Pin-3D: A Physical Synthesis and Post-Layout Optimization Flow for Heterogeneous Monolithic 3D ICs

Sai Surya Kiran Pentapati
sai.pentapati@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Kyungwook Chang
kchang.gatech@gmail.com
Georgia Institute of Technology
Atlanta, Georgia

Vassilios Gerousis
vgerousis24@gmail.com
Samsung Semiconductor Inc.
Austin, Texas

Rwik Sengupta
rwik78@gmail.com
Samsung Semiconductor Inc.
Austin, Texas

Sung Kyu Lim
limsk@ece.gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

In this paper, we present an optimization flow for monolithic 3D ICs called Pin-3D Optimizer. Compared with the state-of-the-art RTL-to-GDS flows that rely on ad-hoc technology file tweaks and RC scaling, Pin-3D offers a streamlined method to run commercial 2D IC tools to obtain high-quality monolithic 3D IC designs. Specifically, Pin-3D supports effective legalization, routing, timing closure, and ECO optimization for monolithic 3D IC designs. We propose a novel optimization methodology where the cells in each tier of a 3D IC are optimized using cell data and constraints of the full 3D design. The optimizations in a tier also directly influence the timing, power in the other tiers, leading to better overall PPA of the 3D IC. With the help of two industry processors designed with a 28 nm technology node, we show that Pin-3D provides up-to 9.0% smaller wirelength and 88% smaller total negative slack than die-by-die M3D flows. We also observe up-to 8.7% lower power and 26% smaller wirelength than 2D ICs. In addition, Pin-3D is the first flow that supports routing and timing optimization in heterogeneous logic-on-logic monolithic 3D ICs. We demonstrate this capability by performing area-balanced tier partitioning, routing, and timing closure of a 3D design with different technologies on each die.

1 INTRODUCTION

Monolithic 3D IC (M3D) is a bare die stacking technology, where standard cells, or even transistors, can be placed on top of each other in the third-dimension. This technology achieves better energy-efficiency and a smaller chip footprint than a 2D IC without shrinking the transistors. Moreover, the inter-tier via that is called monolithic inter-tier via (MIV) is in nano-meter scale (50 nm to 100 nm, typically). Thus, MIV allows ultra-high density logic-on-logic and logic-on-memory stacking, which the micron-scale through-silicon-via (TSV) cannot support easily.

ICCAD '20, November 2–5, 2020, Virtual Event, USA

ACM ISBN 978-1-6654-2324-3/20/11…$15.00
https://doi.org/10.1145/3400302.3415720

In an M3D IC, the dies (tiers) are stacked on top of each other in a sequential fashion [1, 2]. To fully utilize the benefits of M3D ICs, along with the improvements in M3D fabrication techniques, Electronic Design Automation (EDA) tool flows should be developed for placement, clock design, routing, and optimization of the 3D ICs. In our work, we assume only two-tiered 3D ICs due to the limitation of maximum number of metal layers supported in the present commercial EDA tools.

In this paper, we present a new timing closure and optimization engine for M3D ICs called Pin-3D Optimizer that is applicable to homogeneous and for the first time, heterogeneous 3D ICs. We perform die-by-die optimization with full 3D design context to design a commercial quality 3D IC. Specifically, Pin-3D supports effective legalization, routing, timing closure, and Engineering Change Order (ECO) optimizations for M3D designs. ECO in 2D ICs is used to improve and simplify the fine-tuning ability for manual timing closure in corner cases. Pin-3D provides this ECO capability with M3D designs for the first time. Using two industry processors, and open-source benchmarks - LDPC and Netcard - designed in a commercial 28 nm node, we show that M3D ICs have 8-30% power reduction along with 18-38% wirelength reduction compared to 2D ICs. Pin-3D is also the first flow to support heterogeneous gate-level Monolithic 3D IC optimization, we show this by designing and optimizing a 128-bit AES encoder circuit with 45 nm technology node on bottom-die, and 15 nm node on the top.

2 BACKGROUND AND RELATED WORK

There have been several academic 3D placers [5–8] developed in recent years for TSV based 3D ICs. While these placers can be modified to design M3D ICs, they have neither routing nor timing closure capabilities, which is critical in designing commercial quality 3D ICs. In order to overcome these limitations, "pseudo-3D flows" [3, 4] have been proposed that extend commercial 2D EDA capabilities to 3D ICs. In a pseudo-3D flow, 3D designs are built using an "intermediate 2D design" that closely emulates 3D parasitics. This intermediate 2D design is then partitioned into multiple tiers and routed to obtain the final 3D design.

Shrunk-2D [3] relies on "cell and interconnect shrinking" to build a 2D layout that tries to emulate final 3D IC footprint, cell locations, and wire parasitics. After the intermediate 2D design stage, cells

Table 1: Qualitative comparison among state-of-the-art "pseudo-3D" physical design tools for monolithic 3D ICs and this work. "enhanced die-by-die" means the pins from both dies are visible during die-by-die optimization.

	Shrunk-2D [3]	Compact-2D [4] without 3D Optimization	Compact-2D [4] with 3D Optimization	Pin-3D (this work)
Key idea	cell and wire shrinking	placement compaction	row halving	pin projection
3D stack	two separate dies	two separate dies	double metal stack	double metal stack
Strength	first pseudo-3D flow	shrinking unnecessary	3D optimization	overcome weaknesses in [3, 4]
Weakness	shrinking causes DRC issues	under buffering/sizing	DRC due to row halving	–
Placement	2D + tier partitioning	2D + tier partitioning	2D + tier partitioning	2D + tier partitioning
Legalization	die-by-die	die-by-die	die-by-die	enhanced die-by-die
Signal Routing	die-by-die	die-by-die	true 3D (ignoring DRC)	true 3D
Clock Tree Design	2D + tier partitioning	2D + tier partitioning	2D + tier partitioning	2D + tier partitioning
Power/Ground Routing	manual	manual	manual	manual
Post Route Optimization	not supported	not supported	both dies at once	enhanced die-by-die
Engineering Change Order	not supported	not supported	not supported	supported
Heterogeneous 3D ICs	not supported	not supported	not supported	supported

are expanded back to their original sizes which causes overlaps that are resolved with tier partitioning and legalization.

