# Pin-3D: Effective Physical Design Methodology for Multidie Co-Optimization in Monolithic 3-D ICs

Sai Pentapati, Kyungwook Chang, and Sung Kyu Lim, *Fellow, IEEE*

***Abstract*—Three Dimensional fabrication and packaging of Integrated Circuits has been proposed as one of the key drivers for More Moore technologies by the IRDS. Such integration is useful to improve the performance and cost effectiveness of the newer generation of chips. Several consumer chips have been using micro-bump-based 3-D package bonding techniques but such integration is only done at a very high level. To fully utilize the benefits of 3-D integration, we propose an effective optimization methodology for 3-D ICs. In this work, we present an all-round physical design methodology to support 3-D IC timing optimization, with features, such as timing driven placement, clock tree synthesis, 3-D timing optimization, and ECO optimization for 3-D ICs.**

***Index Terms*—3D floorplanning, clock tree analysis, physical design, timing optimization.**

## I. Introduction

THE DOUBLING of the transistor density has been a major force in driving the hardware capabilities of new chips along with Dennard scaling. This trend progressed with the help of technological improvements in the complete fabrication stack from technology and manufacturing, to the electronic design automation (EDA) software used to design chips.

As dimension scaling slows down, IRDS proposed few key technology drivers for More Moore scaling [1]. 3-D stacking and integration is one such driver and is the focus of our work here. Good placement, routing, and timing optimization flows are important to realize the benefits of the increased transistor density in 3-D. For 2-D ICs, commercial place and route (PnR) tools are sophisticated to handle the complex problems of placement, clock tree synthesis, routing, and timing optimization. Leveraging such capabilities, pseudo-3D flows were devised to exploit the intricate optimization algorithms of the commercial PnR tools for 3-D IC designs. Shrunk-2D [2], Compact-2D [3], and Macro-3D [4] are three such flows with different capabilities.

In Shrunk-2D and Compact-2D, the 3-D design is mapped onto a simpler 2D-like footprint. With such mappings, 3-D specific information is lost and subsequent optimization will not be an exact fit for 3-D IC designs. In comparison, Macro-3D provides complete PnR capabilities for 3-D ICs, but it is limited to memory-on-logic partitioned 3-D ICs. By placing only memory macros on one of the dies along with some clever routing layer hacks, Macro-3D converts the 3-D problem to 2D-like without losing any 3-D information.

In this work, we propose a well-rounded flow for creating a sign-off quality 3-D IC for any partitioning or pitch type using commercial PnR tools. The main contributions of Pin-3D are the representation of 3-D die in a format that is easily understandable and processed by a commercial PnR tool. This is done using “transparent die” and “pin-projection.” We also propose ECO for 3-D that supports changes to the partitioning solution, improved EDA flows for placement legalization, timing optimization and clock tree methodology for 3-D. We also perform additional analysis related to path correlation and timing optimization that are useful in understanding the working of 3-D flows. This work is an extension of our previous work [5]. In this revision, we added more functionality, such as clock tree optimization for 3-D, cost savings by optimizing metal layers in 3-D, and more detailed analysis and description of the flows. With the help of commercial process design kit (PDK), and test circuits (that include two Arm Cortex processors), we show that the 3-D IC designs with Pin-3D have a 9%–30% power reduction along with a 24%–38% wirelength reduction compared to 2-D ICs. Compared to our conference paper [5], we see an additional 8% reduction in energy delay product with Pin-3D. Pin-3D supports heterogeneous monolithic 3-D IC designs (w.r.t. process nodes) for any partitioning type. Here, we show a 128-bit AES encoder circuit implemented with Pin-3D and cells from two different technology nodes: 1) 45-nm node on bottom die and 2) 15-nm node on top. The Pin-3D methodology achieves a sign-off quality timing closure even on heterogeneous designs. This opens up an interesting dimension for circuit design and optimization as there are multiple process nodes on each die of 3-D IC.

## II. Background

A summary of the differences between Pin-3D and the previous pseudo-3D flows have been provided in Table I. Compared to native 3-D placers, such as [6], [7], and [8], pseudo-3D flows that make use of commercial tools can

Manuscript received 23 January 2023; revised 7 July 2023 and 21 September 2023; accepted 3 November 2023. Date of publication 29 November 2023; date of current version 21 March 2024. This work was supported in part by the Ministry of Trade, Industry & Energy of South Korea under Grant 1415187652 and Grant RS-2023-00234159; in part by the Semiconductor Research Corporation under Grant CHIMES 3136.002; and in part by the DOE Office of Science Research Program for Microelectronics Co-design (Abisko). This article was recommended by Associate Editor P. Gupta. *(Corresponding author: Sai Pentapati.)*

Sai Pentapati is with the Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30318 USA (e-mail: sai.pentapati@gatech.edu).

Kyungwook Chang and Sung Kyu Lim are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30318 USA (e-mail: k.chang@skku.edu; limsk@ece.gatech.edu).

Digital Object Identifier 10.1109/TCAD.2023.3337193

TABLE I
QUALITATIVE COMPARISON AMONG STATE-OF-THE-ART "PSEUDO-3D" PHYSICAL DESIGN TOOLS FOR MONOLITHIC 3-D ICS AND THIS WORK. "ENHANCED DIE-BY-DIE" MEANS THE PINS FROM BOTH DIES ARE VISIBLE DURING DIE-BY-DIE OPTIMIZATION ON A COMPLETE 3-D METAL STACK

| | Shrunk-2D [2] | Compact-2D [3] w/o 3D Optimization | Compact-2D [3] 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 | more holistic 3D flow |
| 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 | optimized for 3D |
| 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 |

leverage a wide range of EDA capabilities that have been developed over the past few decades. As such, the pseudo-3D based flows can result in a well designed, sign-off level 3-D IC design.

