Pseudo-3D Approaches for Commercial-Grade RTL-to-GDS Tool Flow Targeting Monolithic 3D ICs

Heechun Park, Bon Woong Ku, Kyungwook Chang, Da Eun Shim, and Sung Kyu Lim
School of Electrical and Computer Engineering
Georgia Institute of Technology

ABSTRACT

Despite the recent academic efforts to develop Electronic Design Automation (EDA) algorithms for 3D ICs, the current market does not have commercial 3D computer-aided design (CAD) tools. Instead *pseudo-3D* alternative design flows have been devised which utilize commercial 2D CAD engines with tricks that help them operate as a fairly-efficient 3D CAD tool. In this paper we provide detailed discussions and fair power-performance-area (PPA) comparisons of state-of-the-art pseudo-3D design flows. We also analyze the limitations of each design flow and provide solutions with better PPA and various design options. Our experiments using commercial PDK, GDS layouts, and sign-off simulations demonstrate that we achieve up to 26% wirelength and 10% power consumption reduction for pseudo-3D design flows. We also provide a partitioning-first scheme to partitioning-last design flow which increases design freedom with tolerable PPA degradation.

ACM Reference Format:
Heechun Park, Bon Woong Ku, Kyungwook Chang, Da Eun Shim, and Sung Kyu Lim. 2020. Pseudo-3D Approaches for Commercial-Grade RTL-to-GDS Tool Flow Targeting Monolithic 3D ICs. In *Proceedings of the 2020 International Symposium on Physical Design (ISPD '20), March 29-April 1, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3372780.3375567

1 INTRODUCTION

In order to maximize the benefits of 3D ICs, Monolithic 3D (M3D) ICs [22] with nanoscale Monolithic Inter-tier Vias (MIVs) for inter-tier connections were introduced. M3D offers massive inter-tier connections that are not possible in Through-Silicon-Via (TSV) based 3D ICs. TSV based 3D ICs also suffer from significant area and electrical coupling overhead. To fully benefit from M3D ICs, key technological advancements in design tools are required. However the current market has no commercial vendor that offers them.

Research conducted on 3D IC design flows over time has profoundly impacted the final design quality of a chip. As part of this, Cong *et al.* [7] proposed intuitive transformation techniques such as folding a 2D layout into n stacked tiers, or compressing a 2D design into $1/n$ and splitting each square bin into n different tiers. Goplen *et al.* [11] proposed an expansion of the recursive bisection algorithm that is used in 2D placement by adding a z-axis to the algorithm and there are multiple research groups currently focusing on developing an optimal 3D IC design flow through 3D placement algorithms. There have been attempts to devise a fundamental 3D solution for the 3D placement problem, and we name this approach the *True-3D* placement. These normally extend existing 2D analytic placers [3][5][15][21] to 3D analytic placers [17][12][16][10][13] by considering tier assignment as another dimension with discrete characteristics.

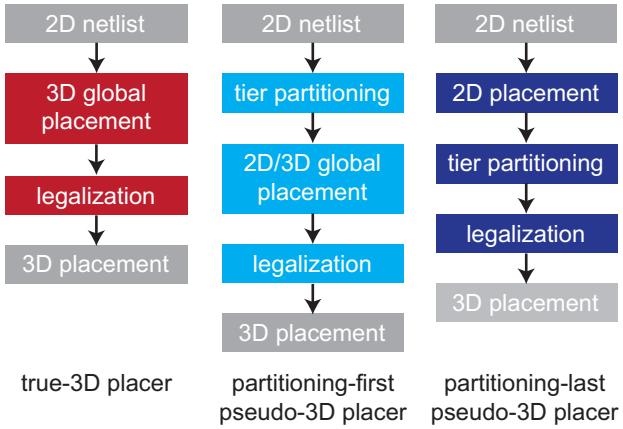

Figure 1: Three approaches to 3D placement.

Another approach, called the *pseudo-3D* design flow, utilizes optimization features of existing 2D CAD engines to produce a 3D IC design. These 3D flows guide the 2D CAD tool to use its high-quality engine on 3D ICs by tricking the engine to operate on a 3D design as if it were working on a 2D design. Pseudo-3D flows concentrate on seamless utilization of 2D CAD tools in 3D IC design instead of developing a completely new 3D CAD tool which needs an extra amount of time for verification and improvement.

In this paper, we introduce the current status quo of various pseudo-3D design flows thoroughly, and provide a comparison of their power-performance-area (PPA) to 2D design. We also present enhancement techniques for each pseudo-3D flow which improves them in different points of view. Our target 3D IC technology is M3D ICs, which is shown to be more effective in logic applications compared with TSV-based 3D ICs [20].

2 DRAWBACK OF TRUE-3D TOOLS

Several academic research groups have proposed their own true-3D design flows [17][12][16][10][13]; each applied their own in-house True-3D placers to perform 3D placement and then verified their results by comparing the half-perimeter wirelength (HPWL) or routed wirelength with the aid of commercial 2D routers applied to

ISPD '20, March 29-April 1, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7091-2/20/03…$15.00
https://doi.org/10.1145/3372780.3375567

Table 1: PPA Comparison of a true-3D placer (Force-3D[13]) with post-placement stages and a pseudo-3D design flow (Shrunk-2D[19]).