Compact-2D [4] avoids geometry shrinking, but relies on "layout compaction". Once a pseudo-3D layout is completed, the design is compacted into a 3D IC footprint that is smaller than that of a 2D IC and subsequent tier-partitioning is performed. In Compact-2D, however, wire RC values need to be scaled by $1/\sqrt{2}$ during the intermediate 2D design as it is later compacted. This decision of scaling down all interconnect RC is ad-hoc, because not all interconnects will be shortened by the same ratio in the final 3D design. A post layout optimization stage in Compact-2D then performs full 3D routing and timing closure in the post tier-partitioning stage. DRC violation fixing in Compact-2D cannot be done with both dies at once, and an additional die-by-die stage is required.

Table 1 summarises various pseudo-3D flows. As we do not have the means to implement 3D Optimization in Compact-2D, we use it without 3D-optimization for M3D PPA comparisons in this paper. Pin-3D has built-in support for 3D optimization and compared to Compact-2D's 3D optimization, Pin-3D's pin projection methodology (discussed in Section 3) also allows us to perform heterogeneous 3D IC routing, optimization, ECO in 3D ICs.

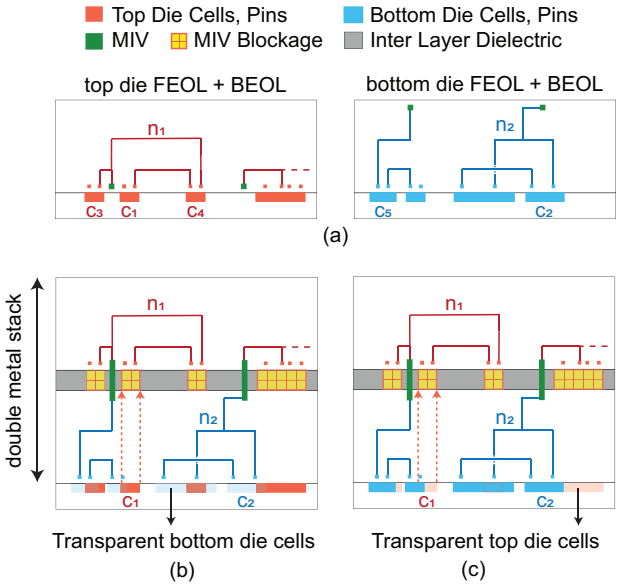

Figure 1: The key idea of Pin-3D: die merging and pin projection. (a) top and the bottom dies separately (b), (c) merged dies for the top die and bottom die optimization respectively. The double metal stack contains pins from both dies to provide the entire 3D context during die-by-die operations. Top die cells are also projected to the MIV layer to ensure no overlap between MIV and routed nets. Pin-3D allows using two different technology nodes as demonstrated in Section 5.7.

3 PIN-3D OPTIMIZER

3.1 Key Idea

The key idea behind Pin-3D flow is called 'pin projection'. This idea allows us to place all the cells in a 3D IC footprint, without scaling the cell geometry. A 3D IC with two-dies stacked on top of each other is shown in Figure 1(a). It is not possible to represent an accurate 3D IC layout in the commercial EDA tools, as they only support a single Front End Of Line (FEOL) layer to place active devices. On the other hand, the Back End of Line (BEOL) in a 3D IC is easily representable (with limitations on metal layer count) as the tools support multiple metal layers. So, we first create a single die with same footprint as the 3D IC, and then place the cells from the top die on the same layer as the bottom die cells. The pins of the top-die cells are projected back to the corresponding top-die metal layers as seen with top-die pins in Figures 1(b), 1(c) to achieve realistic 3D routing with pins on the proper dies according to their cells.

Flattening the 3D cell placement onto a single FEOL layer creates unwanted overlaps between the cells of different tiers. To rectify for this, the cells from the one of the dies are fixed and made "transparent" within the core area, and only the cells from one die are movable at any given stage of the design. In Figure 1(b), the bottom-die cells are made transparent, and in Figure 1(c), the top-die cells

are transparent. In both these cases, the other die is left as it is (non-transparent). To the placement and routing engines, the transparent cells are 'invisible' and only their cell pins are 'visible' looking like a 'projection' of pins on specified metal layer without an associated cell footprint. This is the key pin projection idea that allows us to do optimization in both homogeneous and heterogeneous 3D ICs. The pin layer assignment, placement flattening, and pin projection allows for the cells in non-transparent die to be moved around freely while considering the accurate, full connectivity to cells from both the tiers. This 3D connectivity is possible to achieve with a single FEOL stack as evident from the nets 'n1', 'n2' in Figure 1(b), 1(c). In the split die representation of Figure 1(a), these nets are broken into sub-nets within each die.

3.2 Benefits of Pin-3D Optimizer

With this structure as shown in Figure 1(b), 1(c), the EDA tools obtain full 3D context and perform full 3D-routing in a single pass. Legalization and 3D driven optimization with Pin-3D flow are done in an enhanced die-by-die fashion with the complete design in memory, making them 3D-aware. Consider the layout as shown in figure 1(b), at this stage the bottom-die cells are fixed and transparent. Here, during placement and timing optimizations on the top-die, any change to cell location or sizing of these top-die cells will affect the full-3D routing and connectivity. These changes are observable to the transparent bottom-die, and the input and output timing of the bottom-die cells are correctly updated at every instant based on information provided by the timing libraries. Details on the placement and timing optimization are provided in Sections 3.5, 3.7 respectively.