Shrunk-2D [2] is one of the first pseudo-3D flows to use commercial PnR tools for a 2-tiered 3-D IC design. The key idea here was to use same footprint as the 3-D IC for placement and routing. With similar total silicon area between 2-D and 3-D ICs, the actual footprint of a 2-tier 3-D design is half of the 2-D IC footprint. Shrunk-2D uses scaled front and back end of the line (FEOL, BEOL) to fit all the cells and routing of the larger 2-D footprint in half the area. This is the *Pseudo-3D stage* of Shrunk-2D. Commercial tool flows can be applied to this without any limitations.

To generate the final 3-D GDS, the pseudo-3D stage is transformed into the two tiered 3-D IC by partitioning using the bin-based Fiduccia–Mattheyses min-cut algorithm [9]. The cells are then scaled up to their original sizes and are legalized to remove any overlaps during the transformation process. monolithic intertier vias (MIVs) that connect the nets crossing the two tiers are placed by performing routing of the 3-D nets on the complete 3-D stack. A die-by-die routing stage completes connectivity within the two tiers, and a final optimization is done in each die separately. The independent die optimization does not rectify timing from the other dies and does not create a fully optimized 3-D IC.

In Compact-2D [3], a new pseudo-3D stage was proposed in order to avoid the dimension scaling as it can lead to design rule violations, pin access issues, RC inaccuracies, software licensing limitations. Here, instead of scaling the footprint, it is kept the same as 2-D, in turn, the unit parasitics are scaled by $1/\sqrt{2}$ as a way of replicating the smaller wirelengths and RCs of a 3-D design. While this is a better-pseudo-3D stage than Shrunk-2D, it still has many of the same problems with regards to 3-D timing closure.

An additional post-partitioning optimization has also been proposed for Compact-2D [3] to close timing in 3-D. A custom metal stack spanning both the 3-D tiers is used to represent 3-D BEOL, and two interleaved half-height rows are used to represent 3-D FEOL. While this helps with some timing optimization, the flow is incomplete as there is no placement or clock tree optimization stage based on the 3-D metal stack. In addition, the half height cells cannot fully represent the design rules, and some pin access violations still remain due to the cell halving.

With Pin-3D, we do not alter the physical or electrical properties of nets or cells after the pseudo-3D stage. Moreover, for the first time, we introduce incremental placement, and clock optimization for 3-D ICs with commercial PnR tools. This is in addition to the commercial routing and timing closure for 3-D ICs. We perform a die-by-die placement and buffer insertion operations, but unlike a traditional die-by-die methodology, the entire timing and 3-D physical context is preserved, leading to improved quality of results (QoR). Sections III and IV provide detailed information on how the 3-D aware die-by-die implementation is achieved.

## III. PIN-3D FLOW ENABLEMENT

### A. Key Idea

To enable commercial 2-D PnR tools for 3-D optimization, the cells from both tiers should be in a single FEOL layer without causing overlaps. In Pin-3D, we include both the dies in the physical design database but only keep one "active die" at any time, and the other die is considered inactive during optimization and we term it a *Transparent Die*.

This is created by converting all the standard cells and macros into "COVER" cells. These cells do not occupy any physical area as seen by the tool, effectively making the die "transparent" [Fig. 1(b)]. Cover cells have no active area, and are traditionally used to represent cells that do not contain any logic. Representing cells of the transparent die as cover cells results in cells without placement obstructions or overlaps. This way, cells from both tiers can be present at a single ($x$, $y$) location without causing placement overlaps between the two. But since the timing and power characteristics are controlled by separate library files, the physical alterations to the cell types do not impact their electrical properties, such as rise/fall time, slew, power consumption, and parasitics.

Apart from the cell representation, entire 3-D metal stack from both dies is also required for accurate design as this would allow for high-quality routing in both tiers. For the

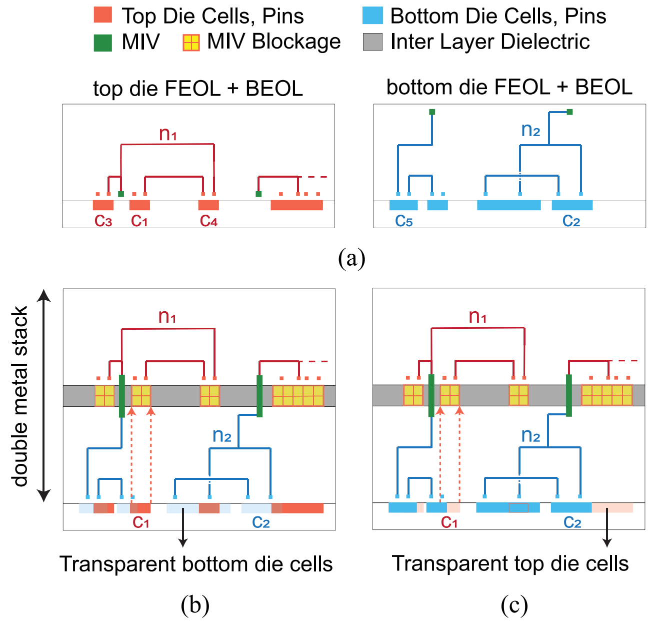


Fig. 1. Key idea of Pin-3D: die merging and pin projection. (a) Top and the bottom dies separately, (b) merged dies for the top die optimization, and (c) merged dies for the bottom die optimization. Our double metal stack contains pins from both dies to provide the entire 3-D context during die-by-die legalization, routing, and timing closure. Top die cells are also projected to the MIV layer to ensure no overlap between MIV and routed nets. Moreover, Pin-3D allows design with two different technology nodes as demonstrated in Section VII-D.

