells. After placing each die, the receiver anchor cell locations on this die are used to place the driver anchor cells on the complementary die. Once the locations of anchor cell pairs are determined, they remain fixed throughout the flow. The logic cell placement is discarded, leaving just the fixed anchor cell pairs. This has the following drawbacks: With fixed locations, the anchor cells cannot be added or moved. So tier partitioning with a comprehensive architectural information is required to avoid timing and optimization bottlenecks. Secondly, the placement to determine anchor cell locations is not performed simultaneously but in a tier-by-tier fashion. Lastly, the size of anchor cells are restricted by the row definition. So, it cannot accurately model an inter-tier in terms or size and creates unnecessary placement blockages.

2.3 Motivation

As mentioned in Table 1, the drawbacks of Shrunk-2D and Compact-2D are inaccurate parasitic estimation due to neglecting the tier-location of cells. Our proposed design flow takes this into account by assigning cell pins according to their tier location in the 3D metal stack used in our Pseudo-3D stage in Figure2. We also overcome Cascade-2D's limitations by placing the cells from both dies simultaneously and inter-dependently. Inter-tier vias are also dynamically created and adjusted based on the full-chip performance, and congestion. Without anchor cells, the high density of inter-tier connections does not affect the white space and can achieve denser placement.

Table 1: Qualitative Comparison between state-of-the-art flows and this work

	Shrunk-2D	Compact-2D	Cascade-2D	Snap-3D (this work)
Key idea	cell and wire shrinking	placement compaction	cascade floorplan	tier-row snapping
3D stack	two dies	two dies	one die, double metal stack	one die, double metal stack
Strength	first pseudo 3D flow	shrinking unnecessary	handle architectural constraints	best PPA
Weakness	shrinking can be misleading	wire RC derating can be misleading	cannot handle dense inter-tier connections	-
Tier-partitioning	partitioning last (automatic)	partitioning last (automatic)	partitioning first (manual)	partitioning first (automatic)
Clock routing	commercial 2D + tier partitioning	commercial 2D + tier partitioning	commercial 2D	commercial 2D
Placement engine	commercial 2D	commercial 2D	commercial 2D	commercial 2D
Legalization	commercial die-by-die	commercial die-by-die	commercial both dies at once	commercial both dies at once
Post-route opt.	not supported	supported	supported	not necessary

In Compact-2D flow, post tier-partitioning optimization [3] fixes the 2D-3D timing degradation using a similar row halving idea where cells are halved in height while their pins are unchanged and lie outside the cells. However, this approach has the following limitations and issues.

(1) The timing closure is performed with fixed placement because commercial tools could not handle cell placement with pins outside cell area. This requires further legalization when cells are resized.
(2) Since the commercial tools always use a single row type to add new buffers or resized cells, this creates area balancing issue as only one tier will have added cells.

By extending the post-tier optimization concept to overcome these issues as discussed in Section 3, we achieve a way to construct a pseudo-3D stage that can almost completely utilize the EDA tool capabilities, and maintains placement during 2D-3D transformation as shown in Figure 1.

3 METHODOLOGY

3.1 Overview

To build high-performance 3D ICs using 2D commercial tools, the Pseudo-3D stage needs to closely replicate the final 3D design. One of the key stage which determines the quality of the design is placement. Our work performs simultaneous placement of multiple dies using a single die (= 2D) placement engine, and our proposed Pseudo-3D layout directly gives a 3D placement without additional displacements.

Commercial 2D tools only provide a single front-end-of-line (FEOL) layer where instances can be placed. In a two-tier 3D, cells are placed on two layers and cannot fit in a single layer of same size. A simple and effective solution is to split the rows into two identical non-overlapping rows corresponding to each tier. By alternating the order of these halved rows, the distance between rows of the same tier is minimized, while creating abutted rows with alternating VDD, VSS. Along with the rows, the standard cells are also halved as illustrated in Figure 5 to fit in this single layer with 3D footprint

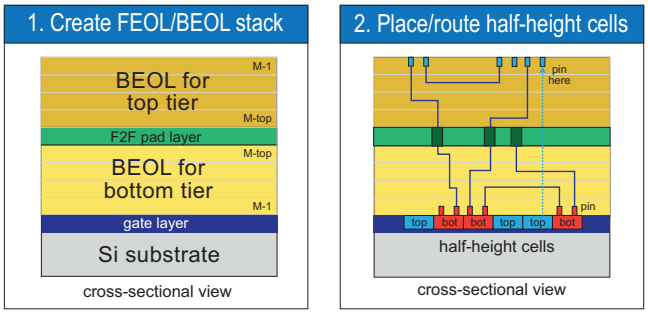


Figure 2: Overview of Snap-3D flow. Cross-section view shown is along the direction perpendicular to rows

size and halved rows. This results in a Pseudo-3D design that closely replicates the final 3D design with Snap-3D flow in Figure 2.