Contrarily, in the die-by-die methodology shown in Figure 1(a), only a single die can be loaded in the tool, and the changes to one die are completely "opaque" to the other die. The timing information cannot be propagated between any two cells that are connected via an MIV through the other die. While some amount of information in independent die-by-die methodology can be updated through timing constraints of the MIV pins after routing in each die, these constraints are not nearly as fine-grained or as synergistic as in the Pin-3D methodology.

Pin-3D also has significant benefits when compared to the simultaneous 3D optimization of post tier-partitioning optimization (3D optimization) in Compact-2D with regards to designs with huge macro blocks, and heterogeneous 3D IC optimization. In the 3D optimization of Compact-2D, the placement is still flattened, but the cell overlaps due to flattening are resolved with row and cell height halving. In case of macro blocks that spawn multiple rows, Compact-2D cannot optimize the cells located on top or bottom of these macro blocks. Due to the pins being located outside the halved cell area, the routing design rules cannot be strictly followed during 3D routing and optimization of Compact-2D. Pin-3D does not scale the geometry and allows for better Design Rule Checks.

Row splitting idea is inapplicable when dealing with heterogeneous 3D ICs where the cells in top and bottom dies belong to different technology nodes with unequal row heights. Heterogeneous 3D IC design is an extremely useful approach that provides a wide array of possibilities in power and performance improvements using a high-performance technology on one die, and low-power

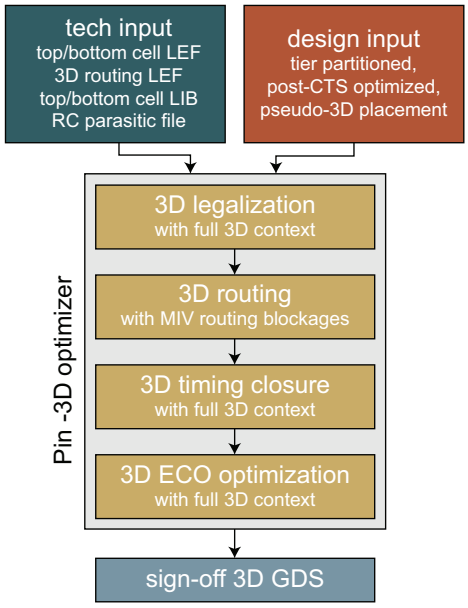

Figure 2: Our Pin-3D optimizer design flow.

technology on the other. Block-level face-to-face integrated heterogeneous 3D IC processor has recently been taped-out [9] by Intel using Foveros 3D-integration [10]. Exploring heterogeneous integration at the finer gate-level integration with monolithic 3D ICs would be of an extreme importance, and is supported for the first time with Pin-3D.

3.3 Overall Design Flow

Pin-3D optimizer is used for 3D placement optimization within each tier, final routing, and timing optimization stages within each tier for the 3D ICs. Therefore, the input design is a tier-partitioned 3D IC with a synthesized clock tree. The routing done in the intermediate pseudo-3D stage is not useful with the 3D metal stack and is discarded. Figure 2 shows the overall flow of the Pin-3D methodology along with the required inputs to the flow. In this paper, the clock-synthesis and optimizations in pseudo-3D stage are done with Compact-2D flow [4], but any other RTL-to-GDS methodology such as Shrunk-2D [3] or any custom partitioned design can be used as the input.

3.4 Technology File Generation

We create M3D-specific FEOL to facilitate pin projection with transparent cells, and M3D BEOL files to create the double metal stack structure. For the M3D BEOL files, we generate the Library Exchange Format (LEF) file with the complete routing rules of the metal layers, and Interconnect Technology (ICT) RC lookup table files for parasitic extractions of the wires in the 3D metal stack. The BEOL files are generated similar to the files in Compact-2D design's post-tier-partitioning optimization.

Second, we create the M3D FEOL files, specifically cell LEF containing cell and pin shapes, and Liberty (LIB) files with cell power, delay look-up tables, and pin capacitances. These cell LEF files are

the main set of files that help us realize the "transparent" cells in placement layer. This is done using "OBS" (obstruction) definitions for the cells. The OBS is traditionally used for creating metal layer obstructions that specify the tool to not route over internal cell pins, but this can also be used for creating custom placement obstructions. In the cell LEF file, the "SIZE" statement specifies the shape of any cell or macro, but it only supports non-zero rectangular bounding boxes. Using the OBS construct for the cells, we define the placement layer (referred to as OVERLAP layer in LEF) obstruction to be a zero-sized rectangle at the origin of the cell. This makes the cells transparent in placement engines and do not contribute to the placement overlaps. The pin shapes are unaffected as they are defined using a separate PORT statement in the LEF. The metal layer assignment of the pins in PORT statements are modified to assign the pins of top-die cells to top-die metal layers, and likewise for bottom-die cells.

Two sets of cell LEF files are created for representing the two different configurations shown in Figure 1(b), 1(c). The LEF to be used during top-die optimizations has 'transparent' bottom-die cells, and vice versa. In both these LEFs, routing obstructions in the MIV cut-layer (called MIV Blockages in Figure 1(b), 1(c)) are added to the top-die cell with size equal to the cell size. This dictates the tool to not place any MIVs within the bounding boxes of the top-die cells as MIVs cannot penetrate through the cells on top-die. As the pin-shapes are unaffected in either configuration, the routing done using the cell LEFs of either configuration will be identical. No changes to the cell information are made in the LIB files, the top-die cells have the same information as the 2D technology of top-die, bottom-die cells have the same information as top-die for homogeneous 3D ICs. In case of the heterogeneous 3D ICs, the cell information (timing, shape, pin definitions) and metal layer information (routing width, space, offset, material properties) come from different technology nodes for each die as the bottom die cells and top die cells will not be similar in terms of physical, electrical charecteristics.