router to connect nets accurately to the cell pins, the pin shapes for each cell should be present on correct layers in the 3-D metal stack. This is achieved through *Pin Projection*, where pin shapes are projected to correct layers. We see this in Fig. 1(b) where the cells of the top-tier (colored in red) are placed in the bottom but their pins are located on the corresponding layer. The cells from the top-tier are projected to the top metal layer rather than placing on metal layers right above the cells as would be expected from a traditional cell design. The Pin Projection is enabled by hacking the BEOL and FEOL tech files for 3-D.

### B. Enabling Commercial Tools for 3-D IC

As with any commercial PDK, only the technology design files for a 2-D IC design are available from the foundry. To create 3-D technology files, we modify these files to support the aforementioned changes required for a good 3-D design as well as keeping the original foundry properties unchanged for the most part. For the 3-D BEOL, we need layout exchange format (LEF) files defining all the routing rules for the metal layers, an interconnect technology (ICT) file with metal layer parasitic information and the complete metal stack information, such as the dielectric information and the thickness, heights of each layer in the BEOL.

To represent the 3-D FEOL for standard cells and hard macros, we require macro LEF files for physical information, a liberty timing file for lookup tables of timing and power characteristics. In addition to these technology related files, design specific files are also needed, such as the netlist (for logical connectivity), clock and timing constraints file (for timing targets and path related information), and optionally a design exchange format (DEF) (for physical information of the circuit such as the placement and routing layout). The design specific files are fairly similar to those of a 2-D design.

### C. 3-D BEOL Creation and Pin Projection

In our work, the 2-tier 3-D ICs are constructed with 6 metal layers per die. These are labeled as M1_bottom, M2_bottom, ... M6_bottom, M1_top, M2_top, ... M6_top in order from bottom to top. In the newly created LEF file with double metal stack, the routing rules for each Mx_bottom and Mx_top are assumed to be same as the Mx layer from 2-D LEF file provided by foundry. This makes sure the generated 3-D LEF also contains all of the routing rules required by the foundry for a 2-D design. A via layer is required between every two consecutive metal layers and is defined following the same methodology we used for metals. Notably, there is no parallel for the MIV in the 2-D PDK as it connects the two tiers together in a 3-D IC design by forming a new cut layer between M6_bottom, M1_top. Only the required routing rules, such as width and spacing, are created in the LEF file for the new cut layer. The ICT file with layer parasitic information and the metal stack structure is similarly extended to 3-D using the corresponding information from 2-D.

### D. 3-D FEOL Creation and Transparent Cells

To accurately represent the cells from multiple tiers together within PnR tools, it is required to differentiate the cell names, and to also have pin shapes on correct layers. To do this, we duplicate the 2-D macro LEF files into two identical files for top and bottom die. The bottom and top die cells are renamed by adding custom suffixes to differentiate between the cells from each die. The layer names in top die cells are further modified according to the top die metal layer naming conventions so that the pin shapes are located on the correct layers in 3-D stack. Since an MIV has to pass through top-tier FEOL to access the M1_top, they cannot be located at the same location of cells in the top-tier. A 3-D specific routing rule for each top-tier cell is added by creating MIV layer obstructions of the same size as the cell. This ensures every instantiation of the top-tier cells in the 3-D design carries along an obstruction that restricts the tool from adding MIVs overlapping with top-tier cells.

*1) Transparent Cells:* A major limitation of the commercial PnR tools for 3-D placement and timing optimization is that they can only have a single FEOL layer. Compact-2D uses the half-height row cells in order to avoid overlaps between cells from different tiers during 3-D routing and optimization. But, as this creates pin access issues, it is not generalizable for placement, clock tree optimization, or heterogeneous 3-D ICs. In Pin-3D, instead of halving the cell, we conditionally convert it into a transparent cell that does not have any active area. By selectively modeling the cells of either top or bottom die as transparent, we successfully remove the overlaps from the cells of transparent dies. The nontransparent die can undergo all of the cell sizing, insertion, deletion operations allowing for full suite of commercial tool capabilities. We introduce two flavors of LEF files to be used during different stages of the flow. Fig. 1(b) and (c) shows the two different scenarios in which

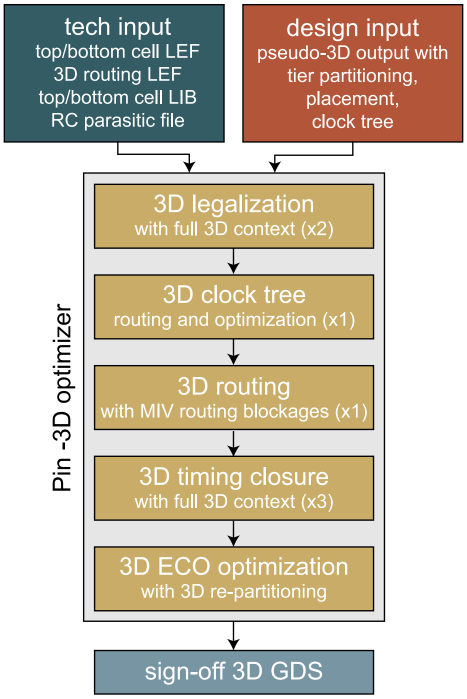


Fig. 2. Our Pin-3D optimizer design flow. The numbers in the braces (× . . . ) specifies number of times a stage is repeated for a 2-tiered 3-D IC. For instance, legalization is performed two times in total, once per die.

cells from different tiers are turned transparent. The bottom die cells are turned transparent during the top die optimization stages and vice-versa.