3.2 Our Key Idea

KEY IDEA OF THIS WORK:

Given an original 2D netlist, we first partition the netlist into two from 2D placement layout. Next, we shrink the height of standard cell layouts by one half. Our target placement footprint is 50% of the 2D counterpart. In addition, the rows in this placement footprint are labeled top vs. bottom in an alternating fashion.

Under this condition, we can use any commercial 2D placer to place these half-height cells onto the target row structure and obtain high quality 2-tier 3D placement. Here the cells that are partitioned to the top tier are placed onto the rows that are marked top, and the bottom tier cells go to the bottom rows.

In our layer stack-up, we use a single layer of transistor device with double the metal layer stacks. This resembles the layer structure in face-to-face 2-tier 3D IC, but contains a single layer device so that a commercial 2D physical design tool can perform placement and routing for both tiers simultaneously.

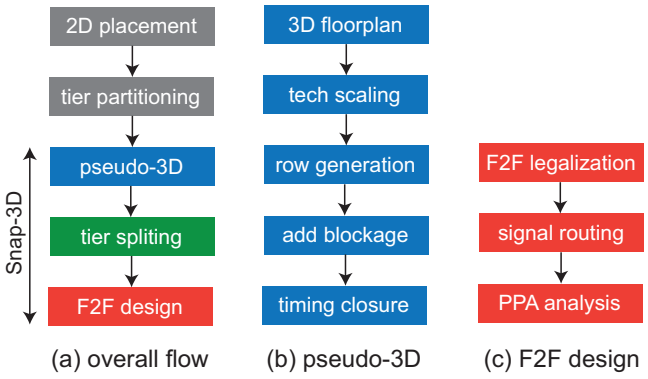


Figure 3: Flow charts. (a) Snap-3D design flow, (b) Pseudo-3D stage of Snap-3D, (c) F2F design stage of Snap-3D.

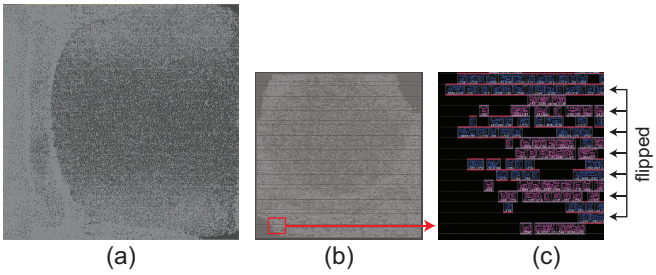


Figure 4: Snap-3D placement layouts. (a) given 2D placement, (b) Snap-3D placement using half-height and 50% footprint, (c) zoom-in shot.

3.3 Snap-3D Flow

The overall Snap-3D flow is shown in Figure 3(a). First, the flow utilizes a 2D layout (placement) as an input for the placement-driven tier partitioning algorithm [10]. This includes manual memory placement and standard cell placement. Next, cells are partitioned onto the two tiers to generate a modified 3D netlist with regards to their tier locations. This is followed by the Pseudo-3D stage, which contains the sub-stages given in Figure 3(b). A 3D footprint is created with half the initial 2D footprint (as there are two tiers in 3D as opposed to single tier in a 2D design). The memory macros are then manually placed and flattened to determine the available regions to place standard cells. Then, placement rows are generated such that placement will be maintained after 3D transformation, and the pseudo-3D design is remarkably close to the final 3D design. Next, the blockages are inserted to constrain standard cell placement. The Pseudo-3D final step is the 3D timing closure. The methodologies for these steps are discussed in detail in the following subsections.

