Pseudo-3D Physical Design Flow for Monolithic 3D ICs: Comparisons and Enhancements

HEECHUN PARK, Seoul National University, South Korea
BON WOONG KU, Synopsys Inc., United States
KYUNGWOOK CHANG, Sungkyunkwan University, South Korea
DA EUN SHIM and SUNG KYU LIM, Georgia Institute of Technology, United States

Studies have shown that **monolithic 3D** (**M3D**) ICs outperform the existing **through-silicon-via** (**TSV**) -based 3D ICs in terms of **power, performance, and area** (**PPA**) metrics, primarily due to the orders of magnitude denser vertical interconnections offered by the nano-scale monolithic inter-tier vias. In order to facilitate faster industry adoption of the M3D technologies, physical design tools and methodologies are essential. Recent academic efforts in developing an EDA algorithm for 3D ICs, mainly targeting placement using TSVs, are inadequate to provide commercial-quality GDS layouts. Lately, *pseudo-3D* approaches have been devised, which utilize commercial 2D IC EDA engines with tricks that help them operate as an efficient 3D IC CAD tool. In this article, we provide thorough discussions and fair comparisons (both qualitative and quantitative) of the state-of-the-art pseudo-3D design flows, with analysis of limitations in each design flow and solutions to improve their PPA metrics. Moreover, we suggest a hybrid pseudo-3D design flow that achieves both benefits. Our enhancements and the inter-mixed design flow, provide up to an additional 26% wirelength, 10% power consumption, and 23% of power-delay-product improvements.

CCS Concepts: • **Hardware → 3D integrated circuits**; **Physical design (EDA)**; **Software tools for EDA**;

Additional Key Words and Phrases: Monolithic 3D IC, pseudo-3D approach, computer-aided design, 3D placement, timing closure

ACM Reference format:
Heechun Park, Bon Woong Ku, Kyungwook Chang, Da Eun Shim, and Sung Kyu Lim. 2021. Pseudo-3D Physical Design Flow for Monolithic 3D ICs: Comparisons and Enhancements. *ACM Trans. Des. Autom. Electron. Syst.* 26, 5, Article 37 (May 2021), 25 pages.
https://doi.org/10.1145/3453480

1 INTRODUCTION

Monolithic 3D (**M3D**) IC [4, 27] is introduced to maximize the benefits of 3D IC technologies with its **monolithic inter-tier vias** (**MIVs**) for the cross-tier connections. MIV overcomes the

This work was supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2021.
Authors' addresses: H. Park, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea; email: phcssl0112@gmail.com; B. W. Ku, Synopsys Inc., 690 E Middlefield Rd, Mountain View, CA 94043, United States; email: bon.ku@synopsys.com; K. Chang, Sungkyunkwan University, 2066, Seobu-ro, Jangan-gu, Suwon-si, Gyeonggi-do 16419, South Korea; email: k.chang@skku.edu; D. E. Shim and S. K. Lim, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30332, United States; emails: daeun@gatech.edu, limsk@ece.gatech.edu.

1084-4309/2021/05-ART37 $15.00
https://doi.org/10.1145/3453480

inefficiency of the existing **through-silicon-via** (**TSV**) -based 3D ICs by replacing the large-area, high-cost TSVs with the nano-scale MIVs. MIVs enable denser via insertion and thus offer more design freedoms. A collaborative study between an academic institute and a major foundry have shown that M3D outperforms TSV-based 3D IC designs in terms of **power, performance, and area** (**PPA**) qualities using commercial RTLs and commercial libraries [22, 23]. The last crux of M3D is the hardness of low temperature process (~400°C) for the top tier metal stacks. Nevertheless, there are various efforts [24, 25, 28] and prototypes [1] to realize the M3D IC fabrication, which are guiding us toward a bright future. In order to facilitate faster industry adoption of the M3D technologies, physical design tools and methodologies are essential.

Placement was the main focus of early day development of 3D IC physical design tools. In the 2000s, several research groups from academia have presented placement algorithms mainly targeting TSV-based 3D ICs. Goplen and Sapatnekar proposed an extension of the force-directed method to handle 3D placement [13], and the recursive bisection method by adding z-axis cuts to the algorithm [14]. Cong et al. [9] proposed 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.

In the early 2010s, more aggressive approaches have been presented to solve the 3D placement problem, still targeting TSV-based 3D ICs. These works extend the existing 2D analytic placers [5, 7, 19, 26] by considering z-dimension as the new freedom to further improve wirelength and congestion while minimizing overlap among cells and TSVs [11, 15, 16, 20].

Starting in the mid 2010s, another approach, so-called *pseudo-3D* design flows, made its debut [6, 18, 21]. Pseudo-3D approaches utilize commercial 2D CAD engines to produce 3D IC designs. Unlike the previous academic 3D IC placers (*true-3D*), these approaches handle routing and timing closure as well. A common theme among these pseudo-3D approaches is that they first build 2D physical design using commercial 2D IC physical design tools. Next, they transform this intermediate 2D layout into 3D layout using various techniques. Pseudo-3D approaches concentrate on seamless utilization of 2D commercial tools to produce 3D IC design, which is the main reason their PPA qualities outperform those of commercial 2D counterparts.

In this article, we discuss the status quo and provide thorough comparisons of various pseudo-3D design flows, which include both qualitative and quantitative points of view. We also analyze the limitations of each design flow and provide solutions to improve them. Moreover, we introduce a hybrid pseudo-3D design flow that takes advantage from all design approaches. Our works provide up to an additional 26% wirelength, 10% power consumption, and 20% of **power-delay-product** (**PDP**) improvements.

2 CURRENT STATUS OF TRUE-3D PLACERS

In this section, we discuss the current status of true-3D design flows. Several academic research groups have proposed their own true-3D design flows [11, 13, 15, 16, 20]; 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 each tier of the 3D design separately. The flowchart on the left of 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 HPWL, $\#TSV$ is the number of TSVs, α is a weight number constant, $\rho_{b,l}$ and $D_{b,l}$ represent the total density of cells and the maximum allowable density in bin b

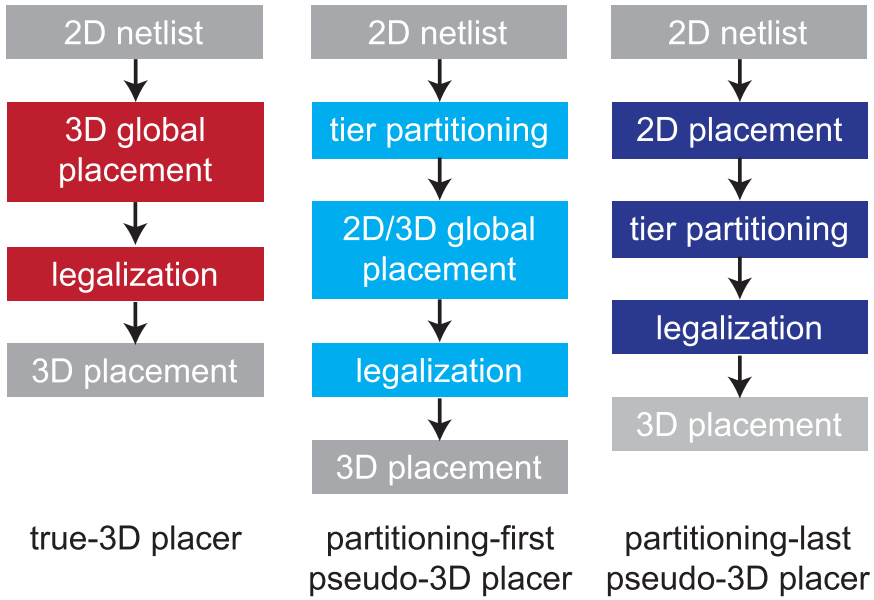

Fig. 1. Three approaches to 3D placement.

of tier t, respectively. After the placement result is generated, TSVs are assigned along with the inter-tier nets (named as *3D nets*) which connect cells on different tiers. This introduces additional overlaps between cells and TSVs which are fixed during the last legalization stage. TSV-related constraints (e.g., TSV-related mechanical stress, electrical coupling, thermal coupling, keep-out zone) are taken into account during or before the TSV insertion to allow enough spaces for TSVs after the 3D placement.

True-3D placers, unfortunately, are not ready for commercial adoption. True-3D placers from academic research groups are standalone and not fully integrated with the entire commercial physical design flow. This means they are only able to report wirelength estimation based on the HPWL. Although some of them made efforts to utilize a commercial routing engine to calculate the actual value of wirelength of the full 3D design, it is still not practical since the routing cannot take the real timing constraints into consideration and thus the result is not as accurate as the commercial quality. In addition, other essential stages for the commercial adoption (e.g., **clock tree synthesis** (**CTS**), timing closure) are missing in the true-3D design flows.

As an effort to make a fair comparison, we integrated one of the true-3D placers [16], which we named *Force-3D*, into the commercial EDA flow.[1] Force-3D is a 3D expansion of the 2D force-directed quadratic placement solution [26]. It first partitions the 2D netlist into 3D, then adds MIVs along 3D nets. Lastly, it performs 2D force-directed analytic placement with each MIV modeled as a small cell in the top tier. While the original work targets TSV-based 3D ICs, we applied it to the M3D ICs by removing all TSV-related constraints (e.g., TSV-related mechanical stress, electrical

[1]Our efforts to integrate 3D placers from other groups into our commercial pseudo-3D flows were not successful for the following reasons: (1) source codes were not available, (2) binaries were not compatible with commercial EDA tools, and (3) the placers were not compatible with commercial **physical design kits** (**PDKs**) and libraries such as a foundry 28 nm we are using in this article.