3.5 Pin-3D Placement Legalization

The design input from either Compact-2D or Shrunk-2D uses placement driven bin-based FM min-cut algorithm for partitioning the cells into two tiers. Due to the nature of these flows, this stage still contains overlaps between cells that need to be legalized before performing timing optimizations. As opposed to the die-by-die placement legalization employed in Shrunk-2D and Compact-2D, in Pin-3D's 3D aware enhanced die-by-die legalization, we use full 3D design context with transparent cells and projected pins to remove overlaps in each die. Using the commercial tool's placement optimization engine, the legalization is done to reduce congestion and improve timing. Without the 3D context, the previous flows use the placement refining engine which just removes the overlaps for the necessary cells. During the top-die legalization in Pin-3D we also employ an additional preferential spacing rule between cells, a placement mode option in the tool, to avoid densely packed local clusters. Such dense local clusters hinder 3D routing as the MIVs have to avoid the top-die cells and be placed outside the clusters. This spacing rule is not needed during bottom-die placement legalization as it does not provide any obstruction to the MIVs.

3.6 Pin-3D Routing

Once the cells are completely legalized, the 3D design is then routed in a 3D fashion with the full 3D metal stack. Even though the cells from only one die are non-transparent here, it doesn't affect the routing engine in any way as all the cell pins are non-transparent. Complete 3D routing also allows for the sharing of metal layers between the two dies. This leads to an optimized use of the MIV layer as the critical nets in top-die can readily access the low-resistance and low-congested top metal layers of the bottom-die. Due to our cell handling using transparent cells technique, all the routing rules are closely followed and incremental routing changes are not needed to do additional violation fixing as in [4].

3.7 Pin-3D Timing Closure

The 3D aware post-route optimization is done in three passes starting with top-die, and then on the bottom-die, and finally re-optimizing the top-die again. To first understand the 3D aware nature of Pin-3D die-by-die optimization and the need for three optimization passes, consider the n_1 net connecting the cells $\{c_3, c_4, c_5\}$ in Figure 1(a). c_5 is connected to the adjacent cell in the same die with one pin (say, input), and the other pin (output) of the cell goes through an MIV and fans out to the two top-die cells c_3, c_4 on n_1. As the full 3D design and the net n_1 are split into two in Figure 1(a), it is impossible to optimize the bottom die cell c_5 without knowing the cell information of c_3, c_4, and the clock skew on its register-to-register path that could be split between the two dies. On the other hand, Figures 1(b), 1(c) contain the full 3D design information within a single die and can consider the clock skew on any path, and information of neighboring cells anywhere in the 3D design.

In the first round of optimization on top-die, the cells c_3, c_4 are optimized based on the input slew, capacitance data from c_5 as well as the c_4 adjacent top-die cell c_1. During this step, the bottom die cells remain fixed, therefore a bad sizing of cells in bottom-die would limit the top-die optimization, as the internal cell delay of c_5 cannot be reduced. In extreme cases, the optimizer adds a top-die buffer to split the net n_1 right after the MIV and reduce the output wire and fan-out cell capacitance load of cell c_5. In the second stage, the bottom die cells are then optimized, and here c_5 can be further optimized to reduce any timing violations. At this stage, the top-die cells are now once optimized with the full 3D context, they would not have cells with unexceptionally large delays unlike the bottom-die cells during first pass. Now that the bottom-die cells are 3D context optimized, a final pass of optimization in the top-die can down-size or remove any aggressively sized cells or buffers from the first iteration. Further optimization iterations lead to very marginal improvements and we stop after the third pass.

3.8 Pin-3D Engineering Change Order

Due to the multi-level die stacking, 3D ICs need to support changing the tier-partitioning of cells along with traditional Engineering Change Order (ECO) changes such as modifying the type of cells, addition of filler cells. With the inclusion of the complete design in memory, we can utilize the ECO utility of EDA tools to change the cell type within each die and to 'push down / pop up' cells to different dies at any stage of the Pin-3D flow. Changing the cell tier is simply done by changing the standard cell type of desired

Algorithm 1: ECO Technique

```
criticalRegs ⟵ launching registers of register-to-register
  paths with slack < 0.0;
nonCriticalRegs ⟵ launching registers of
  register-to-register paths with slack > 150 ps;
foreach reg in criticalRegs do
    if reg is not maximum drive strength then
        Up-size the register;
    else
        Use the corresponding lowest V_th register;
    end
end
foreach reg in nonCriticalRegs do
    if reg is not minimum drive strength then
        Down-size the register;
    else
        Replace with corresponding low-power register;
    end
end
Perform incremental placement and routing
```

cell from its current type to its counterpart in the other tier. By doing so, the cell remains exactly in-place and the pin locations are simply moved from one tier to the other.

ECO performs incremental routing to allow for the changes after cell swapping. Any timing violations that may occur during the ECO stage are fixed easily using incremental optimization. ECO is usually used to manually fix or improve any timing violations that were not closed by the auto optimization engine, or to improve the manufacturability of designs.

In this paper we show the ECO capability of Pin-3D by addressing the following: Cell delay of the launching registers in the Pin-3D optimized design contributes to a significant portion (>10%) of the total delay on the register-to-register paths. This is relatively easy to fix using a simple ECO technique described in Algorithm 1. We create two register arrays 'criticalRegs' and 'nonCriticalRegs' to not only fix the timing violations in 'criticalRegs' group, but also to save power in excessively optimized registers of 'nonCriticalRegs'. For each critical register in the array, we up-size it to the next highest drive strength within the same V_{th} regime. In case the critical register is already at highest driving strength, we change its V_{th} to the least available V_{th} in the same drive strength. Similarly each non-critical register is first down-sized in drive strength, and in the case of least drive-strength register, a low-power register is used within the same V_{th}. Changing the V_{th} of the non-critical registers lead to drastic timing degradation, and we do not change the V_{th} for this group.