3.4 Technology Scaling

In this section, the technology (cell, row, SITE) scaling methodology is presented. The FEOL technology files for any design are LEF for physical information of cells, and LIB for the capacitance, timing, and power information of the cells. The top-down view of the physical structure of an example cell is illustrated in Figure 5(a) for a normal unmodified cell. First, cells are scaled down by one half in height, while keeping their width constant. Second, standard cells

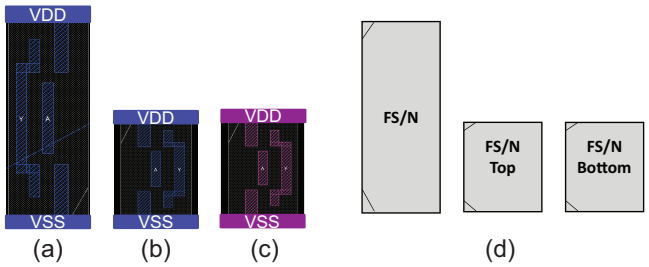


Figure 5: Technology scaling. (a) 2D standard cell, (b) half-height version for top tier, (c) half-height version for bottom tier, (d) placement site definitions for (a), (b), and (c).

are duplicated and their pin layers are modified to represent cells of the top and bottom tiers. These top and bottom die variations of the example cell are illustrated in Figure 5(b),(c). Two different placement SITE definitions are required to separate the rows of the top and bottom tiers, and constrain the cell placement accordingly. Therefore, the original standard cell site is halved and duplicated into two different 3D SITEs as illustrated in Figure 5(d). 3D BEOL remains unchanged to produce precise parasitic estimation (Figure 2). The impact of the half-height cell pin is negligible since cell timing and power characteristics remain untouched in the LIB files.

3.5 Tier Partitioning

In Snap-3D flow, tier partitioning is achieved by applying the placement-driven bin-based FM min-cut algorithm [10]. The required placement for this partitioning is done in a 2D footprint to target desired cell utilization without causing overlaps. Snap-3D's 3D footprint requires a partitioning solution to be implemented without overlaps, and cannot be used at this stage. The partitioning solution is used to modify the cell names in the 2D cells to generate a 3D netlist. The modified cell names refer to the modified top and bottom tier cells mentioned in the Technology Scaling section. Figure 4 shows the layouts from 2D placement in Figure 4(a) to Snap-3D placement in Figure 4(b). Figure 4(c) shows a zoomed-in placement region with pins colored according to their 3D tier.

3.6 Handling Memory Macros

As discussed before, the user pre-places the memory macros manually on the 3D footprint to specify the (X, Y, Z) locations of macros as shown in the example placement in Figure 8. Once memory macros are placed, the memories are flattened by ignoring tier locations to determine available regions for standard cells. The regions where two memories from different tiers overlap (M1/2 at the bottom right corner in Figure 8 are unavailable for cell placement and a full placement blockage is created in the 3D floorplan (Figure 8). In regions where only one memory macro from either die is present (area occupied by macro M3 in Figure 6), standard cells can still be placed in this area on the other die. Let's call them the partial blockage regions.

In the S2D flow[2], a partial blockage is handled by limiting the region density to half the target cell utilization density. However, our flow defines the placement rows for two different tiers separately. So the partial blockage regions due to a memory macro are converted to full placement blockages for tier rows matching the

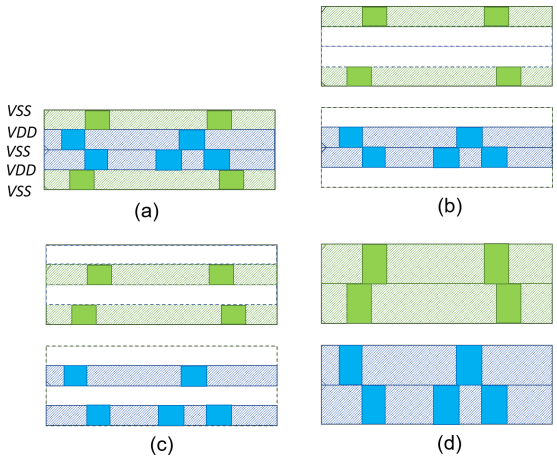


Figure 6: Snap-3D placement row handling. (a) Pseudo-3D placement rows, (b) top and bottom tiers after splitting, (c) after shifting down the even rows, (d) after resizing with respective location from pseudo-3D layout.