Table 1. PPA Comparison of a True-3D Placer (Force-3D [16]) with Post-Placement Stages and a Pseudo-3D Design Flow (Shrunk-2D [21])

		True-3D [16]	Pseudo-3D [21]
AES-128 (3.4 *GHz*)	Cell area (μm^2)	123,526	119,127 −3.56%
	WL (m)	1.495	1.212 −18.93%
	WNS (ns)	−0.077	−0.066
	Power (mW)	274.16	268.00 −2.24%
	PDP ratio	1	**0.95**
LDPC (1.2 *GHz*)	Cell area (μm^2)	68,019	48,752 −28.33%
	WL (m)	2.132	1.226 −42.50%
	WNS (ns)	−0.195	−0.086
	Power (mW)	139.31	81.91 −41.20%
	PDP ratio	1	**0.53**
Nova (625 *MHz*)	Cell area (μm^2)	199,444	187,331 −6.07%
	WL (m)	2.870	2.145 −25.26%
	WNS (ns)	−0.271	−0.193
	Power (mW)	121.60	110.74 −8.93%
	PDP ratio	1	**0.87**
TATE (1.4 *GHz*)	Cell area (μm^2)	243,201	237,869 −2.19%
	WL (m)	2.884	1.853 −35.75%
	WNS (ns)	−0.007	−0.098
	Power (mW)	336.69	318.99 −5.26%
	PDP ratio	1	**1.06**
ECG (1.0 *GHz*)	Cell area (μm^2)	106,810	106,866 +0.00%
	WL (m)	1.273	0.909 −28.59%
	WNS (ns)	-0.050	−0.040
	Power (mW)	107.53	105.70 −1.70%
	PDP ratio	1	**0.97**

coupling, thermal coupling, keep-out zone).[2] For the integration, we extract the 3D placement results from the in-house Force-3D placement tool and perform all post-placement stages (i.e., pre-CTS optimization, CTS, post-CTS optimization, route, and timing closure) for each tier with a commercial 2D **place-and-route** (**P&R**) tool (Cadence Innovus). It enables the sign-off analysis and GDS layout extraction for Force-3D design results.

Table 1 shows a comparison between a true-3D (Force-3D) and a pseudo-3D design flow (Shrunk-2D [21]).[3] The true-3D design flow shows worse results in all factors of cell area, **worst negative slack** (**WNS**), power, and also PDP compared to the pseudo-3D design results. Even for the exception case with better PDP (TATE), pseudo-3D design shows smaller cell area, wirelength, and power consumption than the true-3D design. In-house tool-based true-3D placers are not based

[2]The article [16] provided two different design flows according to when to decide the TSV locations (*TSV co-placement* and *TSV site*). We chose TSV co-placement as Force-3D, which is more suitable as it irregularly locates TSVs, i.e., MIVs.

[3]Our attempts to use the benchmarks reported in academic placers were not successful. This is primarily because the underlying PDK/library information for those gate-level netlists are missing. This makes routing, timing closure, GDS construction, and sign-off PPA calculation impossible. In our experiments, we use open-source RTLs and synthesize them using a commercial 28-nm PDK/library to overcome these limitations and improve the industry relevance and practicality of our results.

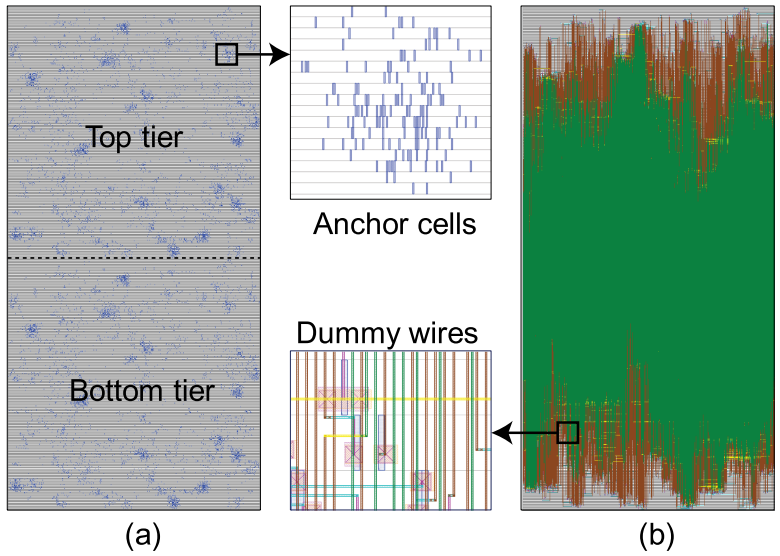

Fig. 2. Cascade-2D [6] intermediate layouts. (a) After anchor cell insertion. (b) After routing anchor cells using dummy wires.

on the commercial design background and thus provide inefficient designs when combined with commercial EDA flow. Therefore, we only focus on CAD-tool-friendly pseudo-3D design flows in this article.

3 SUMMARY OF PSEUDO-3D DESIGN FLOWS

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 design flows: Cascade-2D [6], Shrunk-2D [21], and Compact-2D [18].[4] Cascade-2D is classified as a *partitioning-first* design flow, which performs tier partitioning before performing a global placement. On the other hand, Shrunk-2D and Compact-2D belong to *partitioning-last* design flows which perform tier partitioning after an intermediate placement stage, thereby exploiting placement results to generate more specific partition results. The general differences between 3D placements are shown in Figure 1.

3.1 Cascade-2D Flow

The main motivation of Cascade-2D is to place, route, and close timing of all tiers simultaneously, rather than performing tier-by-tier designs separately. Since 2D CAD tools cannot restore all tiers of a 3D design at the same time, Cascade-2D uses an intermediate 2D structure with a vertically long plane (W:H = 1:2) to represent a 3D design 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 anchor cells at the MIV position in both tiers (Figure 2(a)) and a zero-parasitic dummy wire that connects the anchor cells (Figure 2(b)). These two anchor cells have an identical xy coordinate when we stack up the top and bottom tiers, and become an MIV at last.

Figure 3 shows the overall Cascade-2D design flow. It uses a tier partitioning algorithm to split the 2D netlist into a 3D netlist with two tiers, say top and bottom. Then, all 3D nets are disconnected

[4]There are other pseudo-3D design flows being proposed constantly (e.g., [17]), while we focused on these three flows, since they serve as a pioneer or representative work for each approach.

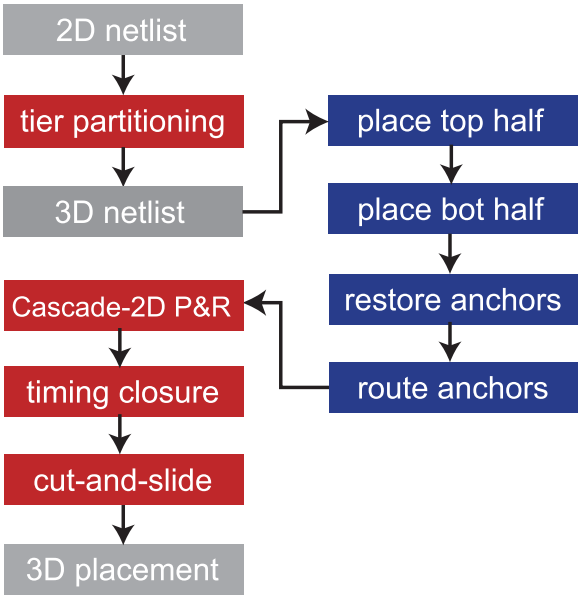

Fig. 3. Overview of Cascade-2D flow [6].

and represented as I/O pins for each tier, and we manually attach a logically dummy anchor cell to each of those I/O pins.

Then, we perform sequential 2D placements of the top and bottom tiers to deduce the anchor cell locations using a commercial 2D placement engine. In detail, during the top tier placement, the CAD tool locates all top-tier cells including the driving anchor cells in the top tier (top → bottom). The anchor cell locations are referred to the following bottom tier placement stage by placing the receiving anchor cells at the same $x - y$ positions of the driving anchor cells. Then, the bottom tier anchor cells guide the bottom tier cells to be placed nearby their corresponding anchor cells. Remaining anchor cells (for bottom → top 3D nets) are placed in a similar way. Then we generate a 2D canvas of 1:2 aspect ratio (Figure 2(a)) and restore anchor cells to top and bottom halves, while all other cell placements are discarded. These anchor cells are routed in pairs with zero-parasitic dummy wires (Figure 2(b)). They only occupy the upper metal layers that are not used by the actual design to avoid interference, and zero-parasitic represent the instant inter-tier connection through MIV.

After all anchor cells and dummy wires are set, we use the commercial 2D CAD tool to perform traditional 2D P&R including timing closure for the rest of cells and wires, as if the tool is running on a 2D design. A hard horizontal fence is placed across the horizontal center to forbid cross-half (i.e., cross-tier) cell movement during the entire flow. Lastly, the final result is transformed to a monolithic 3D design by laying the top half upon the bottom half and replacing each dummy wire to a vertical MIV with a certain RC value.

3.2 Shrunk-2D Flow

The main idea behind Shrunk-2D flow is to fully utilize the commercial 2D P&R tool in the placement of xy coordinates, and then take care of z coordinates (i.e., tier assignment). To achieve this, for an n-tier M3D design, it uses the commercial CAD tool to generate an intermediate placement

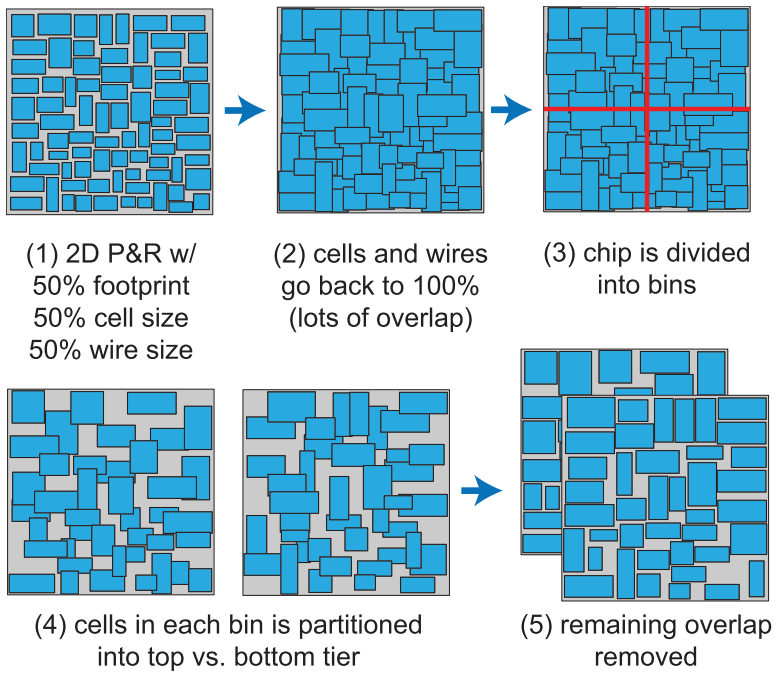

Fig. 4. Illustration of 3D placement in Shrunk-2D flow [21].

result in which all cells, wires, and the total footprint are shrunk to $1/n$ in size compared to its original 2D design.