## IV. Pin-3D Design Flow

The overall flow of the Pin-3D optimizer is shown in Fig. 2. The design input of the Pin-3D optimizer can be from any pseudo-3D flow irrespective of the partitioning type or technology node. This gives the Pin-3D flow a good starting point for 3-D optimization. A bad starting point can significantly worsen the design QoR since the optimization is done in a step by step fashion as will be discussed later in Section IV-D. The technology files required to represent the 3-D design are discussed above, and the various stages of the flow are discussed in this section.

*1) Initial Partitioning* The design input to Pin-3D is from any of the pseudo-3D flows that were discussed in Section II. In pseudo-3D flows, such as [2], the partitioning step is responsible for die assignment of the cells. This is achieved using an area-balanced min-cut algorithm [9] in order to reduce the number connections between the tiers. 3-D connection between any two cells adds additional complexity to routing since the routes need to pass to/from different tiers. So the min-cut algorithm is used to make sure the densely connected cells are not separated across the two tiers. One of the main downfalls of using such connectivity-based min-cut is in terms of the placement of the cells. After an initial optimization stage, as with the pseudo-3D flows, densely connected cells are clustered together in the coordinate space. So the cell placement in each tier after partitioning would be very clustered. This leads to worse-3D QoR, and so a bin-based approach is taken for 3-D partitioning in [9]. Cells are placed into different bins based on their placement in the pseudo-3D design. Min-cut is performed within these bins separately so that large cell clusters are not formed, and area across the two tiers is balanced within each bin separately. Note that, macro placement and how such macros are handled during partitioning is an important part of this stage and we use the same flow that the authors introduced in [9] In this work, we add an incremental placement stage in Pin-3D along with global 3-D routing so that the 3-D routing can be used to guide the placement legalization after partitioning.

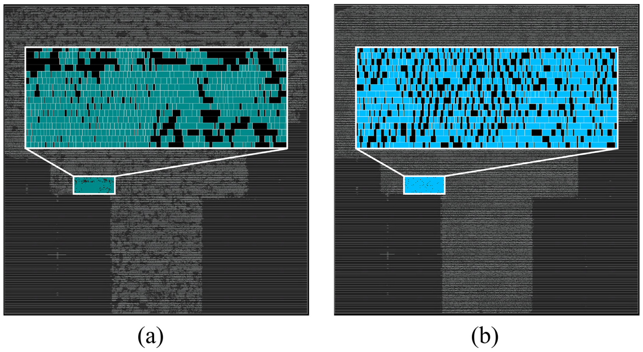


Fig. 3. Standard cell placement of Cortex-A7 and zoom-in at a specific location using (a) Compact-2D legalization and (b) Pin-3D legalization. Dense cell clusters is bad for M3-D routing. Tier-partitioning and prelegalized cell placement is the same between the two.

### A. *Incremental Placement With Global Routing*

During incremental placement, cells are displaced to remove any overlaps and to optimize placement based on the 3-D connectivity. To allow for larger number of white spaces between the cells when possible, the top die cells have an added soft padding of one SITE between the cells in a row. This generates a more porous top-tier cell placement that allows more locations to penetrate through for the MIVs. The top-die standard cell placement in Fig. 3 clearly shows the difference between legalization results of Pin-3D flow compared to C2D. With a soft padding, cells are only padded with white spaces when there is no significant impact on design metrics, such as power, performance, and area (PPA). Placement is still a die-by-die procedure in Pin-3D (indicated in Fig. 2 by the *x2* in legalization step) but with the full 3-D context (timing, connectivity). This is due to the *transparent die* whose standard cells are turned into COVER cells and cannot be used for legal placement. The order of die-by-die placement is not of a concern as long as the partitioning is balanced in terms of timing. If the partitioning is unbalanced with one die having more standard cells and paths with higher-timing criticality, it is better to start the operations on the critical die. This allows for better tuning the critical die directly. Otherwise, since Pin-3D has the full timing information, starting placement optimization on a die with lower criticality, the tool tried to optimize the critical die indirectly and degrades the optimization quality.

Pin-3D's design flow also includes a trial routing stage during the legalization so that the placement and routing are interdependent to create better-congestion-driven and/or timing-driven placement and routing. This is due to the better use of commercial tool's placement engines rather than a simple legalization that has been done in previous 3-D works.

*1) Placement in the Presence of Power Network:* The power network is usually defined before the 3-D legalization stage (not explicitly shown in Fig. 2), and so the tool has to honor the presence of existing routes when placing the cells and dealing with cover cells of the transparent die during placement step as discussed earlier and subsequent routing stages.

The pin-projection step along with the transparent cell creation works well together to honor such preexisting routes in the design. For example, consider a presence of power delivery network in all the layers of the 3-D stack of Fig. 1(c). The cells from the bottom die will have pin shapes in the bottom metal stack, so only the metal shapes, such as PDN or other existing routes, in the bottom tier interact with the bottom tier cells. The top-tier metal shapes will have no direct impact on the placement and routing of cells in the bottom tier. Similarly, the pin-shapes within the transparent top-tier cells are located on the top layer stack and are not impacting the placement and routing of bottom tier cells. Since the top-tier cells are transparent in this step, they do not change their placement at this step. This is one of the limitations of using cover cells for representing a transparent cell. Note that the top-tier cells are then placed and routed during the second stage of place/route in Pin-3D where the die is more accurately represented by Fig. 1(b).