		True-3D[13]	Pseudo-3D[19]
AES (3.4GHz)	WL (m)	1.495	1.212 **-18.93%**
	WNS (ns)	-0.077	-0.066
	Power (mW)	274.16	268.00 **-2.24%**
LDPC (1.2GHz)	WL (m)	2.132	1.226 **-42.50%**
	WNS (ns)	-0.195	-0.086
	Total Pwr. (mW)	139.31	81.91 **-41.20%**
Nova (625MHz)	WL (m)	2.870	2.145 **-25.26%**
	WNS (ns)	-0.271	-0.193
	Power (mW)	121.60	110.74 **-8.93%**
TATE (1.4GHz)	WL (m)	2.884	1.853 **-35.75%**
	WNS (ns)	-0.007	-0.098
	Power (mW)	336.69	318.99 **-5.26%**
ECG (1.0GHz)	WL (m)	1.273	0.909 **-28.59%**
	WNS (ns)	-0.050	-0.040
	Power (mW)	107.53	105.70 **-1.70%**

each tier of the 3D design separately. The flowchart on the left in Figure 1 shows the overview of the True-3D placer. The 3D global placement is formulated as an optimization of total wirelength and through-silicon-via (TSV) count with the overlap constraints as follows:

$$\begin{aligned} min \quad & \{HPWL + \alpha \cdot \#TSV\} \\ s.t. \quad & \rho_{b,t} \leq D_{b,t}, \forall \text{bin } b, \text{tier } t, \end{aligned} \tag{1}$$

where $HPWL$ is the total half-perimeter wirelength, $\#TSV$ is the number of TSVs, α is a weight number constant, $\rho_{b,l}$ and $D_{b,l}$ represents the total density of cells and the maximum allowable density in bin b of tier t respectively. After the placement result is generated, TSVs are assigned along with the inter-tier nets (3D nets) which connect cells on different tiers. This introduces additional overlaps between cells and TSVs which are fixed during the last legalization stage.

Unfortunately, these true-3D placers are not ready for commercial adoption yet. True-3D design flows target TSV-based 3D ICs and thus are not ready to handle monolithic 3D ICs. Most constraints for TSVs (e.g., TSV-related mechanical stress, electrical coupling, thermal coupling, keep-out zone, etc.) do not exist in monolithic 3D ICs which makes the comparison across true-3D and pseudo-3D design results unfair. Lastly, True-3D placers are not integrated into the RTL-to-GDS flow, which makes it difficult to evaluate them using final GDS layouts and sign-off simulations. While our work reports practical PPA values such as routed wirelength, timing information and power consumption, True-3D design results are only able to report wirelength estimation based on the HPWL.

As an effort to make a fair comparison, we integrated Force-3D [13], a true-3D placer, into the commercial EDA flow. Force-3D is a 3D expanded placement method of a 2D force-directed quadratic placement solution [21]. With the integration, all post-placement stages (i.e, pre-CTS optimization, CTS, post-CTS optimization, route, and post-route optimization) are performed for each tier with a commercial 2D CAD tool to perform sign-off analysis after 3D placement. Table 1 shows the comparison between Force-3D and a pseudo-3D design flow (Shrunk-2D [18][19], details

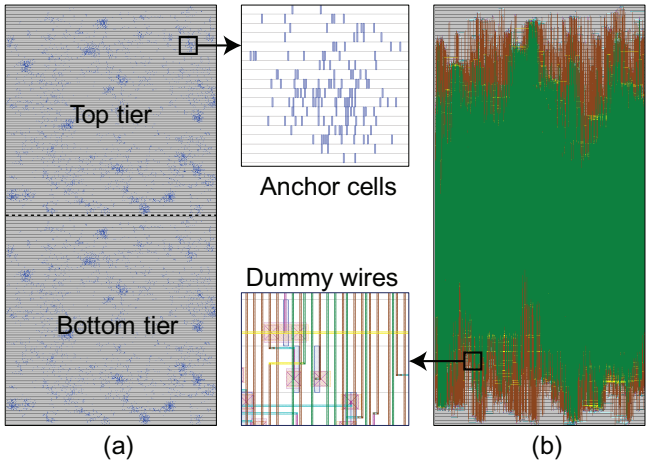

Figure 2: Cascade-2D [4] intermediate layouts. (a) after anchor cell insertion, (b) after routing anchor cells using dummy wires.

in the following section). It is obvious that the True-3D placer does not consider the commercial EDA flow and thus shows worse PPA results compared to the CAD-tool-friendly pseudo-3D design flow.

3 STATE-OF-THE-ART PSEUDO-3D TOOLS

Unlike true-3D, Pseudo-3D design flows focus on benefiting from existing commercial 2D CAD engines. In this section, we discuss three state-of-the-art Pseudo-3D placers: Cascade-2D [4]; Shrunk-2D [19]; and Compact-2D [14]. Cascade-2D is classified as a *partitioning-first* design flow, which perform tier partitioning before placement in $x-y$ coordinates. On the other hand, Shrunk-2D and Compact-2D belong to *partitioning-last* design flows which perform tier partitioning from intermediate placement results, thereby exploiting placement information to generate more specific partition results. The general difference between 3D placement flows is shown in Figure 1.

3.1 Cascade-2D Flow

The main motivation of Cascade-2D [4] is to place, route, and close timing of all tiers simultaneously, rather than performing tier-by-tier designs separately. Since existing commercial 2D CAD tools cannot restore all tiers of a 3D design at the same time, Cascade-2D uses a 2D structure that represents a 3D design in a vertically long floorplan (W:H = 1:2) where the top and bottom half represent the top and bottom tier correspondingly, as shown in Figure 2. Each inter-tier connection is represented with two dummy anchor cells at the MIV position in both tiers (Figure 2(a)) and a zero-parasitic wire that connects the anchor cells (Figure 2(b)).

Figure 3 shows the overall design flow of Cascade-2D. It partitions the 2D netlist into 3D netlists with two tiers (say *top* and *bottom*). The 3D nets are represented as I/O pins in each tier and anchor cells are attached right next to each pin. Then, the MIV locations are determined by anchor cell locations from sequential 2D placement of the top and bottom tier. During the top tier placement,we locate the driving MIVs (top → bottom) at the anchor cell's output ports. The bottom tier placement is performed while

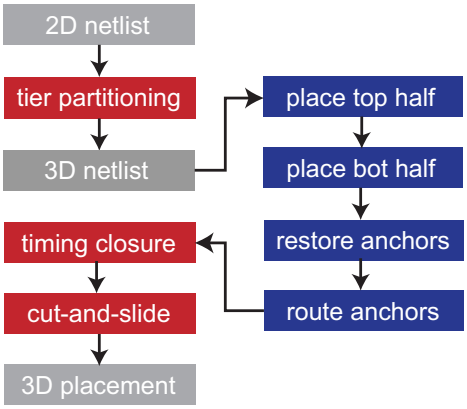

Figure 3: Overview of Cascade-2D [4] design flow.

considering the MIV locations from the top tier placement, which guides the receiving cells to be placed near their driving MIVs. Then the remaining anchor cells and MIVs (bottom → top) are placed in a similar way. MIV locations are then restored to the Cascade-2D floorplan of 1:2 aspect ratio (Figure 2(a)), and routed with zero-parasitic dummy wires (Figure 2(b)). Dummy wires represent the instant connection between both tiers through MIVs. Lastly, it uses a commercial 2D CAD tool to perform P&R and timing closure from scratch as if the tool is running on a 2D design. A hard horizontal fence across the center forbids cross-half (=cross-tier) cell movement during the entire flow.