Figure 4 shows a conceptual view of the 3D placement of Shrunk-2D design flow for a two-tier M3D design. In the intermediate Shrunk-2D P&R result from the commercial 2D CAD tool, all standard cells and wires are shrunk to 50% and its total footprint also has 50% area of its 2D design counterpart (Figure 4(1)). All components are then expanded to their original size, which creates many overlaps as in Figure 4(2). It is then divided into several square bins (Figure 4(3)), and the Fiduccia-Mattheyses (FM) min-cut partitioning algorithm[12] is applied to each bin to evenly partition the cells into both tiers (Figure 4(4)). Remaining cell overlaps in both tiers are removed by tier-by-tier legalization using the commercial tool (Figure 4(5)).

Figure 5 shows the overview of the whole Shrunk-2D design flow. Technology scaling is performed at first to modify **library exchange format** (**LEF**) and **liberty** (**LIB**) files to represent shrunk cells. Then we run the 2D P&R tool based on the downscaled technology files to generate an intermediate Shrunk-2D P&R result, as if it is running for a traditional 2D design with a smaller technology. Then, LEF/LIB blow-up, tier partitioning, and legalization is done as illustrated in Figure 4.

After that, routing is performed in stages from MIV planning to tier-by-tier detailed routing. MIV planning uses a simplified netlist that only includes 3D nets and a modified 2× metal stack as in Figure 6(a). Top and bottom tier cells have their pins at the top and bottom half metal layers, respectively. Then, the CAD tool is used to route 3D nets, and each location of the middle vias (i.e., between the sixth and seventh metal layers in Figure 6(b)) is converted to an MIV location, while discarding other results. MIV locations become I/O pins for both tiers, and all wires are rerouted tier-by-tier in detail using the CAD tool in detailed routing stage. Timing closure is also performed in a per-tier fashion, which only refers to the design constraints from the initial timing analysis.

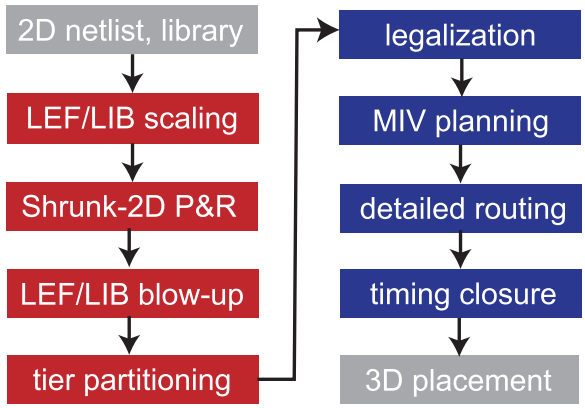

Fig. 5. Overview of Shrunk-2D flow [21].

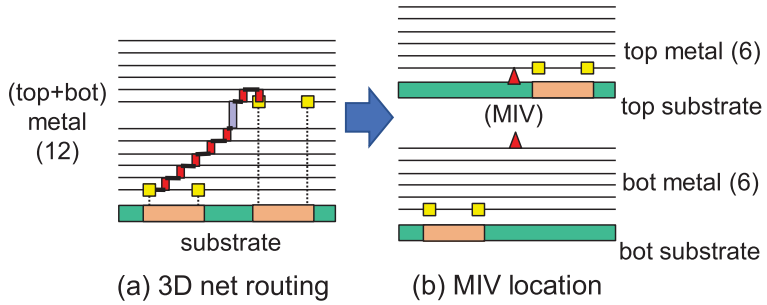

Fig. 6. Illustration of MIV planning.

The problem on such per-tier timing closure is discussed in Section 4.1. Both tiers are reconnected by MIVs at last for the final timing and power analysis.

The important point is that, every step in the Shrunk-2D design flow is aided by commercial 2D CAD tools, except for manual technology scaling and tier partitioning. This means that Shrunk-2D flow fully utilizes the 2D CAD tool for 3D design and generates an efficient M3D design output.

3.3 Compact-2D Flow

Shrunk-2D flow has a fatal risk of shrinking technology smaller than the CAD tool can handle. When we try to apply Shrunk-2D flow with the smallest technology node, say 7 *nm* in the academic license, it is necessary to scale the technology down to a smaller node, say 5*nm*. Then CAD tool (i.e., Cadence Innovus) is impossible to operate the Shrunk-2D P&R stage. Even if the technology node is in a possible range to be operated, the scaled values have to be rounded up or down to the precision that the tool can handle (i.e., MANUFACTURINGGRID in LEF file). This makes Shrunk-2D P&R to run with less accurate technology values, and thus suffer from sub-optimal result and unexpected design rule violations. Compact-2D [18] flow is proposed to resolve this issue by using similar concepts but *not shrinking cells and wires*. Instead, it scales only the RC parasitics by a factor of $1/\sqrt{n}$ for an n-tier M3D design to make the CAD tool treat the interconnect faster than that of the original 2D design. Therefore, a Compact-2D P&R result has smaller path delays compared to the same path in 2D design, which requires less timing closure overhead. When compressing the design footprint into $1/n$, the interconnects return to their original parasitic values and the timing closure resources become sufficient in the end. Compact-2D flow achieves the identical

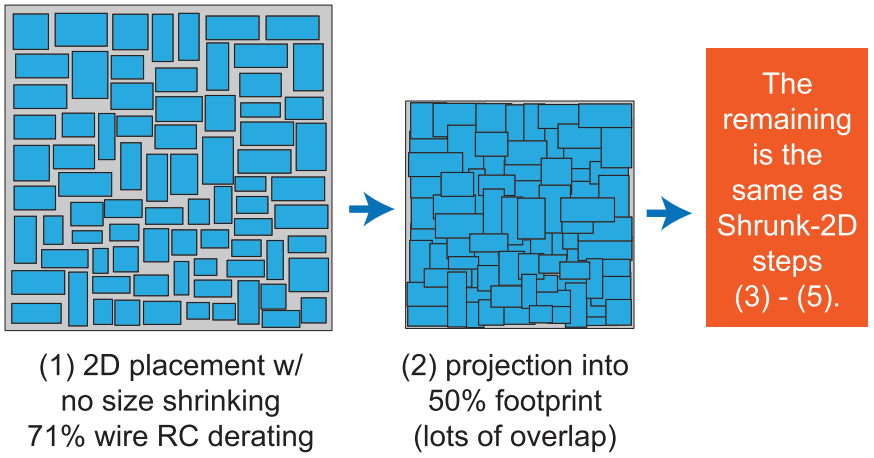

Fig. 7. Illustration of 3D placement in Compact-2D flow [18].

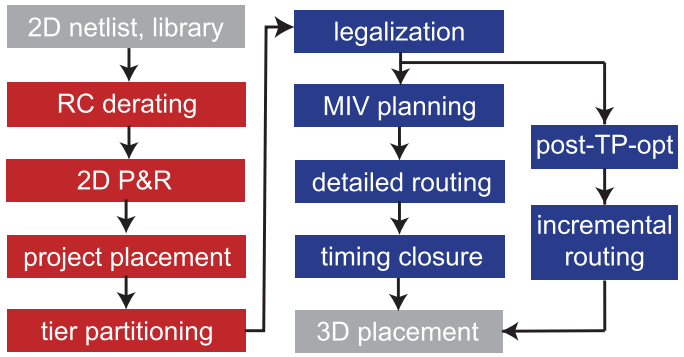

Fig. 8. Overview of Compact-2D flow [18].

effect of shrinking everything without actually shrinking them, and thus 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 3D placement in Compact-2D flow is shown in Figure 7. The intermediate Compact-2D P&R result has the *same* area and cell size as its 2D counterpart, and the main differences are that all wires have smaller RC parasitic values, and that the amount of optimization resources is reduced. In the next stage, it contracts the chip area to 50% and project cells/wires into the downscaled footprint to generate 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 parasitics are scaled up to their original sizes at the same time, overall interconnect parasitics remain consistent.

The rest of the steps are the same as those in the Shrunk-2D flow (i.e., bin-based tier partitioning, legalization). Figure 8 recites the whole Compact-2D flow. It is conceptually similar to the Shrunk-2D design flow, except that Compact-2D flow only adjusts the RC parasitics scaling factor instead of the whole technology files.

In addition, **post-tier-partitioning optimization** (**post-TP-opt**) and incremental routing are introduced in [18] to further improve the M3D design quality. Figure 9 shows the conceptual view. After importing both tiers into the modified metal stack (Figure 9(a)), there are overlaps between

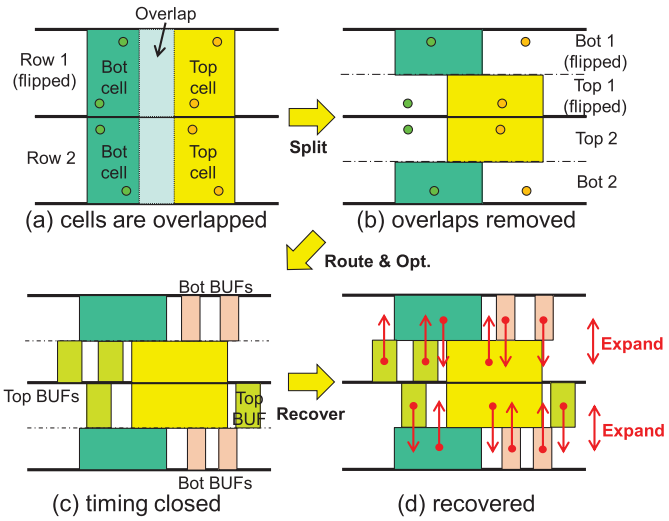

Fig. 9. Illustration of post-tier-partitioning optimization (post-TP-opt).

the top and the bottom tier cells, and it is impossible to run a timing optimization using an existing 2D CAD tool because the tool does not allow overlaps between cells during timing closure (i.e., buffer insertion, cell movement, upsizing or downsizing). To resolve this issue, each standard cell row is split into two rows in that one split row represents one of two tiers, and all standard cells are placed to their corresponding rows by halving their heights while their pins stay as they were (i.e., outside the cell boundary), as in Figure 9(b). Pins do not overlap as they are located at different metal layers. The concept of flipped standard cells for power grid sharing are also covered by alternating the order of representation such as "top-bottom-bottom-top-…." Now the CAD tool is able to insert resources without considering overlaps (Figure 9(c)), and they are recovered by expanding all cells to their original sizes at the last step (Figure 9(d)). Post-TP-opt routes and optimizes the design with all nets, instead of only 3D nets as in MIV planning, and the routing results are preserved until the end by incremental routing.