Cadence Resize table	
Cell group	Cell Name
NAND2	NAND2_1X_1
	NAND2_2X_1
	NAND2_4X_1
	NAND2_1X_0
	NAND2_2X_0
	NAND2_4X_0
...	
XOR2	XOR2_4X_0/1

(a)

Snap-3D Resize table	
Cell group	Cell Name
NAND2_top	NAND2_1X_1
	NAND2_2X_1
	NAND2_4X_1
NAND2_bot	NAND2_1X_0
	NAND2_2X_0
	NAND2_4X_0
...	
XOR2_bot	XOR2_4X_0

(b)

Figure 7: Standard cell grouping for cell resizing. (a) Cadence, (b) Snap-3D. Resizing with (a) leads to area unbalance between the top and bottom tiers in 3D ICs.

tier location of the memory macro, and no blockages in the other tier.

Next, memory scaling is required to allow standard cells to be placed on partial blockages area. The scaling approach is the same as the state-of-the-art design (Shrunk2D)[2] by shrinking memory macros boundary to the size of a SITE row while keeping memory pins in place. Note that the memory macros also have two variations like standard cells, with pins in either top or bottom tier BEOL depending on the variation.

3.7 Snap-3D Place and Route

In this section, placement row generation, resizing constraints and timing closure in Snap-3D are presented. Once 2D-placement, tier partitioning, and netlist modification are completed, a 3D footprint is first created. The user provides the chip dimensions or utilization, and a custom script creates the placement rows in specified order shown in Figure 6 (a)

Then, the macro placement information (location, orientation) are used to create partial (tier specific) and full blockage (blockages on all tiers) regions Figure 8.

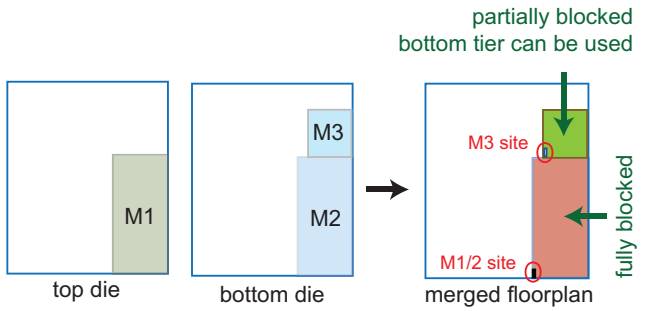


Figure 8: Macro handling in Snap-3D. Snap-3D uses a commercial 2D placer, so the placement area needs to be correctly distinguished. Macros become smallest possible size (= site), not to intervene with gate placement.

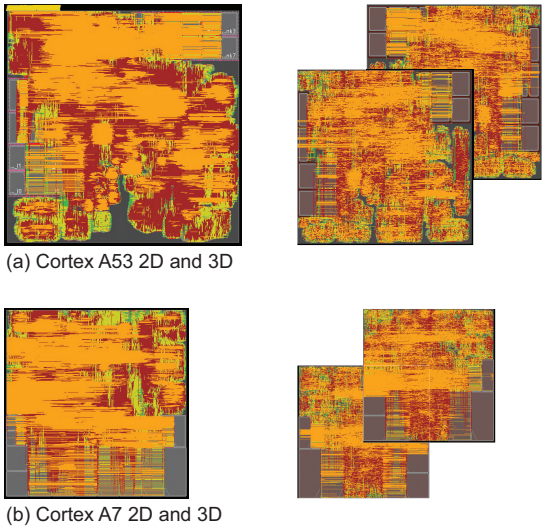
Before the timing closure stage, a tier-area balance setting is important to ensure that the area of cells in each die is similar. A heavy unbalance leads to a non-optimal chip area and the presence of unnecessary white spaces in the design. Since the commercial tools have a single default tier-row. this is used as the cell row for resized or newly added cells. This un-does the partitioning by changing the cell tier of the non-default tier to the default tier when resizing. This leads to an area skewed final design, and significantly reduces the available area for timing closure.

In traditional 2D technology, cells are grouped into equivalent groups like { NAND2, NOR2, XOR2, INV, ...}, and the tool uses the best cell in each group to achieve timing closure as illustrated in Figure 7(a). So, if top-tier NAND2 and bottom-tier NAND2 cells are not differentiated, the tool simply uses the required cell from the NAND2 group with default site row. This is the cause of the area skew. To work around this limitation, the cell groups are further split based on their tier for 3D designs. This doubles the number of groups to { NAND2_<tier0>, NAND2_<tier<1>, NOR2_<tier0>, NOR2_<tier1>, ... }as illustrated in Figure 7(b). With this tier-based resizing, cells are kept at the specified tier location throughout the design.