4 BENCHMARK AND TECHNOLOGY SETUP

4.1 Homogeneous M3D IC

As heterogeneous 3D ICs are specifically applicable to Pin-3D flow, we perform detailed comparisons using homogeneous 3D ICs. We use a commercial 28 nm technology for both 2D and M3D designs, and the M3D BEOL, FEOL technology files are created as described in Section 3.1. We assume the top and bottom-dies are identical to the 2D dies, i.e., without any FEOL or BEOL degradation, this assumption is reasonably valid as variation aware floorplanning techniques presented in [11] can make the design resilient to performance degradation. Previous works on pseudo-3D flows [3, 4] also make such assumptions.

Single-core industry processors used for demonstrating the PPA benefits of the Pin-3D flow are configured with the following main blocks: single core CPU, 32 kB of L1 Cache, Floating Point Unit (FPU). These processors are referred to as Industry-A, Industry-B in the later sections. Under an existing non-disclosure agreement (NDA), we do not provide raw data but normalize them to protect sensitive material. Along with the two processors, two open-source circuits - LDPC, Netcard - that do not contain memory macros are also used as the test-bench circuits. We report raw data for these circuits. LDPC is a small circuit with close to 55,000 gates and nets each and only 2,048 registers. Netcard is a relatively big design containing 240,000 gates and nets each with 67,200 registers.

4.2 Heterogeneous M3D IC

Our heterogeneous M3D technology files are custom generated using technology information using open-source 15 nm academic PDK for the top-die, and a 45 nm academic PDK for the bottom-die cells, metal layer information.[1] The benchmark used is 128-bit AES design with ~ 100,000 gates and 10,688 registers. The cell libraries are purely academic and not of a commercial quality. In this paper, we only show the applicability of Pin-3D optimization to heterogeneous 3D IC. The heterogeneous design shown in this paper is not realistic for multi-voltage designs as we do not add voltage level shifters and it is applicable when the two technologies operate at the same voltage.

5 RESULTS AND ANALYSIS

5.1 Impact on 3D Routing

With Pin-3D, we observe up to 9% lower wirelength compared to the C2D design in Table 3 due to the improved timing and MIV-driven placement legalization. We also see a higher utilization of the MIV layer with Pin-3D routing as mentioned in Section 3.6 due to close approximity of top-most bottom-die BEOL to the top-die FEOL. The top-most metal layers would be relatively less used by the same die FEOL as they are much further away from the cell pins. The top-die FEOL takes advantage of this fact and uses the top metal layer of the bottom-die BEOL. The independent die-by-die routing style of the Compact-2D design cannot exploit this advantage as the routing is done within each BEOL separately. This is the reason we see almost 2× the number of MIVs in Pin-3D routing, compared to Compact-2D. The 3D routing complexity due to the difference in the routing stack between pseudo-3D and the final 3D stage is a major source of discrepancy between pseudo-3D and final 3D stages. This leads to incorrect parasitic and timing optimization in pseudo-3D stages that is corrected with 3D optimization.

[1] Our attempt to use commercial PDKs was not successful as we are not allowed the access to foundry device and interconnect technology files.

Table 2: 3D ECO optimization result on register-to-register paths using Pin-3D ECO. We use Industry-A processor in 28 nm.

Industry-A	w/o ECO	w/ ECO
Frequency	1	1.0
Sequential Cell Area	1	1.000
WNS	1	0.910
TNS	1	0.669
#Violations	449	270
Power	1	1.000

5.2 Impact on 3D Timing Closure

Pin-3D design is optimized as discussed in Section 3.7 with three passes and adjusts for the timing degradation caused by cell displacements, and 3D routing mismatch from pseudo-3D stage. Compared to C2D design of Industry-A, we observe a huge reduction in the total negative slack that is $> 8\times$, and the worst slack decreases 58% as shown in Table 3. In order to fix the timing of the design, the total power increases marginally by ~2% from the addition of more buffers and higher drive strength cells on the critical paths with Pin-3D optimizer. The product of energy and effective delay(= Clock Period−Worst Negative Slack), which captures both the power and the delay variations, is 16.2% lower in Pin-3D. The total positive slack is on par with the 2D design for most of the designs, implying that the designs are not over-optimized with the Pin-3D optimizer. It is only in Industry-A, that the total positive slack (TPS) is ~ 30% more than the 2D counterpart. While this may seem like an aggressive optimization, the total power increase is $< 2\%$ from the Compact-2D, which means that the improvement in the positive slack is not entirely from the additional buffers added, but from the additional routing improvements in Pin-3D.

5.3 Impact on 3D ECO

Table 2 shows the impact of our Pin-3D ECO algorithm presented in Section 3.8. We see that with a very negligible increase in total power and sequential cell area ($< 0.05\%$), almost 40% of the violating / negative slack paths are fixed with ECO, and the total slack is reduced by 33%. The worst slack itself has not decreased significantly because some critical paths already have the maximum drive-strength and minimum V_{th} registers, and thus are not improved by the ECO algorithm. None of the existing pseudo-3D flows offer ECO capability. Thus, we believe that this is the first demonstration of M3D ECO results using a commercial design implemented using a commercial PDK.