After all design steps, the Cascade-2D design result is transformed to a monolithic 3D design by laying the upper half (top tier) upon the lower half (bottom tier) and replacing each dummy wire to a vertical MIV. Cascade-2D fully utilizes the efficiency of commercial 2D CAD tools, which leads to a better timing-closed design compared with the tier-by-tier design.

3.2 Shrunk-2D Flow

The main idea behind Shrunk-2D [19] is to use a commercial 2D CAD tool to generate an intermediate result in which all cells and wires are shrunk in half and placed onto a 2D die with a half footprint. This guarantees the full utilization of a commercial 2D P&R tool in terms of (x, y) location of all cells. All we need to do afterwards is to resize the cells and wires to their original sizes, tier partition the cells and insert MIVs to 3D nets.

Figure 4 shows a conceptual view of the Shrunk-2D flow. In the intermediate Shrunk-2D P&R result from the commercial 2D CAD tool, all standard cells and wires are scaled down to 50% and its total footprint has 50% area of its 2D design counterpart (Figure 4(1)). All components are expanded to their original size afterwards, which creates many overlaps in the design (Figure 4(2)). The chip is then divided into several square bins (Figure 4(3)), and the Fiduccia-Mattheyses (FM) min-cut partitioning algorithm[9] is applied to each bin to evenly partition the cells into both tiers (Figure 4(4)). Remaining cell overlaps in both tiers are removed during tier-by-tier legalization (Figure 4(5)). Shrunk-2D, and Compact-2D [14] in the next subsection, perform tier partitioning after the full P&R of an intermediate 2D design with a 2D CAD tool. This enables the use of (x, y) locations for cells during tier partitioning which

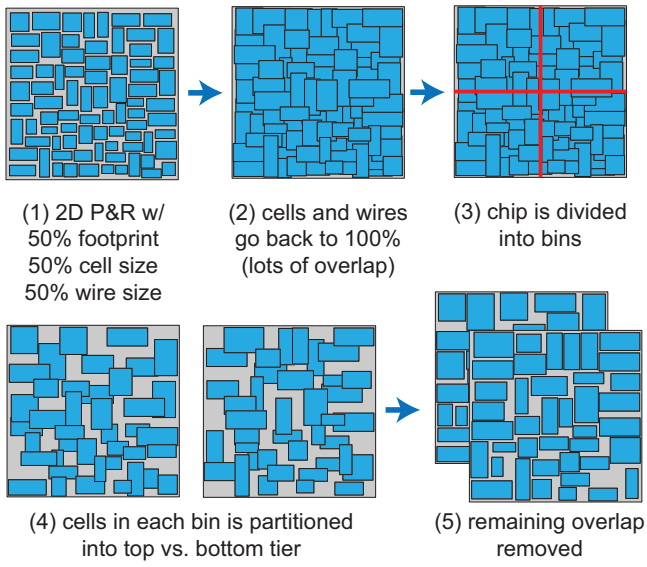

Figure 4: Illustration of Shrunk-2D flow [19].

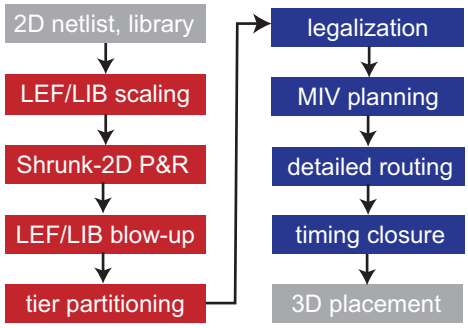

Figure 5: Overview of Shrunk-2D [19] design flow.

improves the partitioning quality to maintain the planar placement result from the commercial tool.

Figure 5 shows the steps in the Shrunk-2D design flow. Technology scaling is first performed to generate the necessary library exchange format (LEF) and liberty (LIB) files for shrunk cells. Then P&R is performed using the shrunk cells as if it were performing 2D P&R. Once the Shrunk-2D design is achieved, LEF/LIB blow-up, tier partitioning and legalization is done as shown in Figure 4. Then, routing is performed sequentially by placing MIVs (MIV planning) and tier-by-tier routing with design constraints from the initial timing analysis (timing closure). An important point to note is that, every step in the Shrunk-2D flow is performed by commercial 2D CAD tools, except for the tier partitioning step. This means that the Shrunk-2D design flow fully utilizes the commercial 2D tool in 3D design and generates an efficient monolithic 3D design output.

3.3 Compact-2D Flow

Shrunk-2D has risks of shrinking cells and wires into an even smaller technology than the commercial tool can handle. It also has inaccurate RC values for shrunk interconnects. This affects the quality of final 3D design results. The Compact-2D [14] design flow

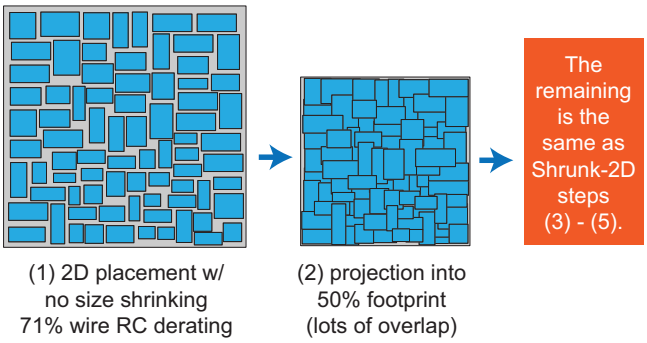

Figure 6: Illustration of Compact-2D [14] flow.