For buffer insertion during timing closure, cells are honored by placement constraints and available spaces. Thus, the number of added buffers in both tiers are balanced. With tier-row placement constraints and tier-based resizing, we enable 2D EDA tools to build the multiple tiers 3D design simultaneously.

3.8 Coping with Pin-access Issues

Since the standard cells in the snap-3D flow are scaled-down by half height-wise, pin access becomes the main concern for an accurate routing estimation in the timing closure stage. Conventional pseudo-3D flows use a single 2D metal stack for routing, and such scaled pin shapes would require scaled tracks and simplified DRCs to complete routing. However, as Snap-3D flow utilizes the 3D BEOL, scaling is not required to create the same routing track structure as 3D. Moreover, as the adjacent placement rows are from different tiers, they have cells with pins in different layers and use different routing tracks. Vias and metal layers connecting to the pins still violate Design Rule Checks (DRCs) due to pin shape scaling within cells, but that is a Pseudo-3D stage and the violations are fixed after the cells and pins are resized back in the F2F 3D design stage.



(a) Cortex A53 2D and 3D

(b) Cortex A7 2D and 3D

Figure 9: Commercial 28nm GDSII layouts. (a) Cortex A53, (b) Cortex A7. Other benchmark layouts are omitted due to the space limitation.

3.9 Tier Splitting

Once the Pseudo-3D design of Snap-3D is completed, the layout is split into two tiers based on placement row type. Figure 6(a) ,(b) present how a layout is split based on placement row. Even-numbered placement rows counted from the lower left of the footprint are shift down before resizing back to their original height as illustrated in Figure6(c),(d). As a result, 3D placement is obtained from the Pseudo-3D stage without any legalization or displacements because the manufacturing grid in the horizontal axis remains unchanged.

3.10 F2F Design Considerations

Once pseudo-3D timing closure is completed, F2F Design is an important step to fix the design rule violations (DRVs) and relocate F2F-pad locations based on the routing changes. The steps are shown in Figure 3(c). As routing from Snap-3D's Pseudo-3D stage has DRVs from half-height pins, the design is re-routed with normal-size cells in true-3D BEOL without any further timing closure. F2F via locations obtained here are used in tier-by-tier routing to further optimize the routing with minimum DRVs. The tier-by-tier routed designs are analyzed as a whole for timing and power analyses.

4 EXPERIMENT RESULTS

In this section, we analyze the impact of Snap-3D design by comparing PPA metrics with 2D design, and State-of-the-art 3D designs– Shrunk2D and Compact2D. Seven design benchmarks including four pure-logic designs and three processor designs containing SRAM caches are used for comparison.

Pure-logic circuits contain two wire-dominated circuits VGA and LDPC_ENC_DEC, and cell-dominated circuits TATE and AES_128. Three memory-and-logic processor benchmarks are single core instantiations of RocketCore, Arm® Cortex®-A7, and Arm® Cortex®-A53.

The experimental results contain iso-performance, maximum performance, memory metric comparisons, and clock metric comparison, providing an in-depth analysis of our Snap-3D flow.

4.1 Experimental Setup

For our designs, we use a commercial 28nm process design kit (PDK) with 6 metal layers for signal routing in 2D, and 12 metal layers (double the 2D metal stack) for 3D design with Face-to-Face via as inter-tier connections. F2F via size, pitch, resistance and capacitance are set to be 0.5um, 1.0μm m, 0.5Ω and 0.2fF respectively. For Compact2D design, post-tier optimization and incremental routing are not performed as they were not supported with the commercial PDK used here.

The maximum clock frequency achievable for 2D implementations are 1.5GHz for VGA, 0.75GHz for LDPC_ENC_DEC, 1.8GHz for TATE, 2.8GHz for AES_128, 1GHz for RocketCore. Values are normalized for Cortex-A7, Cortex-A53 due to an NDA, and protect sensitive content.

3D footprint used in all the 3D flows is exactly half the 2D footprint. Final GDS layouts are shown in Figure 9. The layout illustrates 2D layout and 3D layout of Snap-3D flow.