5.4 Overall PPA Comparisons

Table 3 shows the overall PPA comparisons between commercial 2D, Compact-2D (C2D) [4], and Pin-3D optimized designs of Industry-A, Industry-B, LDPC, and Netcard. The routing and placement layouts for commercial 2D and Pin-3D designs of Industry-A and Industry-B are shown in Figure 3.

C2D Comparison. When compared to C2D, we see that Pin-3D has better results for almost all the metrics across the four

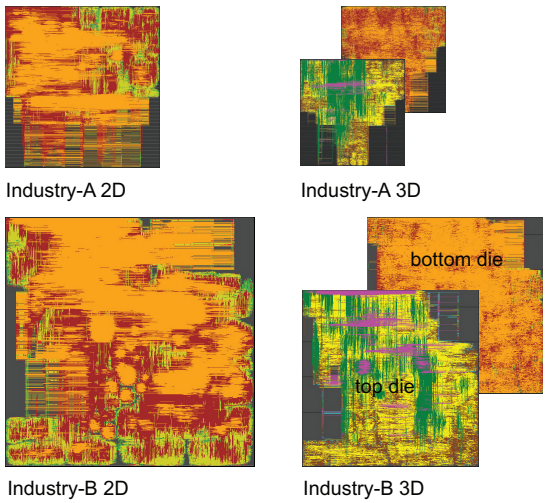

Figure 3: GDS layouts of our Industry-A and Industry-B designs. For 3D designs, we show the placement for the top die, and the routing for the bottom die. We use a commercial 28nm technology in all designs.

benchmarks in Table 3. Moreover, we achieve 12% and 14% better EDP (Energy-Delay Product) than commercial 2D with our Pin-3D for Industry-A and Industry-B circuits respectively. In Table 3, PDP (Power-Delay Product) and EDP uses the effective delay (= Clock Period − Worst Negative Slack = $1/\text{effective frequency}$).

Industry-A Results. Industry-A shows better results with Pin-3D than 2D across most of the metrics except total negative slack. This is due to the clock optimization in the input post-CTS design, which will be later discussed in detail in Section 5.6. Comparing power breakdown between 2D and Pin-3D designs, we see that the most power savings come from the combinational cell power which is 14.5% smaller in the M3D design.

The sequential cell power does not follow the same trend due to high toggle rate at clock pin, leading to large internal power. In 2D Industry-A design implemented, 83% of the flip-flops used are least drive-strength cells and cannot be further down-sized in the M3D design. This limits the sequential power reduction with M3D ICs. This trend is not just specific to the Industry-A design, but is observed in all the circuits design in the 28 nm technology node. Overall, Pin-3D design has 7.9% better total power and 12% better EDP compared to its 2D counterpart.

Industry-B Results. Industry-B follows a similar trend as the Industry-A design, except the wirelength is 2.5% higher in Pin-3D than C2D. Industry-B design has a significant increase in cell utilization(≈8%) with Pin-3D compared to Compact-2D among the four designs considered here. This implies that the cell area also increased by the same percentage in the Pin-3D optimized version to fix the timing violations. Due to this increased amount of cell changes, the amount of routing is increased leading to the increase in wirelength. While the higher wirelength in Pin-3D leads to higher switching power, the higher cell area creates an increased internal, leakage power compared to C2D. Better timing closure, and thus

Table 3: PPA comparisons among commercial 2D, Compact-2D [4], and Pin-3D optimized designs

	Industry-A			Industry-B			LDPC			Netcard		
	2D	C2D	Pin-3D	2D	C2D	Pin-3D	2D	C2D	Pin-3D	2D	C2D	Pin-3D
Target Frequency (GHz)	1	1	1	1	1	1	1.500	1.500	1.500	1.250	1.250	1.250
Footprint (μm^2)	1	0.500	0.500	1	0.499	0.499	111,420	55,692	55,692	525,835	263,272	263,272
Cell utilization (%)	1	0.944	0.963	1	0.882	0.953	77.75	61.04	62.18	73.06	68.02	68.51
Gate Count	1	0.975	0.983	1	0.952	0.962	55,858	48,612	49,248	240,218	229,217	236,330
Total WL (m)	1	0.812	0.740	1	0.786	0.813	2.290	1.526	1.428	9.981	7.308	7.019
MIV Count	–	50,474	104,219	–	168,759	337,032	–	14,574	28,674	–	66,218	141,311
Internal Power	1	0.944	0.976	1	0.915	0.955	90.99	62.41	67.28	73.92	71.29	73.38
Switching Power	1	0.863	0.859	1	0.823	0.868	165.26	112.29	113.76	79.39	62.41	61.07
Leakage Power	1	0.739	0.870	1	0.743	0.819	0.19	0.12	0.13	0.298	0.207	0.225
Sequential Power	1	0.983	0.984	1	0.972	0.997	15.61	15.76	15.88	78.03	71.48	71.01
Macro Power	1	0.999	0.997	1	0.994	0.992		–			–	
Combinational Power	1	0.832	0.854	1	0.810	0.858	236.50	155.40	161.70	54.71	52.44	53.63
Clock Power	1	0.974	1.020	1	0.916	1.008	4.30	3.68	3.56	10.87	9.99	10.05
Mem Input Net Latency	1	0.680	0.724	1	0.558	0.554		—			—	
Mem Output Net Latency	1	0.640	0.578	1	0.632	0.570		—			—	
Mem Net Switching Power	1	1.003	0.837	1	0.752	0.751		—			—	
Total Power (mW)	1	0.905	0.921	1	0.871	0.913	256.44	174.82	181.17	153.61	133.91	134.68
Total Negative Slack (ns)	1	10.786	1.336	1	7.919	1.701	-20.23	-141.21	-47.14	-2.145	-783.74	-2.015
Avg. Negative Slack (ns)	1	1.268	0.714	1	1.761	1.119	-0.011	-0.067	-0.025	-0.012	-0.032	-0.006
Total Positive Slack (ns)	1	0.587	1.345	1	0.508	1.032	662.0	562.8	681.4	5545.3	3955.3	5725.7
Effective Freq. (GHz)	1	0.843	1.023	1	0.844	1.031	1.429	1.221	1.361	1.160	0.969	1.191
Power × Delay (pJ)	1	1.074	0.900	1	1.031	0.885	179.50	143.18	133.21	132.41	138.19	113.13
Energy × Delay (pJ ∗ ns)	1	1.275	0.879	1	1.220	0.859	125.65	117.26	97.91	114.14	142.62	95.02

Table 4: Top 100 critical path averages of register-to-register path group. The metrics for industry processors are normalized w.r.t the clock period.