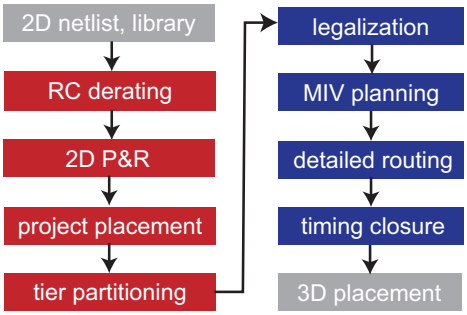

Figure 7: Overview of Compact-2D [14] design flow.

resolves this problem by using similar concepts from Shrunk-2D while not shrinking cells and wires. Instead, it scales only the RC parasitics ($1 \rightarrow 0.707$) by a factor of 0.707 to make the CAD tool treat the interconnect much faster than that of the original 2D design. Therefore a Compact-2D P&R result has smaller timing path delays compared to the same path length in a 2D design, which requires less resources (e.g. buffers, inverters, ...) for timing closure. When compressing the whole design into 50%, the interconnects return to their original RC value and the timing closure resources become sufficient. This flow removes a great amount of potential design rule violations that come from shrinking elements beyond the coverage of CAD tools.

The conceptual view of the Compact-2D flow is shown in Figure 6. The intermediate Compact-2D P&R result has the same area and cell size as a 2D design. In the next stage, it contracts the chip area to 50% and generates a placement result with high overlaps, which is similar to the intermediate stage of Shrunk-2D. Since all interconnects are scaled down and the unit RC parasitics are scaled up to the original size at the same time in and identical proportion, the overall interconnect RC parasitics remain consistent. The rest of the steps are the same as those in the Shrunk-2D flow.

Figure 7 recites the Compact-2D design flow. It is conceptually similar to the Shrunk-2D design flow, except that Compact-2D only adjusts the RC parasitics scaling factor, instead of the whole technology files.[1]

[1]We did not apply the improved post-tier-partitioning timing closure scheme in [14], since it requires modification of commercial-grade physical design kits (PDKs) into an intolerable shape for the CAD tool.

Table 2: Qualitative comparison of pseudo-3D design flows

	Cascade-2D	Shrunk-2D	Compact-2D
Partitioning	Beginning	Middle (After placement)	Middle (After placement)
Fixed tech files	Anchor cells (Easy)	Cells & wire (Hard)	RC factor (Easy)
Macro block	Yes	Yes (partially)	Yes (partially)
Timing closure	Simultaneous	Per-tier	Per-tier

3.4 Qualitative Comparisons

Table 2 lists the characteristics of the three different pseudo-3D design flows. Cascade-2D performs tier partitioning before the CAD tool flow and gives the designer the authority of manual tier assignment, while the Shrunk-2D and the Compact-2D flows merged the bin-based FM min-cut tier partitioning method into their pseudo-3D placement step to pursue the best quality. In terms of modifying technology files, Cascade-2D only needs an extra LEF/LIB of a zero-timing, buffer-like anchor cell for MIV locations. Shrunk-2D, however, needs an extra effort to shrink all geometry numbers related to standard cells and wires, which is perilous for commercial PDKs. Compact-2D solved this issue by only scaling the interconnect RC factor. All flows take care of macro blocks by assigning their tiers and placing them before the CAD-tool-based placement, though Shrunk-2D and Compact-2D add another constraint of 50% partial placement blockage to the partially vacated region (macro on one tier).

Before explaining the last row of Table 2, Figure 8 shows an example of simultaneous and per-tier timing closure for an inter-tier timing path. For simultaneous timing closure (Figure 8(a)), a simple buffer is inserted or a wire is adjusted. This is as accurate as the 2D timing closure since the entire path across both tiers is considered. For per-tier timing closure (Figure 8(b)), however, the path is divided into two segments and treated separately with interface information written in constraint files (i.e., Synopsys delay constraint (.sdc) file). When each tier is modified to close timing, the values in the constraint file become out-of-date since the other tier has been modified. Therefore, it does not guarantee the timing closure of the merged path because separate timing closures has been made under a wrong constraint file.

It is not possible to perform simultaneous timing closure for a 3D design directly with a 2D CAD tool since it is impossible to bring both tiers into a single plane. Cascade-2D made it possible by bringing both tiers vertically with 1:2 floorplan, but Shrunk-2D does not have such a trick and therefore suffers from a large negative slack. Compact-2D introduced a trick for simultaneous timing closure, which is not always available in every PDK.

3.5 PPA Comparisons with Gate-Level Designs

We compared different pseudo-3D design flows in terms of PPA. In this section, design results of 2D (from Cadence Innovus) and monolithic 3D (from Cascade-2D, Shrunk-2D and Compact-2D design flow) are compared using five flattened gate-level benchmark circuits from OpenCore benchmark suites [1], with a commercial-grade 28nm PDK. MIV diameter and RC values are set to $100nm$, 64Ω and $0.2fF$.