While post-TP-opt allows simultaneous timing closure of Compact-2D design, it is an aggressive scheme that requires modification of foundry PDKs into an intolerable shape for the CAD tool. Hence, it encounters fatal errors if used with some delicate PDKs or design benchmarks.

4 COMPARISONS OF PSEUDO-3D FLOWS

This section provides a thorough comparison of pseudo-3D design flows, which includes both qualitative and quantitative factors. For the PPA comparison, we used various types of benchmarks with both gate-level designs and a mixed-size design with macro blocks. We also add a subsection for the summary of all flows.

4.1 Qualitative Comparisons

Table 2 lists the characteristics of the three different pseudo-3D design flows. For Compact-2D, we compare both design flows with original MIV planning followed by tier-by-tier routing and post-TP-opt with incremental routing.

In the beginning, Cascade-2D only needs to prepare 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

Table 2. Qualitative Comparison of Pseudo-3D Design Flows

	Cascade-2D	Shrunk-2D	Compact-2D	
			Orig.	Post-TP-opt
Modified tech files	Anchor cells (Easy)	Cells & wire (Hard)	RC factor (Easy)	+Half cells (Hard)
Macro block	Yes	Partially	Partially	
Partitioning	Beginning	Middle	Middle	
MIV count	Small	Large	Large	Larger
Timing closure	Simultaneous	Per-tier	Per-tier	Simultaneous

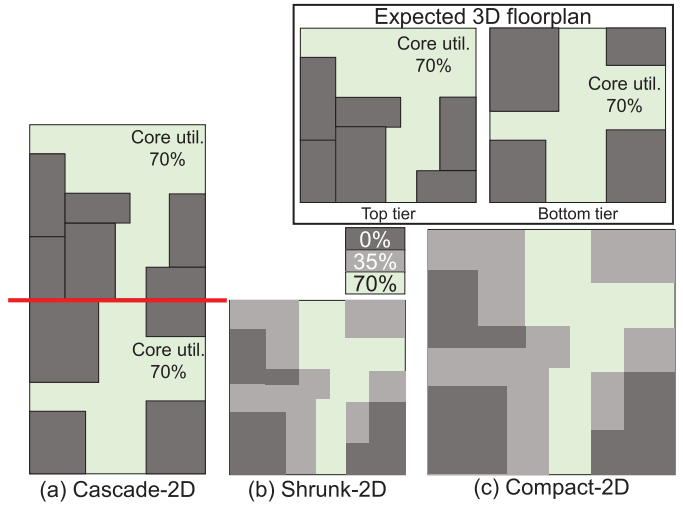

Fig. 10. Illustration of macro block floorplans in the pseudo-3D design flows.

solved this issue by only scaling the interconnect RC factor. Compact-2D with post-TP-opt, on the other hand, still requires another LEF set of half-height standard cells, each of which has some pins staying at the outside of the cell boundary. This becomes a major drawback of post-TP-opt that using such LEFs sometimes meet unexpected faults during the optimization step with CAD tool.

All flows are able to take care of macro blocks by assigning their tiers and placing them before CAD-tool-based placement stage. While Cascade-2D does not have exceptional constraints for the macro blocks, Shrunk-2D and Compact-2D add 50% partial placement blockages to the partially blocked region (macro on one tier) before Shrunk-2D and Compact-2D P&R stage. Figure 10 shows an example of macro block floorplanning in the pseudo-3D flows, and we set the core utilization of each tier as 70%. Cascade-2D flow simply imports each block to its corresponding half (Figure 10(a)), and runs the rest of the steps without any extra concerns. However, Shrunk-2D and Compact-2D have a complex blockage map with mixtures of full (0%) and partial blockages (35%). In the following standard cell placement, the CAD tool tends to assign a smaller amount of cells into the partial blockages than they can actually accommodate (50%), and thus generates a sub-optimal placement result.

As for tier partitioning, Cascade-2D performs it separately before the CAD tool flow, and gives the designer the authority of manual tier assignment. It makes the usage of MIVs in Cascade-2D

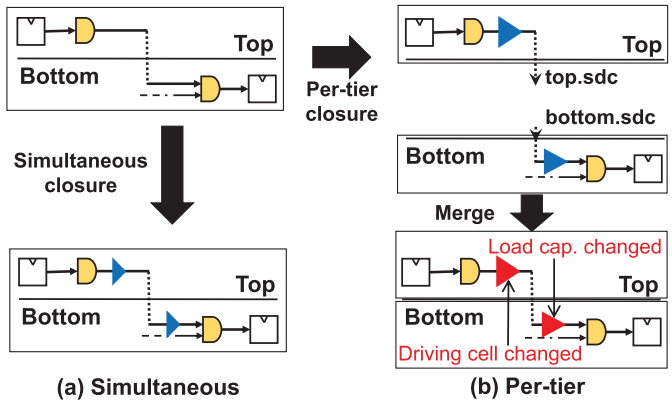

Fig. 11. Example of (a) simultaneous timing closure and (b) per-tier timing closure.

flow relatively small, since the tendency of most tier partitioning algorithms is to be based on minimizing cut size, which goes to minimal 3D net count and minimal MIV count at last. On the other hand, Shrunk-2D and Compact-2D flows merged the bin-based FM min-cut tier partitioning method into their pseudo-3D placement step to pursue better quality. Bin-based tier partitioning makes the resultant cut size larger with more balanced local area skew. Post-TP-opt in Compact-2D flow changes the tier assignment of cells for design optimization, which further increases the final MIV count.

Before explaining the last factor, Figure 11 shows an example of simultaneous and per-tier timing closure for an inter-tier timing path. For simultaneous timing closure (Figure 11(a)), buffer(s) will be inserted to close timing. This is as accurate as the 2D timing closure since the entire path across both tiers is considered. For per-tier timing closure (Figure 11(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. In short, it does not guarantee the timing closure of the merged path because separate timing closures have 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. While Cascade-2D made it possible by bringing both tiers into a 1:2 2D canvas, Shrunk-2D and Compact-2D do not have such a trick and therefore suffer from a large negative slack. Compact-2D with post-TP-opt is able to simultaneously close timing by using half-height cells, though it is risky to be applied to some PDKs and design benchmarks.

4.2 Quantitative Comparisons: Gate-Level Designs

We compare different pseudo-3D design flows in terms of PPA. In this subsection, design results of 2D and M3D (from Cascade-2D, Shrunk-2D, and Compact-2D design flows) are compared using five flattened gate-level benchmark circuits from OpenCore benchmark suites [2], with a commercial-grade 28-nm PDK from a semiconductor foundry. MIV diameter and RC values are set to $100nm$, 64Ω and $0.2fF$. Cadence Innovus and Cadence Tempus are used for P&R and timing/power analysis of normal 2D design flow and CAD-tool-based stages in pseudo-3D design flows.

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 to have half area.

Table 3. Comparison of Commercial 2D and Pseudo-3D Design Flows (Cascade-2D [6], Shrunk-2D [21], and Compact-2D [18]) **and the Proposed Inter-mixed Flow (Details in Section 5.4)**

		2D	Cascade-2D	Shrunk-2D	Compact-2D		**Inter-mixed**
					Orig.	Post-TP-opt.	
AES-128 (3.4 GHz)	Cell area (μm^2)	123,202	123,915 +0.58%	119,127 −3.31%	118,719 −3.64%	131,055 +6.37%	**118,603 −3.73%**
	#MIV	0	1,976	39,521	39,772	70,823	**28,239**
	WL (m)	1.444	1.694 +17.29%	1.212 −16.09%	1.224 −15.24%	1.229 −14.91%	**1.343 −7.02%**
	WNS (ns)	−0.009	−0.013	−0.066	−0.069	+0.0002	**−0.010**
	Power (mW)	274.21	292.33 +6.61%	268.00 −2.27%	266.58 −2.78%	287.22 +4.74%	**270.07 −1.51%**
	PDP ratio	1	1.081	1.162	1.166	1.016	**0.990**
LDPC (1.2 GHz)	Cell area (μm^2)	46,958	47,103 +0.31%	48,752 +3.82%	45,313 −3.50%	46,879 −0.17%	**49,022 +4.39%**
	#MIV	0	5,996	15,390	16,462	26,682	**8,212**
	WL (m)	1.527	1.707 +11.81%	1.226 −19.69%	1.197 −21.61%	1.160 −24.05%	**1.918 +25.63%**
	WNS (ns)	−0.044	−0.056	−0.086	−0.102	+0.005	**−0.038**
	Power (mW)	85.63	107.83 +25.92%	81.91 −4.35%	74.91 −12.52%	79.28 −7.41%	**107.70 +25.77%**
	PDP ratio	1	1.277	1.002	0.933	0.874	**1.249**
Nova (625MHz)	Cell area (μm^2)	188,260	188,006 −0.13%	187,331 −0.49%	187,450 −0.43%	188,049 −0.11%	**187,248 −0.54%**
	#MIV	0	264	33,980	33,658	54,354	**21,462**
	WL (m)	2.257	2.571 +13.91%	2.145 −4.96%	2.120 −6.08%	2.029 −10.12%	**2.830 +25.36%**
	WNS (ns)	−0.060	−0.054	−0.193	−0.205	0.000	**−0.043**
	Power (mW)	110.54	113.47 +2.66%	110.74 +0.18%	110.75 +0.19%	109.54 −0.90%	**112.41 +1.69%**
	PDP ratio	1	1.023	1.082	1.090	0.955	**1.007**
TATE (1.4 GHz)	Cell area (μm^2)	238,023	217,199 −8.75%	237,869 −0.06%	237,714 −0.13%	238,358 +0.14%	**237,449 −0.24%**
	#MIV	0	3,199	52,442	52,672	99,287	**21,553**
	WL (m)	1.935	3.141 +62.38%	1.853 -4.23%	1.843 −4.73%	1.844 −4.68%	**2.766 +42.98%**
	WNS (ns)	−0.012	−0.026	−0.098	−0.129	+0.061	**+0.013**
	Power (mW)	315.44	319.42 +1.26%	318.99 +1.13%	318.53 +0.98%	318.80 +1.06%	**331.24 +5.01%**
	PDP ratio	1	1.032	1.128	1.169	0.912	**1.015**
ECG (1.0 GHz)	Cell area (μm^2)	106,965	96,944 −9.37%	106,866 −0.09%	106,927 −0.03%	107,171 +0.19%	**106,590 −0.35%**
	#MIV	0	1,010	21,384	21,270	40,196	**15,225**
	WL (m)	0.989	1.253 +26.59%	0.909 −8.13%	0.900 −9.04%	0.906 −8.43%	**1.296 +30.94%**
	WNS (ns)	−0.037	−0.020	−0.040	−0.056	+0.065	**+0.112**
	Power (mW)	104.56	104.24 −0.30%	105.70 +1.09%	105.42 +0.83%	105.42 +0.83%	**109.62 +4.84%**
	PDP ratio	1	0.981	1.013	1.027	0.909	**0.898**

% difference is calculated w.r.t. 2D design result (WL = wirelength, WNS = Worst negative slack).