4.2 Iso-performance Comparison: Logic Circuits

From Table 2, we analyze the impact of Snap-3D flow by comparing essential metrics over 2D design with state-of-the-art designs (S2D and C2D) for pure-logic circuits.

In cell-dominated circuit, we observed the best wirelength saving over 2D in Snap-3D with 27.2% and 21% reduction for VGA and TATE design respectively. In VGA design, the total power in Snap-3D beats both S2D and C2D with a margin of 8.7%. In TATE, power saving in Snap-3D over 2D design is 6.1% which is slightly smaller than C2D due to the higher clock power used in Snap-3D to achieve better timing. Considering the performance, Snap-3D achieves the best timing in both designs. With a better combination of timing and power, PDP and EDP in Snap-3D are the smallest.

In LDPC_ENC_DEC design with significant wire connections, we obviously see the significant wirelength saving on Snap-3D over 2D design with 31.2%. This is a consequence of the Snap-3D placement in the Pseudo-3D stage as it perform placement with both dies simultaneously. The difference in placement between Snap-3D and 2D design can be seen in Figure 4(a) and (b) respectively, while S2D and C2D placements are almost identical with 2D placement because of the BEOL setting. In addition, Snap-3D shows the best power saving with a 26.6% reduction compared to 2D as a result of the wirelength reduction. Considering performance metrics, Snap-3D has better worst negative slack than 2D and S2D but is marginally bad compared to C2D.

In pure-logic benchmarks, Snap-3D shows better wirelength saving than S2D and C2D. Moreover, Snap-3D provides better performance for all benchmarks with huge power savings in the wire dominant LDPC_ENC_DEC. The total negative slack of the violating paths is generally best in Snap-3D compared to the other 3D flows.

Table 2: Iso-performance PPA comparison among 2D, S2D[2], C2D[3] and Snap-3D for pure-logic benchmarks. WNS and TNS respectively denote the worst and total negative slack.

	VGA				TATE				LDPC_ENC_DEC			
	2D	S2D	C2D	Snap3D	2D	S2D	C2D	Snap3D	2D	S2D	C2D	Snap3D
Target Frequency (GHz)	1.5				2.0				0.75			
Footprint (mm^2)	0.09	0.05	0.05	0.05	0.36	0.18	0.18	0.18	0.27	0.135	0.135	0.135
No of Cells	37556	36799	35526	35379	211088	210790	210507	211238	108,045	101,090	94,506	135,240
F2F Via #	0	13385	13283	14254	0	59481	58679	58848	0	35,972	37,572	53,976
Wirelength (m)	1.21	0.92	0.91	0.88	2.19	1.82	1.80	1.73	7.91	5.8	5.72	5.44
Total Power (mW)	33.33	31.01	31.18	30.44	352.00	342.00	339.85	343.42	234.61	191.47	179.92	172.29
WNS (ns)	0.04	0.05	0.09	0.04	0.05	0.08	0.09	0.05	0.141	0.134	0.0890	0.0950
TNS (ns)	0.59	1.35	11.48	0.95	23.00	89.00	104.00	29.20	37.94	0.75	11.60	23.80
Power Delay Prod	23.63	22.10	23.63	21.43	214.02	218.20	220.90	207.12	345.11	280.31	255.31	245.51
Energy Delay Prod	16.76	15.75	17.91	15.09	130.12	139.21	143.59	124.91	507.66	410.38	362.28	349.86

Table 3: Iso-performance PPA comparison among 2D, S2D [2], C2D [3] and Snap-3D for processor designs with memories. The values for ARM Cortex-A7 and Cortex-A53 commercial processors are normalized w.r.t. 2D design. F2F Via count is normalized w.r.t. S2D. WNS and TNS respectively denote the worst and total negative slack.