Table 3: Comparison of commercial 2D and pseudo-3D design flows (Cascade-2D [4], Shrunk-2D [19] and Compact-2D [14]). % difference is calculated w.r.t. commercial 2D. (WL = wirelength, WNS = Worst negative slack)

		2D	Cascade-2D	Shrunk-2D	Compact-2D
AES-128 (3.4GHz)	MIV count	0	429	39521	39772
	WL (m)	1.444	1.637 +13.37%	1.212 -16.09%	1.224 -15.24%
	WNS (ns)	-0.009	-0.020	-0.066	-0.069
	Power (mW)	274.21	292.92 +6.82%	268.00 -2.27%	266.58 -2.78%
LDPC (1.2GHz)	MIV count	0	5996	15390	16462
	WL (m)	1.527	1.707 +11.81%	1.226 -19.69%	1.197 -21.61%
	WNS (ns)	-0.044	-0.056	-0.086	-0.102
	Power (mW)	85.63	107.83 +25.92%	81.91 -4.35%	74.91 -12.52%
Nova (625MHz)	MIV count	0	264	33980	33658
	WL (m)	2.257	2.571 +13.91%	2.145 -4.96%	2.120 -6.08%
	WNS (ns)	-0.060	-0.054	-0.193	-0.205
	Power (mW)	110.54	113.47 +2.66%	110.74 +0.18%	110.75 +0.19%
TATE (1.4GHz)	MIV count	0	3199	52442	52672
	WL (m)	1.935	3.141 +62.38%	1.853 -4.23%	1.843 -4.73%
	WNS (ns)	-0.012	-0.026	-0.098	-0.129
	Power (mW)	315.44	319.42 +1.26%	318.99 +1.13%	318.53 +0.98%
ECG (1.0GHz)	MIV count	0	1010	21384	21270
	WL (m)	0.989	1.253 +26.59%	0.909 -8.13%	0.900 -9.04%
	WNS (ns)	-0.037	-0.020	-0.040	-0.056
	Power (mW)	104.56	104.24 -0.30%	105.70 +1.09%	105.42 +0.83%

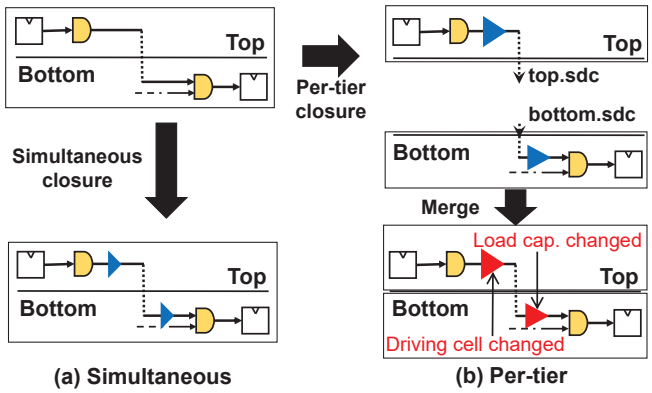

Figure 8: Example of (a) simultaneous timing closure and (b) per-tier timing closure

Table 3 summarizes all results. For each benchmark, an identical 3D footprint is used for all pseudo-3D designs, which occupies 70.7% of width and height compared to 2D. For the tier partitioning algorithm of Cascade-2D, we used various partitioning algorithms [9][6][8], and chose a design with the least power consumption.

Although Cascade-2D is the only pseudo-3D design flow that matches the timing closure of the commercial 2D design flow, its results are the worst among all designs in wirelength and power even when compared to the 2D design counterpart. The main reason is that its tier partitioning strategy and MIV planning method is not suitable for flattened gate-level designs. It results in large wire detouring of 3D nets and increased net switching power.

Shrunk-2D and Compact-2D show better results than 2D designs in most terms, especially in wirelength. Between the two design flows, Compact-2D shows slightly better wirelength and power for most cases. They successfully reduce wirelength by down-scaling the HPWL and using more MIVs to convert long 2D interconnects into short vertical 3D nets. However, their results have larger negative slacks than 2D and Cascade-2D designs, which come from the incompleteness of per-tier timing closure.

Table 4: Comparison of commercial 2D and pseudo-3D design flows on RISC-V Rocketcore processor. % difference is calculated w.r.t. commercial 2D. (WL = wirelength, WNS = Worst negative slack, TNS = Total negative slack, TPS = Total positive slack.)

Rocketcore (1.67GHz)	2D	Cascade-2D	Shrunk-2D	Compact-2D
Footprint(mm^2)	0.563	0.281 -50%	0.281 -50%	0.281 -50%
Std. cell area(mm^2)	0.238	0.241 +1.0%	0.234 -1.9%	0.231 -3.1%
WL (m)	2.896	3.517 **+21.4%**	2.542 **-12.2%**	2.592 **-10.5%**
MIV Count	0	1313	46088	45624
WNS (ns)	-0.046	**-0.033**	**-0.197**	**-0.320**
TNS (ns)	-4.996	**-12.284**	**-606**	**-1091**
TPS (ns)	1763	2025	1360	1106
Cell Power (mW)	334.5	343.0 +2.5%	334.6 +0.04%	332.8 -0.5%
Net Power (mW)	115.2	142.6 **+23.8%**	107.0 **-7.1%**	106.0 **-8.0%**
Leak Power (mW)	47.8	49.06 +2.6%	45.7 -4.4%	44.3 -7.4%
Tot Power (mW)	497.5	534.6 +7.5%	487.3 -2.0%	483.0 -2.9%

3.6 PPA Comparisons with Mixed-size Design

We performed pseudo-3D flows on a mixed-size design benchmark with memory macro modules. Figure 9 Shows the layouts of RISC-V Rocketcore processor [2] with 64KB L1 cache memory, using 2D

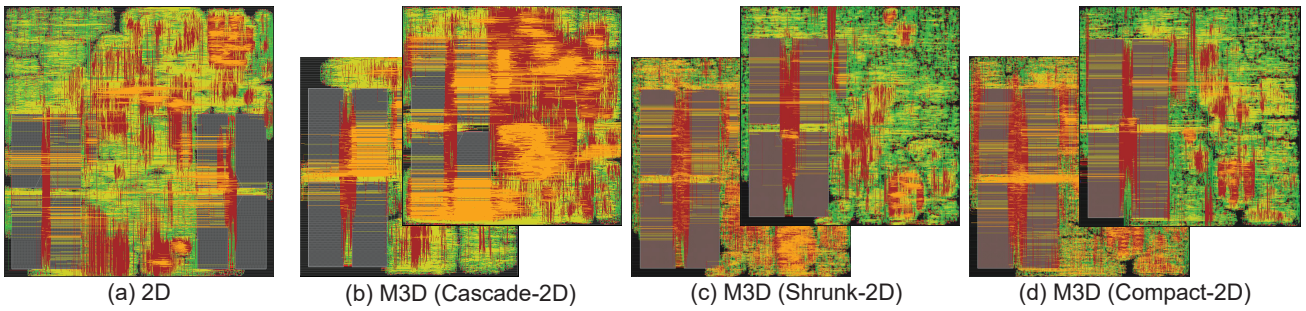

Figure 9: GDSII layout images of RISC-V Rocketcore processor in 2D and M3D designs using different pseudo-3D design flows.

and all pseudo-3D design flows. There are 10 memory macro blocks in the design, 5 of them belong to L1 data cache and the others to instruction cache. In monolithic 3D designs (Figure 9(b) (d)), Instruction cache blocks are assigned to the top tier and data cache blocks are placed on the bottom tier. For tier partitioning in Cascade-2D, we applied the clustering based LR algorithm [6] that gathers cells related to each other in the same tier as much as possible, making the least inter-tier connections.