For the tier partitioning algorithm in Cascade-2D, we use existing partitioning algorithms [8, 10, 12] and choose a design with the least PDP value. For the bin-based tier partitioning algorithm in Shrunk-2D and Compact-2D, we set the bin size as the average net length of the intermediate P&R result. It makes the algorithm have a high tendency to split cells connected by one net into different tiers, and thus uses MIVs aggressively for design optimization. Note that the last column shows the results of an inter-mixed design flow, which will be explained and analyzed in detail later in Section 5.4. Before the detailed analysis, note that all M3D designs already have 50% footprint benefit compared to their 2D counterparts.

Cascade-2D is the only pseudo-3D design flow that matches the timing closure of the commercial 2D design flow and it shows a remarkable reduction in standard cell usage for some benchmarks (TATE, ECG), which is achieved by simultaneous timing closure to remove more

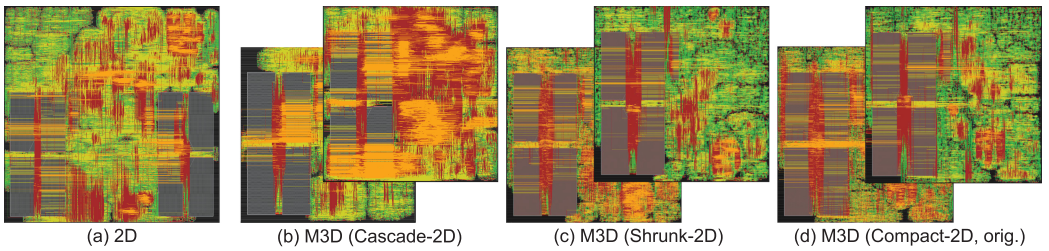

Fig. 12. GDSII layout images of RISC-V Rocketcore processor in 2D and M3D designs with different pseudo-3D design flows.

timing violations and exclude redundant optimization resources. However, its wirelength and power results are worse for most benchmarks, even compared to the 2D design counterpart. The main reason is that its tier partitioning and MIV planning strategies are not well coupled with the CAD-tool-based stages. It results in excessively detoured 3D net wires and increased net switching power consumption. When the design hierarchy is analyzed thoroughly and a delicate tier partitioning method is applied, Cascade-2D shows significant improvements even on wirelength and power consumption, which is verified in [6].

Shrunk-2D and Compact-2D design results show less wirelength and similar or less power compared to the 2D designs for most cases, and it is more obvious in the wire-dominated LDPC benchmark design. They successfully reduce wirelength by downscaling the HPWL and using more MIVs to convert long 2D interconnects into short vertical 3D nets. However, their results have larger negative slacks compared to 2D and Cascade-2D design results, which are even more than 20% of the target clock period for some benchmarks (e.g., AES-128). It means that per-tier timing closure is not enough to cover the timing violations in the M3D designs.

Table 3 also shows the design results from post-TP-opt of Compact-2D, next to the column with original Compact-2D results. We delicately modified the commercial-grade 28-*nm* PDK to generate the LEF file with half standard cell heights and a Quantus RC tech file (*qrcTechFile*) for the 3D metal stacks without touching any core values of the PDK (i.e., parasitics, etc.), and performed the pseudo-3D design flows with post-TP-opt and incremental routing. It shows that, for all benchmark circuits, post-TP-opt makes the worst slack as zero or positive, with the expense of extra MIVs during the optimization. Post-TP-opt has more opportunity to utilize the MIV usage and reduce more slacks than Cascade-2D timing closure because it has an additional option of changing the tier assignment of cells. With a little increase on the power consumption in most cases because of the additional optimization resources, the PDP value decreased for all benchmarks compared to the original pseudo-3D design result as well as the 2D design counterpart.

4.3 Quantitative Comparisons: Mixed-Size Design

We apply pseudo-3D design flows on a mixed-size design benchmark with memory macro modules. Figure 12 shows the layouts of RISC-V Rocketcore processor [3] with 64 KB L1 cache memory, using 2D and all pseudo-3D design flows. There are 10 memory macro blocks generated from the commercial memory compiler; 5 of them belong to L1 data cache and the others to L1 instruction cache. In the M3D designs (Figure 12(b)–(d)), Instruction cache blocks are assigned to the top tier and data cache blocks to the bottom tier. For a fair comparison, we overlap macro blocks in different tiers as much as possible to occupy identical regions in xy coordinates, to minimize the effects of partial blockages in Shrunk-2D and Compact-2D flows. For tier partitioning in Cascade-2D, we applied the clustering-based LR algorithm [8] that gathers cells related to each other in the same

Table 4. Comparison of Commercial 2D and Pseudo-3D Design Flows on RISC-V Rocketcore Processor **(Details of Inter-mixed Flow (Mixed) are in Section 5.4)**

	Rocketcore (1.67 GHz)	2D	Cascade-2D		Shrunk-2D		Compact-2D				**Inter-mixed**	
							Orig.		Post-TP-opt.			
Design metrics	Footprint (mm^2)	0.563	0.281	−50%	0.281	−50%	0.281	−50%	0.281	−50%	**0.281**	**−50%**
	Cell area (mm^2)	0.238	0.241	+1.0%	0.234	−1.9%	0.231	−3.1%	0.241	+1.3%	**0.237**	**−0.35%**
	WL (m)	2.896	3.517	**+21.4%**	2.542	**−12.2%**	2.592	**−10.5%**	2.507	**−13.4%**	**3.546**	**+22.4%**
	MIV Count	0	1,313		46,088		45,624		82,516		**13,349**	
Performance	WNS (ns)	−0.046	**−0.033**		**−0.197**		**−0.320**		**−0.029**		**−0.042**	
	TNS (ns)	−4.996	**−12**		**−606**		**−1091**		**−11**		**−6**	
	TPS (ns)	1763	2025		1360		1106		2963		**2282**	
Power	Cell Pwr. (mW)	334.5	343.0	+2.5%	334.6	+0.04%	332.8	−0.5%	339.5	+1.5%	**336.7**	**+0.7%**
	Net Pwr. (mW)	115.2	142.6	**+23.8%**	107.0	**−7.1%**	106.0	**−8.0%**	111.5	−3.2%	**135.5**	**+17.6%**
	Leak Pwr. (mW)	47.8	49.1	+2.6%	45.7	−4.4%	44.3	−7.4%	50.0	+4.6%	**47.1**	**−1.6%**
	Tot Pwr. (mW)	497.5	534.6	+7.5%	487.3	−2.0%	483.0	−2.9%	501.1	+0.7%	**519.2**	**+4.4%**
PDP ratio		1	1.052		1.207		1.382		0.979		**1.036**	

% difference is calculated w.r.t. 2D design result (WL = wirelength, WNS = Worst negative slack, TNS = Total negative slack, TPS = Total positive slack).

tier as much as possible, which uses the least inter-tier connections. Also note again that the last column (inter-mixed design flow) will be explained and analyzed in Section 5.4.

Table 4 compares the PPA values of all design flows, which shows a similar trend to the gate-level design results (Table 3). Cascade-2D design result 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: better wirelength/power and worse WNS/TNS. The table also shows the Compact-2D design results with post-TP-opt, which improves timing and consumes more power.

4.4 Summary

Cascade-2D works well on the benchmark circuit with full information that enables delicate tier partitioning by applying it at the beginning of the flow. In addition, it treats pre-placed hard macro blocks perfectly as the designer wants (e.g., Figure 10), and thus works best on mixed-size designs as well. Shrunk-2D and Compact-2D, on the other hand, works the best for flattened gate-level designs. However, Shrunk-2D requires manual manipulation of technology files in PDK, which has a risk of fatal error during design process. Commercial-grade PDKs are highly refined, and it becomes much harder to generate flawless non-integer (i.e., 0.707 for two-tier design) scaled technology files by hand. It is even impossible to shrink them if the PDK is already in the smallest node that the CAD tool can handle. Compact-2D removed such risk by simply derating RC values. Moreover, they are based on tier-by-tier routing and it makes timing closure incomplete with considerable WNSs. Post-TP-opt in Compact-2D somehow solves this, although it is not fundamental since it usually requires one to modify the parasitic extraction file, which is usually encrypted for commercial PDKs. Later in Section 5.4, we introduce an inter-mixed pseudo-3D design flow that avoids all the weaknesses and maintain their strengths as much as possible.

Overall, M3D design results from all pseudo-3D design flows show similar or larger PDP results than 2D design counterparts in most of the benchmark circuits. Cascade-2D design results have large wirelength which increases the total power consumption ("**power**" factor). The PDP result of Cascade-2D design is the worst in LDPC benchmark, which is the most wire-dominated circuits

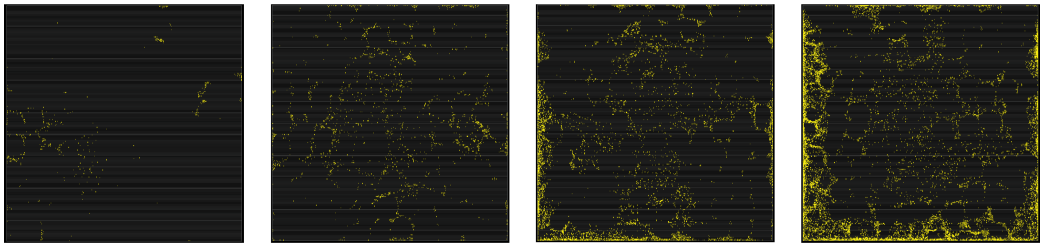

Fig. 13. MIV placement for various MIV counts. Yellow dots represent MIV locations.

among all benchmarks. In contrast, Shrunk-2D and Compact-2D design results have large negative slack values that blow up the "**delay**" factor. Post-TP-opt enabled simultaneous timing closure and covered the intrinsic drawback of Compact-2D flow, with a risk that it is impossible for highly encrypted PDKs.