	Industry-A			Industry-B			LDPC			Netcard		
	2D	C2D	Pin-3D	2D	C2D	Pin-3D	2D	C2D	Pin-3D	2D	C2D	Pin-3D
Clock Period (ns)	1	1	1	1	1	1	0.667	0.667	0.667	0.800	0.800	0.800
Path Slack	-0.041	-0.285	-0.115	-0.157	-0.260	-0.170	-0.025	-0.123	-0.053	-0.019	-0.140	-0.015
Clock Skew	0.050	0.025	0.191	0.135	0.117	0.086	0.005	0.055	0.026	-0.026	0.017	0.003
Setup Time	0.007	0.024	0.014	0.004	0.010	0.004	0.019	0.017	0.018	0.023	0.100	0.032
Path Delay	0.984	1.236	0.912	1.108	1.133	1.080	0.668	0.717	0.675	0.821	0.824	0.780
Cell Delay	0.891	1.188	0.840	0.912	1.066	0.997	0.578	0.663	0.619	0.597	0.688	0.641
Wire Delay	0.093	0.048	0.072	0.105	0.067	0.083	0.090	0.054	0.056	0.224	0.136	0.139

better effective frequency in Pin-3D benefits the EDP, resulting in 14.1% savings with Pin-3D compared to 2D.

LDPC and Netcard Results. LDPC and Netcard are open-source benchmarks, so the design metrics are not normalized. All the power values reported are in mW. When compared to its baseline 2D design, LDPC Pin-3D shows a high power savings of ≈30% with only a relatively small frequency degradation. LDPC is a wire-dominated circuit, as can be seen from the high portion of switching power in the design. So, the wirelength reduction in M3D significantly reduces the output load of the cells, which can then be down-sized without exceeding the delay targets. This leads to 26% reduction in the internal power, which is the most reduction across all the designs. Netcard does not have as high of a switching power proportion and still has a modest power savings of 12.3% with Pin-3D.

5.5 Memory Net Analysis

In this section, we report memory access latency and energy for Industry-A and Industry-B designs[2] as placing cells on top of each other is especially helpful for the macro blocks. In the 3D configuration, the macro blocks are placed in smaller footprint with easier access to standard cells on both tiers. We can see from the Industry-A, Industry-B 2D routing layouts in Figure 3 that the wires over the memory blocks are long due to the size of macros. With 3D placement, these wires become much shorter decreasing the delay of the memory nets. In both the Industry designs, memory net latency in 3D is more than 40% smaller than in 2D. The macro placement hugely impacts this improvement, and it is useful to explore 3D

[2]To the best of our knowledge, this is the first time that detailed 2D vs. 3D IC memory net statistics are reported using GDS layouts and sign-off simulations of commercial RTLs and a commercial PDK.

benefits in designs where memory access through the nets become critical. The C2D memory latency results are quite similar to the Pin-3D designs as the memory net latency is only dependent on the routing and not on the cell optimization. This C2D and Pin-3D latency similarity is also observed in the wire delays of the Table 4.

5.6 Critical Path Analysis

To explore the impact of Pin-3D optimization we analyze the critical paths of the designs from Table 3. We report the average metrics of top-100 critical paths of register-to-register paths. Paths containing the memory or I/O ports have a significant portion of the delay from memory blocks or external delays, and can skew the delay averages and are not included in the analysis.

From Table 4, we see that across all the designs, path slack(= cell delay + wire delay + clock skew + setup time - clock period) is generally the best in 2D, even though the path delay(= cell delay + wire delay) is usually better in Pin-3D designs by as much as 9% in Industry-A. This is because of the clock skew in 2D and 3D designs. A positive clock skew implies that the launch register receives the clock signal later than the capture register, and it decreases the available time for the path delays. Pin-3D clock skew values are worse than 2D in all the cases except Industry-B leading to worse slacks. So, a better 3D clock tree can further improve the timing in M3D designs. Lastly, in C2D designs, we see a notable increase in path delay, which is primarily caused by cell delay increase and the lack of 3D optimization. Similar to the memory net latency, the wire delay here in C2D is comparable to Pin-3D as it is only dependent on the routing, and not the cell types. Since the wirelengths in C2D and Pin-3D are relatively close to each other, we do not see a major difference in this aspect.

5.7 Heterogeneous 3D IC Optimization

Using the technology and design setup as mentioned in Section 4.2, we perform legalization, routing, and timing closure of a heterogeneous 3D IC. The input pseudo-3D design is designed solely with the 15 nm PDK, as no pseudo-3D flow can support multiple technology nodes in a single tier. Area-balanced min-cut tier-partitioning in heterogeneous 3D IC is slightly modified to account for the difference in cell areas in the 15 nm, 45 nm technology nodes. This partitioning achieves correct area-balance with unequal cell sizes within each die. We can observe this in the cell count and cell area of each die in Table 5. The cell area imbalance is within 2% of the total area, while the cell count is split between the two dies in a 2.24:1 ratio. Clock-network and sequential cells are extremely important in overall timing, and are fixed on the top-die.