Table 4 compares the PPA values of all design flows, which shows a similar trend to the gate-level design results (Table 3). The monolithic 3D design created by the Cascade-2D flow shows better timing related values, while its wirelength and corresponding power consumption become worse. Shrunk-2D and Compact-2D based designs show the opposite results with better wirelength/power and worse WNS/TNS.

4 PSEUDO-3D ENHANCEMENTS

In this section, we present enhancement techniques for each pseudo-3D design flow to overcome the weaknesses we have found from their present status.

4.1 Placement-Aware MIV Planning

Although both the tier partitioning and MIV planning steps in the Cascade-2D could be improved, we focused on enhancing MIV planning while retaining the traditional tier partitioning algorithms [9][6][8]. Normally, the 2D placement engine in a CAD tool has the tendency to place cells at the chip *periphery* if they are logically close to the I/O pins. Since anchor cells are logically attached to the I/O pins, in the MIV planning proposed in [4], the anchor cells are pushed towards the chip periphery during sequential placement. This is not a big problem in a hierarchical design with a design-specific tier partitioning (e.g. all cells in a functional block belong to a same tier) because logic gates in a functional block are clustered and they have a strong attraction to pull corresponding anchor cells towards them. However, in an arbitrary design where the design hierarchy is unknown, anchor cells are affected more by the pulling attraction towards the chip periphery than that of their correlated cells.

This trend is clearer when the MIV count is large, which means related cells of an anchor cell are more likely spread over different tiers, and therefore anchor cells are more attracted to the periphery. Figure 10 shows that a higher MIV count results in MIVs cornered

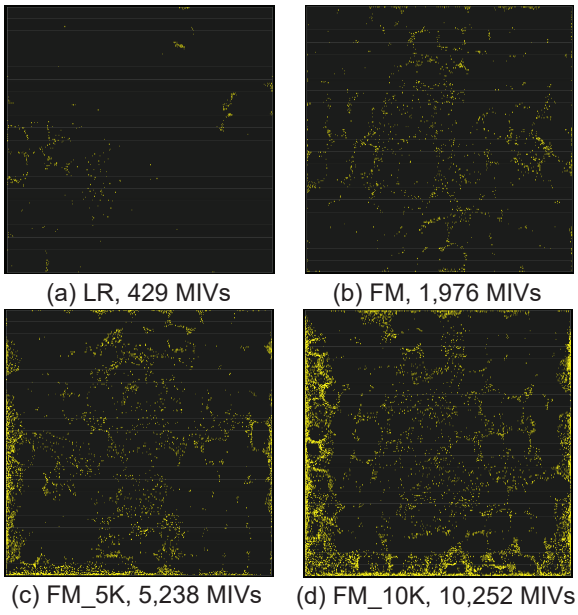

Figure 10: MIV placement for various MIV counts. Yellow dots represent MIV locations.

at the periphery. 3D net routing with such MIV placement results in large detoured wires increasing the total wirelength significantly.

To remove such effect, we devised a *placement-aware MIV planning* to evenly distribute the MIVs. As shown in Figure 12(a), the results from tier partitioning are used along with the 2D P&R result of the design to find out the location of the 3D nets (connected cells assigned to different tiers would mean a 3D net). MIV positions are then determined as the center-of-mass of the 3D nets. Then, they are projected onto the 3D footprint. Figure 12(b) and (c) show the enhanced MIV placement results of Figure 10(c) and (d). MIVs are not crowded at the chip periphery any more but rather spread over the entire chip region.

Table 5 shows the PPA improvement of both cases in Figure 12. Placement-aware MIV planning places MIVs at better locations reducing total wirelength up to 13.29% for the FM_5K tier partitioning algorithm (5,238 MIVs). The FM_5K tier partitioning algorithm performs FM min-cut algorithm [9] until the cut size reaches 5,000. It also reduces redundant buffer insertion for timing closure and the total power consumption is reduced.

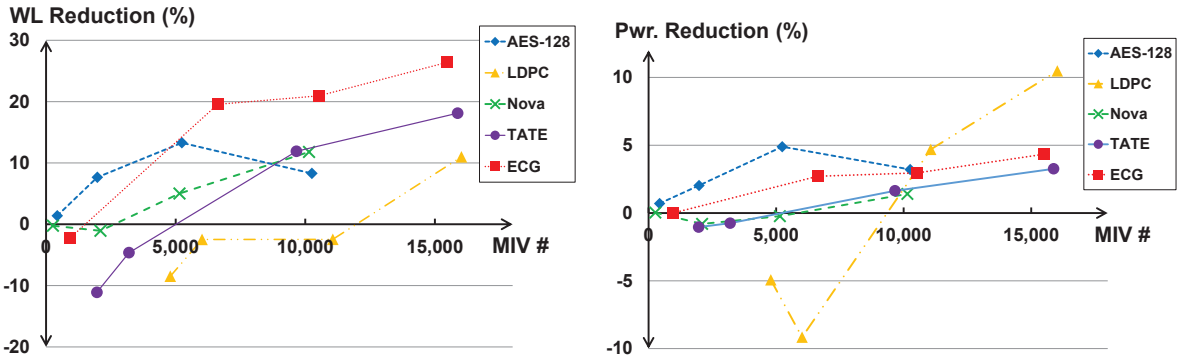

Figure 11: Wirelength and power savings with placement-aware MIV planning w.r.t. MIV count. Each benchmark is designed with tier partitioning results of 4 different MIV counts