5 ENHANCEMENTS OF PSEUDO-3D DESIGN FLOWS

In this section, we present various enhancement techniques for pseudo-3D design flows to overcome the weaknesses we have found from their present status. For Cascade-2D, we provide placement-aware MIV planning to reduce detoured wire. For Shrunk-2D, we apply post-TP-opt that is only adopted in Compact-2D. We then present a partitioning-first scheme for design freedom for both partitioning-last flows (Shrunk-2D, Compact-2D). In addition, we propose an inter-mixed pseudo-3D design flow that utilizes both advantages of partitioning-first and partitioning-last approaches. Lastly, we discuss the future directions of monolithic 3D design.

5.1 Cascade-2D: Placement-Aware MIV Planning

Here we focus on enhancing MIV planning of Cascade-2D flow, while retaining the traditional tier-partitioning algorithms [8, 10, 12] in the beginning. 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. This makes the MIV planning of Cascade-2D to push the anchor cells toward the chip periphery during sequential placement, as anchor cells are logically next to the I/O pins in each tier. This is not a problem in module-aware tier partitioning (e.g., all cells in a module belong to a same tier) because logic gates in a functional block tend to clump together by placement engine, and thus have a strong attraction to pull corresponding anchor cells toward them. On the other hand, in an arbitrary design with unknown hierarchy, especially when the MIV count is large, such an attraction is weakened as the logical connectivity is more likely to be broken into different tiers. This results in a fatal problem in the final location of MIVs from the MIV planning proposed in [6]. Figure 13 shows that a higher MIV count results in MIVs located at chip periphery. 3D net routing from such MIV placement induces large detoured wires that increases the total wirelength significantly.

Thus, we have to bring anchor cells back from periphery to remove this effect, and also they should not be located randomly inside the chip as it will cause severe wire overhead. We therefore devise a **placement-aware MIV planning (PA-MIV)** to evenly distribute MIVs in the plane and concurrently locate them to proper positions. The main idea is to find 3D net locations from a rough 2D placement result, and then place the anchor cells inside the bounding box of each 3D net. Rough 2D placement helps in finding relative locations of 3D nets, and inserting an MIV nearby each 3D net will not cause excessive usage of detouring wire.

Figure 14(a) shows the Cascade-2D flow that includes PA-MIV. Starting with the 2D netlist, placement (and trial route for finding net bounding boxes) and tier partitioning are performed

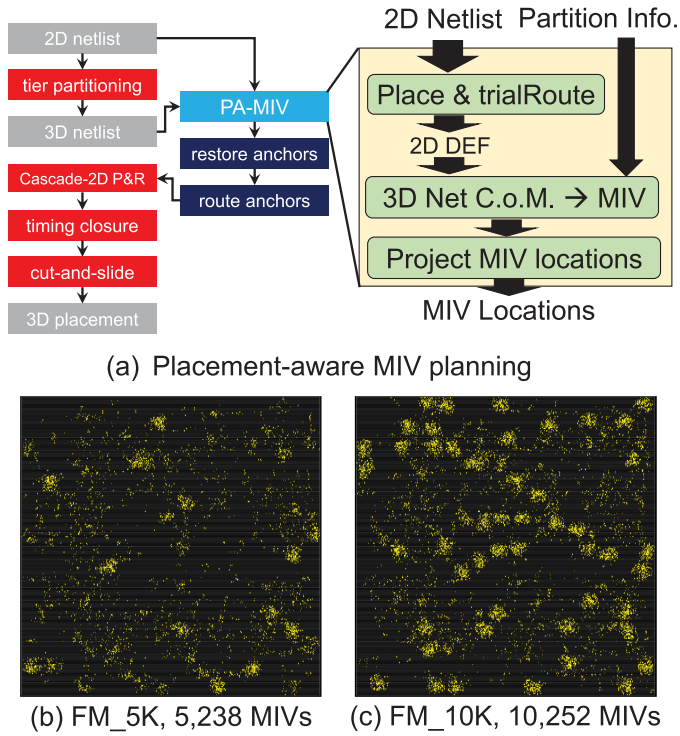

Fig. 14. (a) Placement-aware MIV planning flow. (b),(c) Enhanced MIV placement results from Figure 13(c) and (d).

Table 5. Comparison of Cascade-2D with Original MIV Planning and Placement-Aware MIV Planning (PA-MIV) for AES-128 Benchmark

	Orig.	PA-MIV		Orig.	PA-MIV	
MIV #	5,238			10,252		
Cell area (μm^2)	127,186	122,603	−3.60%	132,582	126,105	−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	295.80	−3.21%

in parallel to generate 2D placement result and tier assignment of each cell. Then we import tier-partitioning results into the 2D placement (in DEF format) to find out 3D net locations in the design. MIV positions are then determined as the center-of-mass of the 3D nets. Lastly, all positions are projected onto the 3D footprint (e.g., all coordinates are multiplied by 0.707 for two-tier design), and transferred to anchor cells at the same coordinates for both tiers to continue the rest of the Cascade-2D flow.

Figure 14(b) and (c) show the enhanced MIV placements of Figure 13(c) and (d). MIVs are spread over the plane instead of clustering at the chip periphery. Table 5 shows the PPA improvements from Figure 14(b) and (c). Placement-aware MIV planning places MIVs at better locations and reduces total wirelength up to 13.29% for the FM_5K tier partitioning algorithm (5,238 MIVs), which performs FM min-cut algorithm [12] until the cut size decreases to 5,000. It also reduces the buffer count for timing closure, and the power consumption.

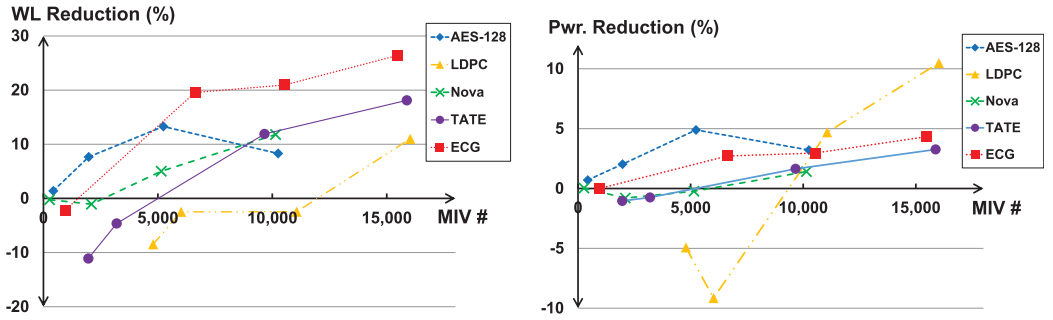

Fig. 15. Wirelength and power savings with placement-aware MIV planning w.r.t. MIV count. Each benchmark is designed with tier partitioning results of four different MIV counts.

Figure 15 plots the wirelength and power savings of PA-MIV. It does not operate well for the cases with small number of MIVs, as they have less problem of MIVs at periphery and additional adjustment of their positions causes design inefficiency. It also goes worse especially on wire-dominated LDPC benchmark. When the MIV count goes high, on the other hand, a problem like Figure 13 happens on original MIV planning and is fixed by PA-MIV. 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).

5.2 Shrunk-2D: Post-TP-opt Application

The main weakness of Shrunk-2D flow (and the original Compact-2D flow) in PDP is the "delay" factor, which is represented by the worst negative slack. Compact-2D flow adds post-TP-opt that enables simultaneous timing closure with risky modifications of commercial PDK. In this subsection, we modified the original Shrunk-2D flow to be able to apply post-TP-opt as well. We used the same technology files (i.e., half-height LEF, 3D qrcTechFile, etc.) that are modified for post-TP-opt of Compact-2D. As observed from both design flowcharts (Figures 5 and 8), there is no critical barrier that prevents the adoption of post-TP-opt into Shrunk-2D as the latter half of both flows is identical.

Table 6 shows the comparison of Shrunk-2D design results from the original flow and from the design flow with post-TP-opt. For all benchmark circuits, regardless of the existence of macro blocks, post-TP-opt in Shrunk-2D eliminates nearly all timing violated paths with almost 75% more MIVs than the original Shrunk-2D design. It also achieved up to **22.9%** of PDP improvement (in TATE benchmark) compared to the original flow.

Post-TP-opt in Shrunk-2D, as well as in Compact-2D, covers the intrinsic drawback of worse timing in original flows by simultaneous timing closure. As their "power" values are already similar or less than 2D designs, reducing the "delay" makes them more efficient than 2D design counterparts in terms of PDP value. Still, note that there is a critical flaw in post-TP-opt that it requires a complicated modification on the PDK, which is even impossible for the highly encrypted PDKs.

5.3 Shrunk-2D and Compact-2D: Partitioning-First Adoption

Another design constraint that both Shrunk-2D and Compact-2D flow have in common is that we don't have multiple choices for tier partitioning but only bin-based FM min-cut algorithm.

Table 6. Comparison of Shrunk-2D: Original vs. Post-TP-opt

		Orig.	Post-TP-opt.			Orig.	Post-TP-opt.
AES-128 (3.4 GHz)	MIV #	39,521	67,684	LDPC (1.2 GHz)	MIV #	15,390	25,483
	WL (m)	1.212	1.229 +1.40%		WL (m)	1.226	1.197 −2.35%
	WNS (ps)	−65.54	**+2.21**		WNS (ps)	−85.74	**+4.35**
	#Viol. path	1, 552	**0**		#Viol. path	706	**0**
	Tot Power (mW)	268.00	283.76 +5.88%		Tot Power (mW)	81.91	83.81 +2.32%
	PDP ratio (vs. 2D)	1.162	**0.997**		PDP ratio (vs. 2D)	1.002	**0.925**
Nova (625 MHz)	MIV #	33,980	54,008	TATE (1.4GHz)	MIV #	52,442	99,728
	WL (m)	2.145	2.099 −2.15%		WL (m)	1.853	1.875 1.21%
	WNS (ps)	−192.98	**−29.50**		WNS (ps)	−97.61	**+69.99**
	#Viol. path	2, 213	**95**		#Viol. path	1, 320	**0**
	Tot Power (mW)	110.74	110.49 −0.22%		Tot Power (mW)	318.99	318.99 +0.00%
	PDP ratio (vs. 2D)	1.082	**0.981**		PDP ratio (vs. 2D)	1.128	**0.899**
ECG (1.0 GHz)	MIV #	21,384	40,627	Rocketcore (1.67 GHz)	MIV #	46,088	82,516
	WL (m)	0.909	0.914 +0.57%		WL (m)	2.542	2.457 −3.37%
	WNS (ps)	−39.64	**+45.71**		WNS (ps)	−196.70	**−39.02**
	#Viol. path	58	**0**		#Viol. path	14, 701	**945**
	Tot Power (mW)	105.70	105.54 −0.15%		Tot Power (mW)	487.3	499.31 +2.46%
	PDP ratio (vs. 2D)	1.013	**0.929**		PDP ratio (vs. 2D)	1.207	**0.992**