### *B. Clock Tree Optimization*

Clock tree synthesis and optimization are crucial for any clock-based IC design, and none of the previous flows have considered optimizing the entire clock tree for 3-D IC design. There have been a few studies that cluster or partition the clock buffers or flip-flops to minimize detrimental skewing between the flip-flops in 3-D, but these are simply heuristics and only have a limited effect. In Pin-3D we enable clock tree synthesis and optimization for 3-D ICs. Therefore, instead of manually tuning and controlling clock tree with clustering techniques, we can use the commercial PnR tool's clock tree capabilities for the 3-D IC design.

In our flow, clock tree synthesis and optimization takes place on the top-die after placement legalization. To generate a high-quality clock tree in 3-D, we identify and move all the clock tree buffers and inverters to the top-die before performing clock optimization. In a typical clock-tree network, the number of buffers is much smaller compared to the sequential cells at the leaf nodes of clock tree. So simply changing the clock combinational cells does not significantly alter initial partitioning solution. Proper clock skew assignment with clock optimization is important and can drastically reduce the total negative slack of the entire design as the clock signal reaches every launching and capturing flip-flops. This allows the tool to fully optimize the clock tree network in a single step, rather than in a die-by-die fashion that can hurt the overall clock tree. During this optimization, the clock latency to the leaf cells/clock sinks as well the clock skew between any pair of sinks are freely optimizable by the tool.

TABLE II
WORST AND TOTAL NEGATIVE SLACK TREND IN PIN-3D, AND THE EFFECT OF THE CLOCK OPTIMIZATION STAGE FOR CORTEX-A7. ALL SLACKS ARE NORMALIZED W.R.T. THE CLOCK PERIOD

| | Worst Negative Slack | | Total Negative Slack | |
|---|---|---|---|---|
| Design Stage | No ccopt | With ccopt | No ccopt | With ccopt |
| Legalization | -0.753 | -0.753 | -1799 | -1799 |
| Clock | – | -0.029 | – | -7 |
| Route | -0.741 | -0.039 | -2157 | -82 |
| Timing opt1 | -0.052 | -0.020 | -103 | -4 |
| Timing opt2 | -0.037 | -0.008 | -41 | -1 |
| Timing opt3 | -0.032 | -0.004 | -24 | -0 |

The resulting clock tree from this methodology shows a beneficial impact on the timing closure. As shown in Table II, the implementation with 3-D clock optimization (ccopt) stage has achieved significantly improved timing closure. The place stage in Table II is combination of both bottom and top die legalization based on the 3-D global routing solution, after which the timing information is extracted. We see that the clock tree stage has effectively removed most of the negative slack after the global route stage. By accurately modeling 3-D design data, the clock tree optimizer was able to generate useful skews from timing paths with positive slacks to be used toward negative slack paths. Clock Tree optimization is only done in a single die as breaking up the clock tree into different stages created undesired affects during the optimization. By splitting the clock tree into two tiers, and optimizing them in a die-by-die fashion, the skews between the clock sinks from different tiers were not optimized well and significantly affected the overall design quality. While the clock tree is restricted to a single die, the skews to all the paths can be addressed with the same ease as a 2-D clock tree. The additional complexity due to the clock buffer and leaf being placed on different dies is the additional routing across the dies required to make full connectivity, and this can be easily modeled in the Pin-3D flow by the commercial tools. A clock buffer cell on the top die can have leaf cells on different tiers without any issue in achieving routing connectivity of clock optimization.

### *C. Routing*

Routing of the 3-D IC is a one-step process unlike the placement. The transparent cell method that mandates die-by-die placement does not affect the routing, which is simply done to lay down physical wires connecting the entire 3-D design. Pin projection method takes care of accurate routing to cells, and since the transparent cells do not have any timing modifications, the router can easily consider their delay and loading effects when routing both the standard cells and the transparent cells in a single pass. We see that from Table II, the worst and total negative slacks both see a slight degradation after the detail routing step. This is because of the inaccuracies in timing estimation between global and detail routing stages. This is not specific to 3-D IC, and is an artefact present in any design flow with modern PnR tools due to their split global and detail routing stages.

### D. Timing Closure

Timing closure of any design is crucial for achieving a sign-off quality commercial design. With Pin-3D, as the cell insertion can only occur in one die at a time, the timing closure is done in three distinct steps as hinted in Table II.

After routing the complete 3-D design in the previous stage, the design flow works on the timing optimization of paths by initially inserting cells in the top die. During top-die cell insertion, the tool can still modify the routing of the nets connected to the cells on bottom die. As a result, a timing limited net on the bottom die can be split up by adding a buffer on the top die and resolving some of the timing issues. This will not yet be optimal as the bottom die cells are left untouched. A second round of optimization stage then concerns with the cell resizing, removal and insertion of buffers on the bottom die which would then solve most of the timing issues present in the design.

But a third round of optimization is run, this time on the top die, to finish the timing closure stage of Pin-3D. As the top die cells had to compensate for the nonoptimized nature of bottom die cells in the first iteration, some of the cells will be aggressively scaled during the initial iteration of the design. Therefore, we run this third optimization stage on the top die, after optimizing the bottom die to reclaim any leakage, area by resolving such aggressively scaled cells.