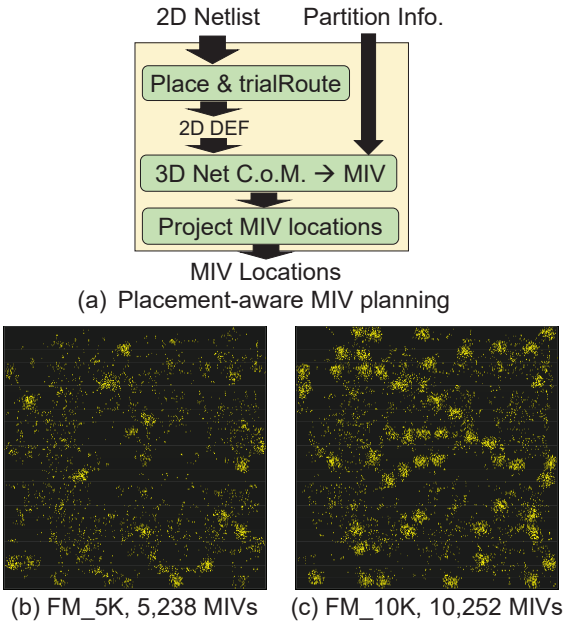

Figure 12: (a) Placement-aware MIV planning flow. (b)(c) Enhanced MIV placement results from Figure 10(c)(d)

Table 5: Comparison of Cascade-2D with original MIV planning and placement-aware MIV planning for AES-128 benchmark.

	Orig.	Place-aware		Orig.	Place-aware	
MIV #	5,238			10,252		
Cell area (mm^2)	0.127	0.123	-3.60%	0.133	0.126	-4.89%
WL (m)	1.817	1.575	-13.29%	1.894	1.737	-8.29%
Total Pwr. (mW)	301.00	286.29	-4.89%	305.61	295.8	-3.21%

Figure 11 plots the wirelength and power savings of the placement-aware MIV planning. The largest wirelength reduction is obtained in the ECG benchmark with FM_15K tier partitioning (15,485 MIVs), which reduces the wirelength by 26.42% compared to the original MIV planning method. The largest power reduction is 10.48% in LDPC benchmark with FM_15K tier partitioning (16,022 MIVs). It is obvious that our placement-aware MIV planning works much better when the MIV count is larger, since it worsens the results of original MIV planning.

4.2 Partitioning-First Adoption

One design constraint that both the Shrunk-2D and the Compact-2D flow have in common is that they have an internal tier partitioning method (bin-based FM min-cut algorithm). Although this guarantees maximal efficiency it has less freedom to control the design flow. This is a concern especially when it comes to designs that the designer wants to manually manipulate tier partitioning based on enough information. In this subsection, we applied the partitioning-first scheme into these two design flows to determine the effects.

Our partitioning-first scheme in both Shrunk-2D and Compact-2D performs RTL-level tier partitioning at the beginning, and imports this partition result into the tier assignment stage after P&R. Additional optimization resources (i.e., buffers, inverters) that are inserted during P&R are assigned to the tier with more connected cells by applying the bin-based FM min-cut algorithm only to those added cells. Figure 13 shows the flowcharts of the partition-first 3D placement scheme, and the remaining stages are the same as their original design flow (Figure 5 & Figure 7).

In Figure 14, while cells are evenly distributed in both tiers of the original Shrunk-2D design (left), partitioning-first Shrunk-2D design (right) has more whitespaces between cell groups, which incurs inefficient resource utilization and thus degraded PPA results. The partitioning-first scheme performs tier partitioning without the future placement information of cells and has no technique to avoid expected local area skews in the final placement result.

Table 6 shows that design results based on the partitioning-first scheme has longer wirelength than that of the original Shrunk-2D and Compact-2D designs resulting in net switching power increase. For the cell-dominated benchmark AES-128, the amount of PPA degradation is tolerable as an exchange of more design freedom. However, the inefficiency is considerable in the wire-dominated LDPC benchmark. In LDPC, cell placement has more effects on the routing results than the cell-dominated AES-128, which makes wirelength and net switching power change more sensitive to tier partitioning.

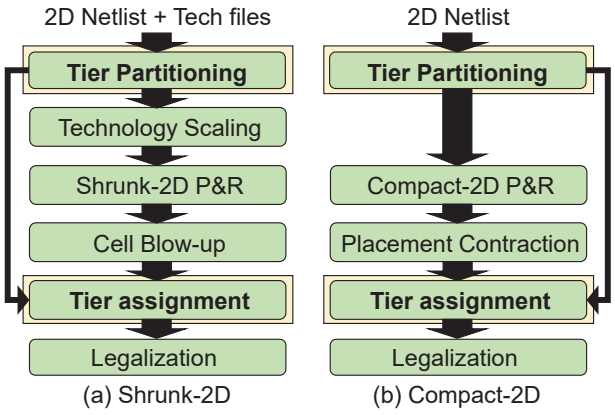

Figure 13: Partition-first 3D placement of (a) Shrunk-2D and (b) Compact-2D design flow.

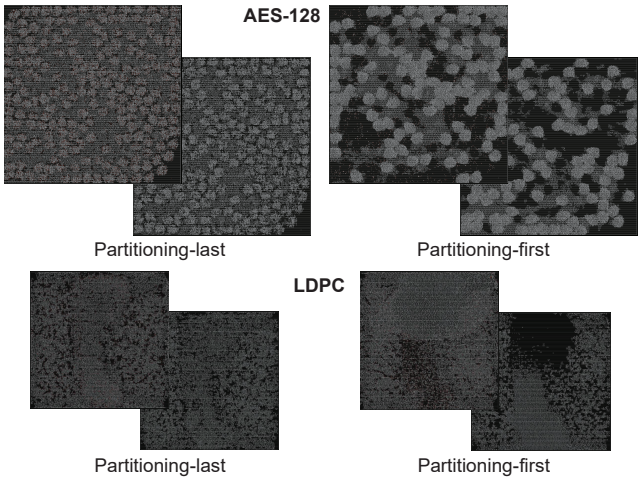

Figure 14: Partitioning-last and partitioning-first placement comparison for AES-128 (top) and LDPC benchmark (bottom).