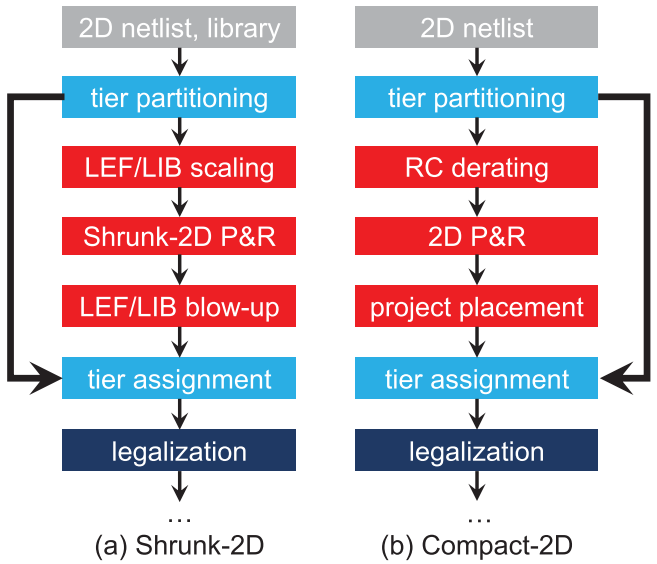

Fig. 16. Partition-first 3D placement of (a) Shrunk-2D and (b) Compact-2D design flow.

Although this guarantees maximal efficiency, it has less freedom to control the design flow. This becomes a concern especially when it comes to designs that the designer wants to manipulate tier partitioning based on sufficient information. In this subsection, we apply the partitioning-first scheme into these two design flows to determine the effects.

Figure 16 shows the flowchart of our partitioning-first scheme in both Shrunk-2D and Compact-2D. We perform RTL-level tier partitioning at the beginning, and import this result

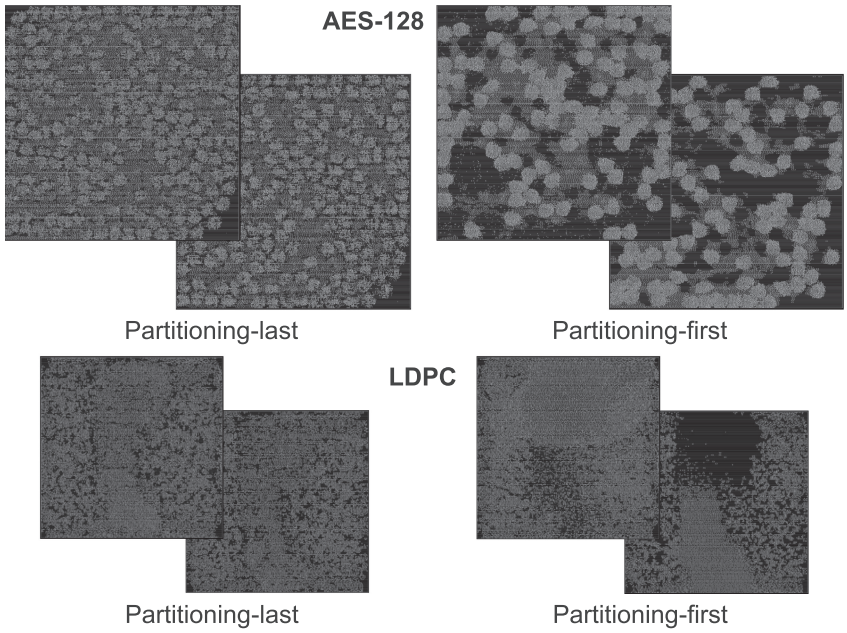

Fig. 17. Partitioning-last and partitioning-first placement comparison for AES-128 (top) and LDPC benchmark (bottom).

into the intermediate P&R result later in the tier assignment stage by allocating each cell to its corresponding tier. Here, additional optimization resources inserted during P&R (i.e., buffers, inverters) are assigned to the tier with more connected cells, which is implemented by applying the original bin-based FM min-cut algorithm only to the added cells. Remaining stages are the same as their original design flows (Figures 5 and 8).

In Figure 17, while cells are evenly distributed in both tiers of the original Shrunk-2D design (left), partitioning-first Shrunk-2D design (right) has whitespaces between cell groups, which incurs inefficient resource utilization and thus degrades PPA results. The main reason is that the partitioning-first scheme performs tier partitioning without the cell placement information, and thus has no technique to avoid expected local area skews in the final placement result.

Table 7 shows that design results based on the partitioning-first scheme has longer wirelength and more power consumption than that of the original pseudo-3D designs. For the cell-dominated benchmark AES-128, the amount of PPA degradation is tolerable as an exchange of more design freedom. The inefficiency becomes considerable in the wire-dominated LDPC benchmark, in that the partitioning-first design requires more wirelength which incurs more power consumption, especially in net switching power. 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.

In summary, partitioning-first scheme sacrifices PPA for more design freedom to apply the designer-defined tier assignment. As the design can be more efficient if the designer knows the background well and is able to apply a better tier partitioning result that outweighs this degradation, it is worth adopting the partitioning-first approach in such case, especially if the target benchmark is a cell-dominated design. Partitioning-first approaches (i.e., Cascade-2D, partitioning-first adoption of Shrunk-2D, and Compact-2D) seem less attractive as our main target benchmarks are gate-level flattened designs with unknown design hierarchy for a straight and fair comparison

Table 7. 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 (ps)	−65.54	−71.96		−85.74	−622.72	
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 (ps)	−68.70	−77.40		−101.83	−484.83	
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%

of all design flows. On the other hand, experimental results from [6] show the superiority of the partitioning-first approach when the design hierarchy is analyzed thoroughly and an elaborated tier partitioning method is applied accordingly. With our work in this subsection, it is also applicable to Shrunk-2D and Compact-2D design flows.

5.4 Inter-Mixed Pseudo-3D Design Flow

From the thorough analysis of the above design flows, we found that Shrunk-2D or Compact-2D flow has the advantage of placement-aware tier partitioning and the abundant MIV usages for better 3D placement, and Cascade-2D flow has the merit of simultaneous timing closure. In this subsection, we propose a *hybrid pseudo-3D design flow* in which it inter-mixes both strong points, in that it uses Compact-2D at front-end and Cascade-2D at back-end. Both Shrunk-2D and Compact-2D are able to be used at the front-end, and we choose Compact-2D for easier fusion with Cascade-2D. Risky PDK modification for post-TP-opt is not required in this flow.

Figure 18 shows the flowchart of the inter-mixed pseudo-3D design flow. We apply Compact-2D front-end stages until the tier partitioning stage, and mount the Compact-2D P&R result and partitioning result to invoke PA-MIV of Cascade-2D. PA-MIV skips the first "Place & trialRoute" stage and uses the Compact-2D P&R result as the intermediate input (i.e., "2D DEF" in Figure 14(a)), which makes the PA-MIV to refer to the actual placement information of logics and thereby generate improved MIV planning results even for the designs with small MIV count. In the meantime, the 3D placement from the tier partitioning stage is not discarded but legalized tier-by-tier, and are merged altogether with the anchor cell locations from the PA-MIV result into the Cascade-2D canvas. We treat this as a start point and perform rest steps (i.e., CTS, route, timing closure, ...) as the Cascade-2D back-end stages. This flow is able to use the partitioning-last approach of Compact-2D that performs better tier partitioning with placement information, and at the same time utilizes simultaneous timing closure of Cascade-2D that is stronger at fixing negative slacks.

The last column of Tables 3 and 4 shows the design comparison of inter-mixed flow along with other pseudo-3D design flows. In terms of PDP value, inter-mixed flow shows better results compared to the original Cascade-2D, Shrunk-2D, and Compact-2D for all cases except LDPC. For some cases (AES-128 and ECG), it is even better than Compact-2D with post-TP-opt and 2D

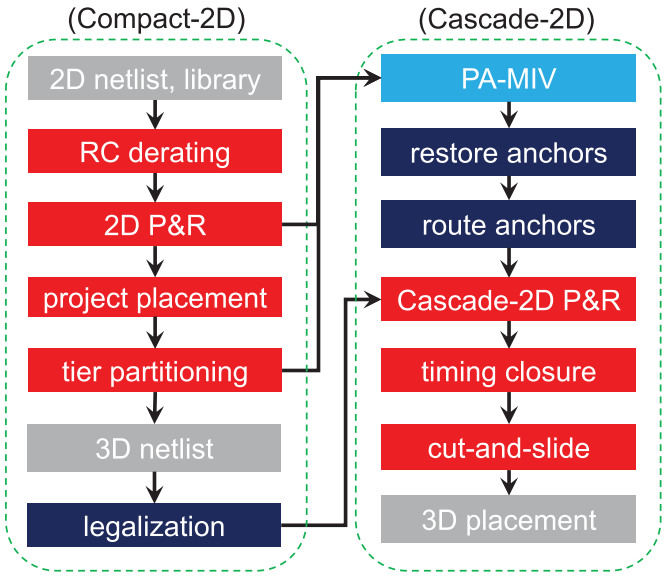

Fig. 18. Inter-mixed pseudo-3D design flow that connects the front-end of Compact-2D and the back-end of Cascade-2D.

design counterpart as well. Design results from inter-mixed flow shows a similar trend to those of Cascade-2D (i.e., better timing, worse power), but WNS is even smaller than Cascade-2D for most cases. It is because inter-mixed flow, which is combined with Compact-2D front-end, is able to utilize more MIVs than the original Cascade-2D for more flexible timing closure. However, as an anchor cell ($135\,nm \times 1200\,nm$) has to be larger than an MIV ($100\,nm \times 100\,nm$), the MIV area is always overestimated and thus excessive usage of MIVs in this flow harms the design quality at last. Therefore, we limit the MIV count to a certain degree below Shrunk-2D or Compact-2D by controlling the bin size during tier partitioning, and this makes the inter-mixed flow have worse design results in wire-dominated benchmarks (i.e., LDPC) like Cascade-2D.