We see that both the worst and total slack improve with number of optimization stages. But the marginal improvement we get from more iterations comes at a significant run-time cost. As seen in Table II, the timing is mostly resolved at the end of the second timing iteration, making the final stage of optimization not very useful. This is not the case for the design without ccopt, whose worst slack has started to plateau at around −0.032 ns, but the total slack is still recovering albeit slowly.

### E. ECO

Finally, ECO is an important stage of the commercial design flows, which is used to manually adjust the path timing, leakage, or other violations that might need to be addressed using custom scripts. In 3-D, ECO is also necessary to change the tier allocation of cells in the post-route stage along with traditional 2-D functions, such as add/modify cells within a tier. Pin-3D's technology and design setup allows all of the aforementioned moves and is shown using a simple flip-flop sizing algorithm.

As the cells of the two tiers are differentiated based on names, we can move an inverter in the bottom tier to the top-tier by replacing it with the top-tier cell using the "eco" commands of PnR tools. Since the MIV blockage is defined within the cell definition, replacing the cell type from bottom to top would automatically create required blockages for routing. Similarly as the pin projection is also done within the cell definition, they also appear on the correct tier after the eco change. The incremental eco routing provided within the PnR tools can then route to the newly moved cell with accurate routing restrictions.

Here, we first note that the clock-to-output delay of the flip-flops contribute to more than 10% of the total path delays in our 3-D designs. This is significant when considering that the critical paths are in some cases have a logic cell depth of ≈40. To address this, we perform ECO using Algorithm 1.