	RocketCore				Cortex-A7*				Cortex-A53*			
	2D	S2D	C2D	Snap3D	2D	S2D	C2D	Snap3D	2D	S2D	C2D	Snap3D
Target Freqency (GHz)	1.0				1 (no units)				1 (no units)			
Footprint (mm^2)	0.31	0.15	0.15	0.15	1.00	0.50	0.50	0.50	1.00	0.50	0.50	0.50
No of Cells	123204	122247	122652	120791	1.00	1.00	0.98	0.99	1.00	0.94	0.93	0.94
F2F Via #	0	35238	35802	35634	0.00	1.00	1.01	1.05	0.00	1.00	1.01	1.15
Wirelength (m)	1.87	1.47	1.47	1.46	1.00	0.75	0.73	0.73	1.00	0.69	0.70	0.73
Total Power (mW)	151.37	145.76	145.30	142.55	1.00	0.92	0.89	0.91	1.00	0.67	0.66	0.67
WNS (ns)	0.06	0.10	0.14	0.08	1.00	1.31	1.48	0.85	1.00	0.57	1.12	0.33
TNS (ns)	14.47	58.67	185.16	43.22	1.00	7.21	15.47	0.47	1.00	0.96	1.46	0.11
Power Delay Prod	159.85	160.10	165.50	153.98	2.10	2.21	2.30	1.78	2.10	1.12	1.46	0.97
Energy Delay Prod	168.80	175.86	188.50	166.33	4.41	5.33	5.93	3.47	4.41	1.86	3.24	1.39

Table 4: Max performance PPA comparison for Cortex-A7. Values are normalized w.r.t. 2D design.

	2D	S2D	C2D	Snap3D
Target Frequency	1.00	1.00	1.00	1.11
No. of Cells	1.00	1.00	0.98	1.00
F2F Via #	0.00	1.00	1.01	1.26
Wire length	1.00	0.75	0.73	0.75
Total Power	1.00	0.92	0.89	1.04
Worst Neg. Slack	1.00	1.31	1.48	1.06
Total Neg. Slack	1.00	7.21	15.47	1.15

4.3 Iso-performance Comparisons: Processor Designs

Here, we present the impact of Snap-3D flow in the processor designs with memory macros to identify the importance of honoring memory pins in Table 3.

In memory benchmark designs, we observe the comparable wirelength and total power in Snap-3D among 3D design except RocketCore design.Both S2D and C2D have very bad timing in 3D. So, the power numbers do not reflect the impact of achieving timing closure in 3D. In terms of performance and timing closure, Snap-3D achieves the best performance with the least absolute WNS and TNS among all designs. The main reason is that we honor the pins of pre-placed memory during the timing closure stage, and properly consider partial and full blockages for cell placement.

Table 5: Clock tree metrics comparison for AES_128

Clock Metrics		S2D	C2D	Snap3D
Target Frequency	(GHz)	2.8		
Clock Latency	(ps)	181.5	177.6	166.1
Clock Skew	(ps)	11.7	11.3	8.5
Clock WL.	(mm)	42.15	41.33	38.99
Clk. F2F via	(#)	674	671	731
Clock Buffer	(#)	910	849	862
Worst Negative Slack	(ns)	0.05	0.08	0.0289

4.4 Max-performance Comparison

In Table 4, we present the maximum performance implementations of Cortex-A7 design. We allow a WNS of 10% of the clock period at the maximum frequency. We do so since WNS=0 implies the tool could further improve the timing if a smaller clock period is provided. Choosing a negative value shows that the timing closure was

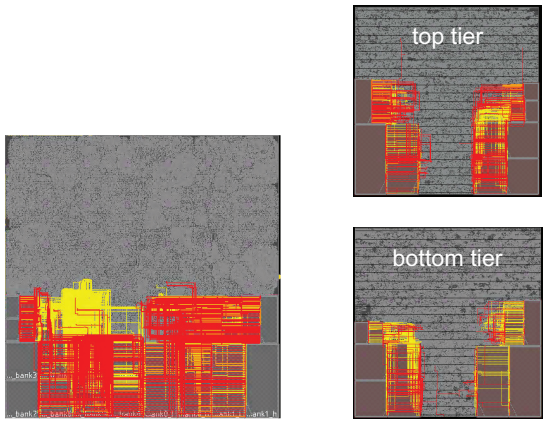


Figure 10: Cortex-A7 routing to and from memory macros. (a) 2D, (b) Snap-3D. Yellow is input nets to memory, and red is output nets from memory.

Table 6: Memory metrics comparison for Cortex-A7 at the maximum 2D frequency.

Memory Metrics	2D	S2D	C2D	Snap3D
Target Freq (GHz)	0.9			
Max mem. Latency (input)	1.000	0.931	1.083	0.972
Max mem. Latency (output)	1.000	0.500	0.286	0.296
Avg Max mem. Latency	1.000	0.682	0.624	0.582
Mem. Access Power	1.000	0.586	0.579	0.576

not able to further reduce the slack. A fixed percentage of the clock period is chosen for a fair comparison. All the implementations in Table 4 use the same initial netlist that is synthesized at maximum 2D frequency. In Snap-3D flow, the maximum achievable frequency is 11% higher than any other implementation.