Taken overall, the proposed inter-mixed flow achieves the best PPA results among all the pseudo-3D flows, except Compact-2D with post-TP-opt. However, the main flaw of post-TP-opt is that it is impossible to be applied for some commercial PDKs, in that we usually have no access to modify the internal source codes of the encrypted parasitic extraction data. On the other hand, inter-mixed flow provides design results comparable to post-TP-opt regardless of PDK. Thus, the inter-mixed pseudo-3D design flow is a valuable option to be chosen because of the superiority on design quality compared to the original design flows and simple technology file modification.

5.5 Future Directions for Monolithic 3D Design

Based on the insight of pseudo-3D design flows, this subsection introduces future directions of pseudo-3D flow and M3D design. Although most problems related to 3D design (e.g., 3D placement, simultaneous timing closure, ...) are treated by some or all pseudo-3D design flows, there are still issues that left unsolved in terms of M3D design, which are the keys for realizing a commercial true-3D P&R tool.

(1) Tier partitioning: A pseudo-3D design flow is a mixture of 2D-based commercial CAD engines and the in-house tools that enable the utilization of these engines for 3D design. As in Table 8, per-tier placement and 3D routing stages are well equipped with commercial CAD tools, while tier partitioning is based on a standalone algorithm implemented with an in-house executable (e.g.,

Table 8. Type of Tools (In-House or Commercial CAD Tool) for Each Stage of Pseudo-3D Flows

Cascade-2D		Shrunk-2D		Compact-2D	
Tier partitioning	**in-house**	LEF/LIB scaling	**in-house**	RC derating	CAD tool
MIV planning	**in-house**	Shrunk-2D P&R	CAD tool	P&R	CAD tool
Cascade-2D P&R	CAD tool	Tier partitioning	**in-house**	Tier partitioning	**in-house**
Timing closure	CAD tool	Legalize & MIV plan & routing & closure	CAD tool	Legalize & MIV plan & routing & closure	CAD tool
Cut-and-slide	CAD tool				
				Tech fix	**in-house**
				Post-TP-opt	CAD tool

C++, Python). In general, the main objective of such partitioning algorithm is to achieve minimal cut size between partitions and is widely applied to tier partitioning of TSV-based 3D IC to reduce the high-overhead TSV count. However, this is not of most concern to monolithic 3D design since the cost of an MIV is as negligible as a single via. Shrunk-2D and Compact-2D applies a bin-based algorithm to divide the problem space into bins and use the placement information for advanced tier partitioning. It is a CAD-friendly approach that the user can sweep the bin size to control the MIV count, but it still is not an "optimal" partitioning method. Therefore, a tier partitioning method that considers the physical location of each cut (MIV) optimally instead of just minimizing it should be devised.

(2) Precise MIV effect with 3D PDK: In Table 8, Cascade-2D flow has another stage with in-house tool, which is MIV planning. For both original MIV planning of sequential placement and proposed PA-MIV, an anchor cell is used to represent an MIV and its position is fixed during the rest stages. But the minimum required size of an anchor cell is a site size of PDK, which is always larger than an MIV. In Shrunk-2D and Compact-2D flows, a 2D metal stack is duplicated to the top and bottom tier to represent a 3D metal stack, and MIV is treated as a normal via between the top-most metal layer of the bottom tier and the bottom-most metal layer of the top tier. They are written in the modified technology files for M3D design (i.e., tech LEF, 3D qrcTechFile), which are also performed by in-house tools (Shrunk-2D: LEF/LIB scaling, Compact-2D: tech fix). However, accuracy of such modification is not guaranteed, as the metal stack of both tiers may have different parasitic values and MIV also could have unique characteristics when penetrating the silicon substrate. These problems can be solved at once by PDK vendor to provide a 3D metal stack with precise specification of an MIV. There was no provision of 3D PDK because M3D design was not active, but recently the attention to M3D design arises and the demand for 3D PDK is also increased. Therefore, a commercial-grade 3D PDK will boost up the research on monolithic 3D design automation and optimization.

(3) Hierarchical (block-level) design: So far, the pseudo-3D design flows mainly targeted flattened design for better PPA. But as the M3D design is becoming real, it is time to consider the design hierarchy and block-level designs for practical application of M3D design flows. While Cascade-2D flow is somehow friendly to such designs as it can reflect design hierarchy or module assignment in the beginning of the flow (tier partitioning), it is not optimized well because of the reasons we stated in the above sections (i.e., Section 4.4, Section 5.1). The original Shrunk-2D and

Compact-2D have no option of doing so, and adopting a partitioning-first approach as in Section 5.3 is still not enough since the design quality is degraded a lot for some cases (e.g., wire-dominated benchmarks). As we now have effective M3D design flows for gate-level designs as shown in this article, we should also devise a constructive design flow specialized to block-level hierarchical designs.

6 CONCLUSIONS

This article 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 qualitative factors and quantitative PPA. We also introduced enhancement techniques for each pseudo-3D design flow and an inter-mixed flow that takes advantages from both approaches. 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. As for Shrunk-2D, we applied post-TP-opt to enhance the original flow by reducing up to 20% of PDP under the risk of extreme PDK modification. Partitioning-first adoption of Shrunk-2D and Compact-2D added another dimension of design freedom with a tolerable amount of PPA degradation. In addition, proposed inter-mixed flow also outweighs the original pseudo-3D design flows in terms of PDP. Lastly, based on our detailed analysis on pseudo-3D design flows, we introduced the future directions of monolithic 3D IC design.

REFERENCES

[1] [n.d.]. DARPA ERI Summit 2019. Retrieved on Dec. 1, 2020 from https://eri-summit.darpa.mil/2019-archive-keynote-slides.

[2] [n.d.]. OpenCores benchmark suite. Retrieved on Dec. 1, 2020 from http://www.opencores.org/.

[3] Krste Asanovic et al. 2016. *The Rocket Chip Generator*. Technical Report UCB/EECS-2016-17 *EECS Department, University of California, Berkeley,*.

[4] P. Batude et al. 2009. Advances in 3D CMOS sequential integration. In *2009 IEEE International Electron Devices Meeting (IEDM)*. 1–4.

[5] Tony F. Chan et al. 2006. mPL6: Enhanced multilevel mixed-size placement. In *ISPD*. 212–214. DOI: https://doi.org/10.1145/1123008.1123055

[6] K. Chang et al. 2016. Cascade2D: A design-aware partitioning approach to monolithic 3D IC with 2D commercial tools. In *ICCAD*. 1–8. DOI: https://doi.org/10.1145/2966986.2967013

[7] T. Chen et al. 2008. NTUplace3: An analytical placer for large-scale mixed-size designs with preplaced blocks and density constraints. *TCAD* 27, 7 (2008), 1228–1240. DOI: https://doi.org/10.1109/TCAD.2008.923063

[8] J. Cong et al. 1997. Large scale circuit partitioning with loose/stable net removal and signal flow based clustering. In *ICCAD*. 441–446. DOI: https://doi.org/10.1109/ICCAD.1997.643573

[9] J. Cong et al. 2007. Thermal-aware 3D IC placement via transformation. In *ASP-DAC*. 780–785. DOI: https://doi.org/10.1109/ASPDAC.2007.358084

[10] J. Cong and Sung Kyu Lim. 2004. Edge separability-based circuit clustering with application to multilevel circuit partitioning. *TCAD* 23, 3 (2004), 346–357. DOI: https://doi.org/10.1109/TCAD.2004.823353

[11] J. Cong and G. Luo. 2009. A multilevel analytical placement for 3D ICs. In *ASP-DAC*. 361–366. DOI: https://doi.org/10.1109/ASPDAC.2009.4796507

[12] C. M. Fiduccia and R. M. Mattheyses. 1982. A linear-time heuristic for improving network partitions. In *DAC*. 175–181. DOI: https://doi.org/10.1109/DAC.1982.1585498

[13] B. Goplen and S. Sapatnekar. 2003. Efficient thermal placement of standard cells in 3D ICs using a force directed approach. In *ICCAD*. 86–89. DOI: https://doi.org/10.1109/ICCAD.2003.1257591

[14] B. Goplen and S. Sapatnekar. 2007. Placement of 3D ICs with thermal and interlayer via considerations. In *DAC*. 626–631.

[15] M. Hsu et al. 2011. TSV-aware analytical placement for 3D IC designs. In *DAC*. 664–669.

[16] D. H. Kim et al. 2009. A study of through-silicon-via impact on the 3D stacked IC layout. In *ICCAD*. 674–680.

[17] S. S. Kiran Pentapati et al. 2020. Pin-3D: A physical synthesis and post-layout optimization flow for heterogeneous monolithic 3D ICs. In *2020 IEEE/ACM International Conference on Computer Aided Design (ICCAD'20)*. 1–9.

[18] Bon Woong Ku et al. 2018. Compact-2D: A physical design methodology to build commercial-quality face-to-face-bonded 3D ICs. In *ISPD*. 90–97.
[19] Jingwei Lu et al. 2015. ePlace: Electrostatics-based placement using fast Fourier transform and Nesterov's method. *TODAES* 20, 2, Article 17 (2015), 34 pages. DOI: https://doi.org/10.1145/2699873
[20] Jingwei Lu et al. 2016. ePlace-3D: Electrostatics based placement for 3D-ICs. In *ISPD*. 11–18. DOI: https://doi.org/10.1145/2872334.2872361
[21] S. Panth et al. 2014. Design and CAD methodologies for low power gate-level monolithic 3D ICs. In *ISLPED*. 171–176. DOI: https://doi.org/10.1145/2627369.2627642
[22] Sandeep Samal et al. 2016. Monolithic 3D IC vs. TSV-based 3D IC in 14nm FinFET technolog. In *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*.
[23] Jiajun Shi et al. 2016. A 14nm FinFET transistor-level 3D partitioning design to enable high-performance and low-cost monolithic 3D IC. In *IEEE International Electron Devices Meeting*.
[24] M. M. Shulaker et al. 2014. Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs. In *IEDM*. 27.4.1–27.4.4.
[25] M. M. Shulaker et al. 2017. Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. *Nature* 547, 8 (2017), 74–78.
[26] P. Spindler et al. 2008. Kraftwerk2—A fast force-directed quadratic placement approach using an accurate net model. *TCAD* 27, 8 (2008), 1398–1411. DOI: https://doi.org/10.1109/TCAD.2008.925783
[27] Simon Wong et al. 2007. Monolithic 3D integrated circuits. In *VLSI-TSA*.
[28] J. Zhang et al. 2012. Carbon nanotube robust digital VLSI. *TCAD* 31, 4 (2012), 453–471.

Received December 2020; revised February 2021; accepted February 2021