In summary, although the partitioning-first scheme is not as efficient in PPA as the original partitioning-last scheme in Shrunk-2D and Compact-2D, it brings design freedom to both design flows. Therefore, it is worth choosing the partitioning-first approach for cell-dominated designs where PPA degradation is acceptable.

5 CONCLUSIONS

This paper has provided a comprehensive study on pseudo-3D design flows for monolithic 3D ICs. The current status of the state-of-the-art pseudo-3D design flows were researched and compared thoroughly based on PPA. We also introduced enhancement techniques for each pseudo-3D design flow. Experiments showed that placement-aware MIV planning in Cascade-2D reduced wirelength by up to 26%, reducing power consumption up to 10% from the original flow. Partitioning-first adoption of Shrunk-2D and Compact-2D design also added another dimension of design freedom with tolerable amount of PPA degradation.

Table 6: Comparison of partitioning-first design to original Shrunk-2D and Compact-2D design.

Shrunk-2D	AES-128		LDPC	
	Orig.	PF_FM	Orig.	PF_FM
MIV #	39,521	4,704	15,390	13,529
WL (m)	1.212	1.278 +5.43%	1.226	1.365 +11.36%
WNS (ns)	-0.066	-0.072	-0.086	-0.623
Net Pwr. (mW)	62.05	63.27 +1.96%	41.29	45.24 +9.58%
Cell Pwr. (mW)	181.21	181.92 +0.39%	30.61	31.26 +2.12%
Compact-2D	**AES-128**		**LDPC**	
	Orig.	PF_FM	Orig.	PF_FM
MIV #	39,772	4,984	16,462	13,256
WL (m)	1.224	1.270 +3.74%	1.197	1.266 +5.78%
WNS (ns)	-0.069	-0.077	-0.102	-0.485
Net Pwr. (mW)	62.46	63.58 +1.79%	39.45	41.72 +5.74%
Cell Pwr. (mW)	179.56	179.92 +0.20%	27.20	28.07 +3.19%

6 ACKNOWLEDGEMENT

This research is funded by the DARPA ERI 3DSOC Program under Award HR001118C0096.

REFERENCES

[1] OpenCores benchmark suite. http://www.opencores.org/.
[2] K. Asanovic et al. The Rocket Chip Generator. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2016-17*, 2016.
[3] T. F. Chan et al. mPL6: Enhanced Multilevel Mixed-size Placement. In *ISPD*, pages 212–214, 2006.
[4] K. Chang et al. Cascade2D: A Design-aware Partitioning Approach to Monolithic 3D IC with 2D Commercial Tools. In *ICCAD*, pages 1–8, 2016.
[5] T. Chen et al. NTUplace3: An Analytical Placer for Large-Scale Mixed-Size Designs With Preplaced Blocks and Density Constraints. *TCAD*, 27(7):1228–1240, 2008.
[6] J. Cong et al. Large scale circuit partitioning with loose/stable net removal and signal flow based clustering. In *ICCAD*, pages 441–446, 1997.
[7] J. Cong et al. Thermal-Aware 3D IC Placement Via Transformation. In *ASP-DAC*, pages 780–785, 2007.
[8] J. Cong and S. K. Lim. Edge separability-based circuit clustering with application to multilevel circuit partitioning. *TCAD*, 23(3):346–357, 2004.
[9] C. M. Fiduccia and R. M. Mattheyses. A Linear-Time Heuristic for Improving Network Partitions. In *DAC*, pages 175–181, 1982.
[10] B. Goplen and S. Sapatnekar. Efficient Thermal Placement of Standard Cells in 3D ICs Using a Force Directed Approach. In *ICCAD*, pages 86–89, 2003.
[11] B. Goplen and S. Sapatnekar. Placement of 3D ICs with Thermal and Interlayer Via Considerations. In *DAC*, pages 626–631, 2007.
[12] M. Hsu et al. TSV-Aware Analytical Placement for 3-D IC Designs Based on a Novel Weighted-Average Wirelength Model. *TCAD*, 32(4):497–509, 2013.
[13] D. H. Kim et al. Study of Through-Silicon-Via Impact on the 3-D Stacked IC Layout. *TVLSI*, 21(5):862–874, 2013.
[14] B. W. Ku et al. Compact-2D: A Physical Design Methodology to Build Commercial-Quality Face-to-Face-Bonded 3D ICs. In *ISPD*, pages 90–97, 2018.
[15] J. Lu et al. ePlace: Electrostatics-Based Placement Using Fast Fourier Transform and Nesterov's Method. *TODAES*, 20(2):17:1–17:34, 2015.
[16] J. Lu et al. ePlace-3D: Electrostatics Based Placement for 3D-ICs. In *ISPD*, pages 11–18, 2016.
[17] G. Luo et al. An Analytical Placement Framework for 3-D ICs and Its Extension on Thermal Awareness. *TCAD*, 32(4):510–523, 2013.
[18] S. Panth et al. Design and CAD Methodologies for Low Power Gate-level Monolithic 3D ICs. In *ISLPED*, pages 171–176, 2014.
[19] S. Panth et al. Shrunk-2D: A Physical Design Methodology to Build Commercial-Quality Monolithic 3D ICs. *TCAD*, 36(10):1716–1724, 2017.
[20] S. Samal et al. Monolithic 3D IC vs TSV-based 3D IC in 14nm FinFET Technology. In *S3S*, pages 1–2, 2016.
[21] P. Spindler et al. Kraftwerk2 - A Fast Force-Directed Quadratic Placement Approach Using an Accurate Net Model. *TCAD*, 27(8):1398–1411, 2008.
[22] S. Wong et al. Monolithic 3D Integrated circuits. In *VLSI-TSA*, 2007.