4.5 Memory Metric Comparison

Memory metrics are analyzed among 3D designs using two main parameters. The first parameter is the worst memory access latency for nets connected to both input (to memory) and output (from memory) pins of the memories. The second is the total switching power of memory nets, which measures memory access power of the design. These two memory metrics reflect the placement and routing quality of cells and nets connected to the memory blocks. In Table 6, we report these memory metrics for Cortex-A7 benchmark as it contains the most memory macros. First, the worst memory access latency for input in S2D is minimum, followed by Snap-3D and C2D, while the memory access latency for output in C2D is minimum followed by Snap-3D and S2D. For a fair comparison, we average the worst memory access latency for input and outputs, and Snap-3D has the best average memory access latency.

As for memory net switching power, all 3D designs show a similar memory access power within small error margins compared to 2D design (0.581 ± 0.005). Snap-3D has the smallest memory net switching power among these designs with a 2D normalized value of 0.576. In summary, Snap-3D flow has a slightly better standard cell placement quality with connection to memory macros compared to state-of-the-art 3D designs. When compared to 2D, the 3D designs show a significantly better reduction in memory net switching power, and memory net latencies. The memory routing layouts are illustrated in Figure 10, where yellow and red lines are memory input and output nets respectively.

4.6 Clock Metrics Comparison

Clock metrics are analyzed among 3D designs using AES_128 circuit due to high number of sequential cells and clock frequency. In Table 5, Snap-3D achieves best clock latency, clock skew and clock wirelength because of compact placement and true 3D timing closure. This results in higher performance (WNS). Furthermore, the clock buffer inserted is not extensive and even less than S2D design, which confirms the impact of clock tree in Snap-3D.

5 CONCLUSION

In this paper, we have proposed a novel physical design methodology named Snap-3D flow to overcome timing degradation and displacement issue in Pseudo-3D stages. We present tier-swapping rows, placement constraints, resizing constraints, and tier-splitting ideas, which allows us to overcome Pseudo-3D issues and provide high-quality 3D designs. Snap-3D flow provides 3D-ready placement without any placement legalization. With a precise parasitic estimation of the 3D routing and no cell displacement after tier splitting, performance is optimal and not degraded from Pseudo-3D to 3D stages as in state-of-the-art design flows.

ACKNOWLEDGMENTS

This research is partially funded by the DARPA ERI 3DSOC Program under Award HR001118C0096, the Semiconductor Research Corporation under Task 2929, and the National Research Foundation of Korea under NRF-2020M3F3A2A02082445.

REFERENCES

[1] Khushu et al. Lakefield: Hybrid cores in 3D package. In Hot Chips Symposium, pages 1–20, 2019.
[2] S. Panth et al. Shrunk-2-D: A Physical Design Methodology to Build Commercial-Quality Monolithic 3-D ICs. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2017.
[3] B. W. Ku et al. Compact-2D: A physical design methodology to build two-tier gate-level 3-D ICs. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 39(6):1151–1164, 2020.
[4] Kyungwook Chang et al. Cascade2D: A design-aware partitioning approach to monolithic 3D IC with 2D commercial tools. International Conference on Computer-Aided Design, pages 1–8, 2016.
[5] Jingwei Lu et al. ePlace-3D. ISPD, 2016.
[6] D. H. Kim et al. Study of Through-Silicon-Via Impact on the 3-D Stacked IC Layout. TVLSI, 21(5):862–874, 2013.
[7] J. Cong and G. Luo. A multilevel analytical placement for 3D ICs. In ASPDAC, pages 361–366, 2009.
[8] M. Hsu, Y. Chang, and V. Balabanov. TSV-aware analytical placement for 3D IC designs. In DAC, pages 664–669, 2011.
[9] Jingwei Lu, Hao Zhuang, Ilgweon Kang, Pengwen Chen, and Chung-Kuan Cheng. ePlace-3D: electrostatics based placement for 3D-ICs. In ISPD, pages 11–18, 2016.
[10] S. Panth et al. Placement-driven partitioning for congestion mitigation in monolithic 3D IC designs. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 34(4):540–553, 2015.