The optimizer was able to add buffers on both dies, and close the timing efficiently by reducing the worst slack -0.615 ns to ~ -0.051 ns. The fully routed, and optimized layouts are shown in figure 4. Due to the smaller pitch of the 15 nm top-die, and the increased cell count we see major portion of the total power ($\sim 80\%$), wirelength ($\sim 70\%$) on the top-die. The routing difference between the dies can be seen in the zoomed-in shot in figure 4. The huge power in the top-die is due to the imbalanced cell count, advanced technology node, and presence of power-hungry sequential cells on the top-die. This imbalance in total power also allows for an extremely efficient heat dissipation from the top-die.

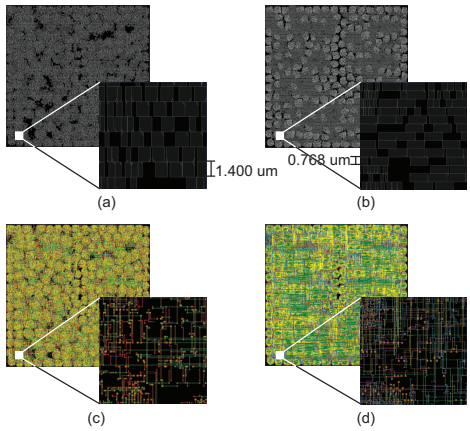

Figure 4: Layouts of 45 nm+15 nm heterogeneous 3D IC design of 128-bit AES with Pin-3D. (a), (b) Full placement in bottom and top dies respectively (c), (d) Full routing of the bottom and top dies respectively with zoom-in windows for each.

Table 5: PPA results of 45 nm+15 nm heterogeneous 3D IC design of 128-bit AES with Pin-3D at 2 GHz clock frequency.

Design Metric	Total	Top-Die	Bottom-Die
Technology Node	Hybrid	15 nm	45 nm
Number of Cell Rows	–	285	156
Cell Area (μm^2)	60,887	29,832	31,055
Gate Count	107,201	74,203	32,998
Buffers Added	3,160	1,091	2,069
Wirelength (mm)	832.2	572.3	259.9
MIV Count	39,237	–	–
Total Power	123.24	104.09	19.15
Critical Path Delay (ns)	0.553	0.051	0.502
Critical Path Cell Count	18	7	11
Footprint (μm^2)		48,246	
Pre Opt. WNS (ns)		-0.615	
Pre Opt. TNS (ns)		-278.4	
Final WNS (ns)		-0.051	
Final TNS (ns)		-1.418	
Clock Tree Statistics			
Buffer Count	463	463	0
Wirelength (mm)	19.64	19.45	0.19
Max Latency (ns)		0.116	
Max Skew (ns)		0.045	

6 CONCLUSION

Our Pin-3D optimizer maximizes the applicability of commercial 2D EDA tools in legalization, routing, optimization, and ECO partitioning of homogeneous and heterogeneous M3D designs. We compared the optimization capabilities with homogeneous M3D ICs and showed its applicability for heterogeneous 3D ICs. Various benefits of heterogeneous 3D IC design and commercial-quality designs will be explored in future work with the Pin-3D flow.

ACKNOWLEDGMENTS

This work is funded partially by DARPA ERI 3DSoC Program under Award HR001118C0096, Samsung Semiconductor Inc., and the Semiconductor Research Corporation under GRC Task 2929.

REFERENCES

[1] L. Brunet and P. Batude et al. First demonstration of a CMOS over CMOS 3D VLSI CoolCube™ integration on 300mm wafers. In 2016 IEEE Symposium on VLSI Technology, 2016.

[2] L. Brunet and C. Fenouillet-Beranger et al. Breakthroughs in 3D Sequential technology. In 2018 IEEE International Electron Devices Meeting (IEDM), Dec 2018.

[3] S. Panth, K. Samadi, Y. Du, and S. K. Lim. Shrunk-2-D: A Physical Design Methodology to Build Commercial-Quality Monolithic 3-D ICs. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2017.

[4] Bon Woong Ku, Kyungwook Chang, and Sung Kyu Lim. Compact-2D: A Physical Design Methodology to Build Commercial-Quality Face-to-Face-Bonded 3D ICs. In Proceedings of the 2018 International Symposium on Physical Design, 2018.

[5] B. Goplen and S. Sapatnekar. Efficient thermal placement of standard cells in 3d ics using a force directed approach. In International Conference on Computer Aided Design, 2003.

[6] M. Hsu, V. Balabanov, and Y. Chang. Tsv-aware analytical placement for 3-d ic designs based on a novel weighted-average wirelength model. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2013.

[7] G. Luo, Y. Shi, and J. Cong. An analytical placement framework for 3-d ics and its extension on thermal awareness. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2013.

[8] Jingwei Lu, Hao Zhuang, Ilgweon Kang, Pengwen Chen, and Chung-Kuan Cheng. Eplace-3d: Electrostatics based placement for 3d-ics. In Proceedings of the 2016 on International Symposium on Physical Design, 2016.

[9] W. Gomes et al. 8.1 lakefield and mobility compute: A 3d stacked 10nm and 22ffl hybrid processor system in 12×12mm2, 1mm package-on-package. In 2020 IEEE International Solid- State Circuits Conference - (ISSCC), 2020.

[10] D. B. Ingerly et al. Foveros: 3d integration and the use of face-to-face chip stacking for logic devices. In 2019 IEEE International Electron Devices Meeting (IEDM), 2019.

[11] S. Panth, K. Samadi, Y. Du, and S. K. Lim. Power-performance study of block-level monolithic 3d-ics considering inter-tier performance variations. In ACM/EDAC/IEEE Design Automation Conference, 2014.