**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
```

By up-sizing the registers on the critical paths we can improve overall timing. Simultaneously, by down-sizing the registers on paths with large negative slacks, the overall power consumption can be kept in control. We identify these two sets of flip-flops based on the path slacks. A 150 ps threshold is used for the positive timing path groups as the path delay would degrade when replacing it with a smaller drive-strength register.

The critical registers are swapped with register of same type with higher drive strength. In case, the register under consideration is of the highest drive strength, we replace it with the same register, but from the lowest-threshold voltage type. The same process is done with the non critical register but in the opposite direction (replacing with lower-drive-strength cells, or with highest-threshold voltage type).

## V. Experimental Setup

### A. Homogeneous 3-D ICs

The performance and efficiency of the Pin-3D design flow is compared with the Compact-2D (C2D) flow. We use C2D without the post-tier partitioning as the version we have does not supported for technology nodes with complex design rules. We tried to the best of our effort to fix the parsing issue to no avail. We also add the 2-D designs for reference so that we can see the complete picture of 2-D versus 3-D, as well as C2-D versus Pin3-D. All the designs are implemented with a commercial 28-nm PDK, with the 3-D technology files generated as specified in Section III.

Two application processors: 1) Cortex-A7 and 2) Cortex-A53, are used as test circuits that validate our flows with commercial designs. The results for these circuits are normalized w.r.t. corresponding 2-D designs as per our NDA with

TABLE III
PIN-3D VERSUS COMPACT-2D [3] ON DIFFERENT ASPECTS OF THE 3-D DESIGN AFTER LEGALIZATION. WE USE CORTEX-A7 IN 28 NM. HIGHLIGHTED VALUES IN RED SHOW THE METRICS OF INTEREST

| Cortex-A7 | | C2D [3] | Pin-3D |
|---|---|---|---|
| Legalization<br>Displacements | Top-Die Avg.<br>Bottom-Die Avg. | 1<br>1 | 0.42<br>0.40 |
| Routing<br>Metrics | Total Wirelength<br>MIV Count | 1<br>50,474 | 0.910<br>104,219 |

Arm. Both of the Arm processors are configured with one core with 32-kB L1 instruction cache, 32-kB L1 data cache, floating point unit, and Arm's NEON SIMD unit. Both the netlists are dominated by logic cell area rather than memory macro area. OpenPiton [10] is another open-source processor design that we have used in some analyses along with the Cortex-A cores.

Two open-source pure-logic test circuits: 1) LDPC and 2) netcard, are also used for analyses. These help us to show the raw design metrics of different design implementations for an in-depth analysis, and to add variety to the test-bench.

### B. Heterogeneous 3-D ICs

Pin-3D flow enables the design and optimization of heterogeneous 3-D IC designs. A proof-of-concept of the design is shown using the open-source nangate 45-nm and nangate 15-nm PDKs. While, the 3-D technology file generation is same as the process described in Section III for homogeneous files, there are a few other restrictions that constrain the design process. This is discussed in detail in Section VII-D.

*a) Tool setup:* In all the implementations shown in this article, RTL synthesis is done using Synposys' Design Compiler and the physical design using Cadence Innovus. Power and timing analysis were done with Cadence Tempus.

## VI. PPA BENEFITS OF THE PIN-3D STAGES

### A. Placement Legalization for 3-D

From Table III, we see that the legalization using Pin-3D manages to achieve 50%–60% lower-average displacements compared to the Compact-2D (C2D) flow. This is because the legalization in each die is guided by the full 3-D placement of the IC. With the incremental placement of Pin-3D, the tool not only removes cell overlaps but also does so to improve timing, congestion, and other design metrics using global routing. With die-by-die legalization of C2D, once the dies are separated after partitioning, the displacements in one die do not impact timing or connectivity of the other die. The displacements resulting from the legalization stage contribute to the mismatch of placement information between the pseudo-3D and the final M3-D stages. For example, if even a single cell on the critical path has large displacement it increases the wirelength and therefore the wire load connected to the cell, and leading to a worse-critical path delay. For the Cortex-A7, legalization takes ≈0.75 h in total.

### B. 3-D Clock Tree Optimization

The run-time and timing closure benefits of clock tree optimization were discussed in Section IV-A1. Here, we analyze different clock tree metrics and the impact of fixing the clock buffers and clock logic on the top die. Fig. 4 shows the two different scenarios during the clock tree optimization stage. As mentioned in Section IV-A1, during this stage, we fix the clock combinational cells to the top-die as depicted in Fig. 4(b). As clock optimization is done during top-die optimization in our implementation, it is better to fix the clock cells to this die. Having entire clock tree (expect the sinks) in a single die allows for a much easier skewing of the clock paths to the sequential cells. The tier assignment of sequential cells is not important for clock skewing. Since cells are correctly represented in terms of timing, and routing can be done to any top or bottom die cells without any issues, the sequential cells (sinks) do not need to be modified to balance skew.

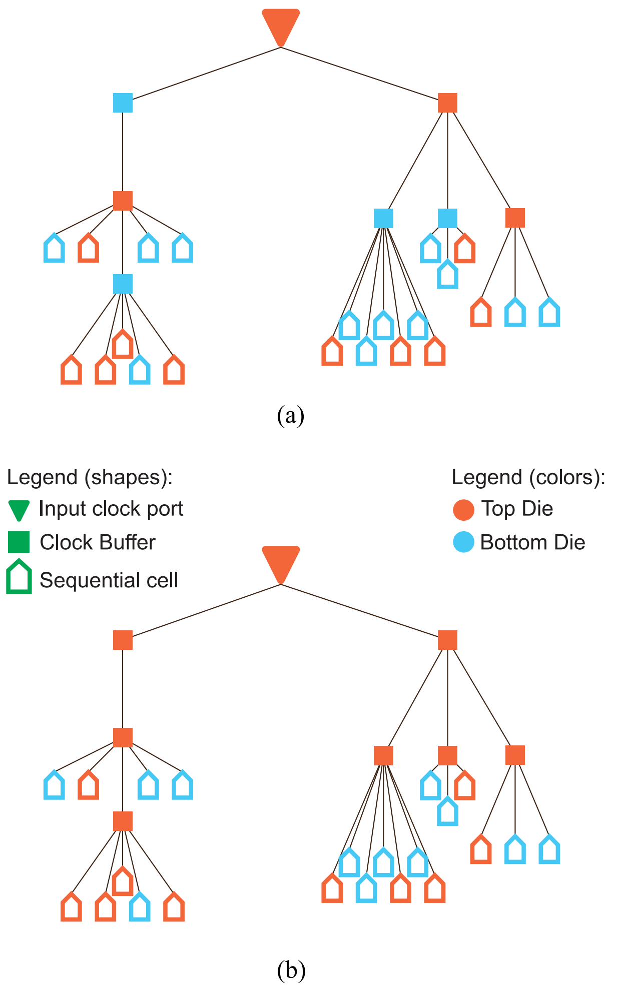


Fig. 4. Example clock tree network showing input clock, clock buffers, and sequential cells. (a) Clock buffers allowed to be placed on both tiers. (b) Clock buffers moved to the top-tier.

As opposed to clock buffer fixing, if the buffers are allowed to be placed freely as shown in Fig. 4(a), the buffers on the bottom die would be nonmodifiable and break the clock tree. So the skew between pairs of sequential cells that are driven by bottom-die buffer cannot even be adjusted directly. To control skew between such sink cells, the tool breaks the nets and adds new clock buffers on the top die. Because of this, we see that in Table IV, the clock tree obtained for a same netlist is much larger when the clock buffers are not fixed to a single tier. The total number of cells is more than 50% of the clock tree

TABLE IV
CLOCK TREE STRUCTURE AND RELATED METRICS OF A NETLIST WITH AND WITHOUT FIXING CLOCK COMBINATIONAL CELLS ON TOP-DIE

| Metric | Units | Without fixing | With fixing |
|---:|---|---:|---:|
| Total Cells | - | 855 | 538 |
| Top Die Cells | - | 507 | 506 |
| Bottom Die Cells | - | 348 | 32 |
| Max depth | - | 24 | 17 |
| Total Area | $\mu m^2$ | 1217.3 | 871.2 |
| Wirelength | mm | 86.4 | 77.6 |
| Max Latency | ns | 0.818 | 0.685 |
| Max Skew | ns | 0.416 | 0.353 |

with buffer fixing. This in-turn creates the larger wirelength, worse latency and max skew of this clock tree. As clock nets are some of the most active signals with the highest-toggle rates, minimizing the clock network area is important to reduce overall power, as they reduce the input pin capacitance contributed by the clock tree cells. Clock optimization takes ≈1 h for Cortex-A7.

### C. 3-D Routing

The global routing aware placement of Pin-3D allows for a better-routing solution. Table VII show that these improvements lead to smaller routed wirelength by up to ∼8% with Pin-3D compared to the C2D design, and around 25% smaller wirelength than the 2-D designs. This leads to better-switching power with 3-D due to the decreased wirelength and wire capacitance load. Another important difference in the routing between the C2D and Pin-3D is the number of MIVs used for routing. Pin-3D designs have ∼2× the number of MIVs of the C2D design. This is due to the full 3-D routing using complete metal stack in Pin-3D. From Fig. 1, we see that the M5_bottom, M6_bottom layers are closer to the FEOL layer of the top-tier than M5_top or M6_top. Additionally, the bottom tier FEOL does not use much of the tracks in M5_bottom, M6_bottom as the track usage of any design decreases further away from FEOL. This leaves the routing tracks in top most metal layers of the bottom BEOL to be utilized by the top-tier nets. Such track sharing adds additional MIVs on nets that do not require MIVs and helps to reduce congestion by distributing routing more evenly across the 3-D metal stack. As the routing done in the 3-D metal stack is very different from the routing in the pseudo-3D stage. So cell sizing and buffering is necessary to achieve timing closure. This is also seen in Table II, where the no clock opt design with just placement and legalization has a large negative slack. Routing stage takes an average of ≈ 1.5 h for Cortex-A7.

### D. Timing Closure for 3-D

Fig. 5 show the evolution of timing path delays between the pre and post partitioning stages. The final pseudo-3D stage with just the global routing parasitics is used for the prepartitioning stage. At this stage, all the cells are placed in a single tier and the routing is done only on a 2D-like BEOL. After partitioning, the cells now have a new z-location corresponding to the tier assignment. The significant changes to routing BEOL and the placement leads to the deviations that are shown in Fig. 5. Here, the worst-critical path to each register in the design is plotted pre and post partitioning on the $X$ and $Y$ axes, respectively. The points are colored by the number of 3-D nets (nets having MIVs) on the critical path, and we see that the paths with more 3-D nets deviate further away from the ideal (solid line in red, where timing of a path pre and post partitioning remain unchanged, i.e., $x=y$). Even in paths with no or few 3-D nets, some paths exist that deviate significantly from the ideal as seen in the path near the bottom left. This is because, on short paths, the 3-D legalization can lead to relatively large displacements in cell positions and affect the short nets significantly.

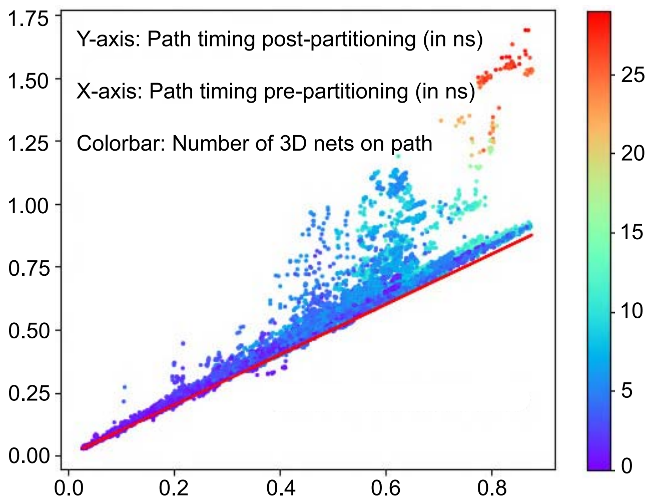


Fig. 5. Path delays of a design before and after tier partitioning. The red line represents the line along which the delays are equal, i.e., the path timing does not change after partitioning.

For example, consider a short net on one of these small delay paths, driven by a buffer of minimum drive strength in the pseudo-3D stage. Even if the cells connected to this net remains on a single tier, after partitioning the pseudo-3D design the legalization leads to movement. Even if the displacement is as small as 2 μm, a short net with a half-perimeter wire length (HPWL) of 1 μm before partitioning can transform into net with HPWL around 3 μm. Due to the small drive strength of driving cells, the cell delay is more sensitive to the increase in wirelength.

As we go further right along the $x$-axis, the deviation between the path delays pre and post partitioning increases as the number of 3-D nets increase. The more 3-D nets are on the path, the higher is the deviation from the ideal since the routing difference adds up over the entire nets. In addition, the long paths generally have large cell drive strength, and are generally more tolerant to additional routing overhead. This is why, the long paths only show deviations from ideal at higher number of 3-D nets per path.

Using the three stage optimization methodology presented in Section IV-D, we perform the timing closure to resolve the deviations caused from 3-D partitioning, placement, and routing. The clock and timing optimization stages together result in a total slack reduction of up to ∼91% compared to C2D design of the Cortex-A7 benchmark. This is from better-skew assignments and cell sizing, but the increase in

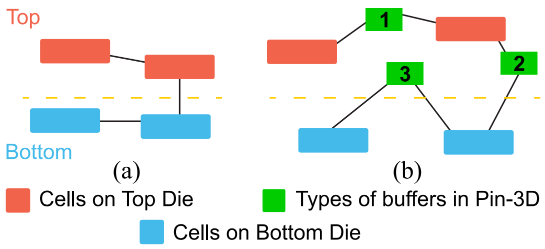


Fig. 6. Example logical connectivity of netlist (a) Before top die optimization, (b) After top die optimization showing three different types (1, 2, 3) of buffer insertions shown in green.

power consumption is $<2\%$ showing the efficiency of the Pin-3D methodology. The worst slack also improves significantly resulting in a 20% better frequency (=18% lower-effective delay). Overall timing optimization takes ≈3 h for Cortex-A7. In total, Pin-3D stage takes ≈6 h. Additionally, the pseudo-3D stage takes around 5 h. This is comparable to the 2-D stage runtime of ≈9 h.

*Efficiency of Pin-3D Optimization:* A key benefit of Pin-3D is the presence of entire design information at each stage of the flow. This has two advantages in terms of design optimization.

1) It allows for any change in the design to be fully 3-D aware.
2) During optimization of, say, bottom tier, the nets in the top-tier can still be modified by allowing for buffers to be inserted on the net.

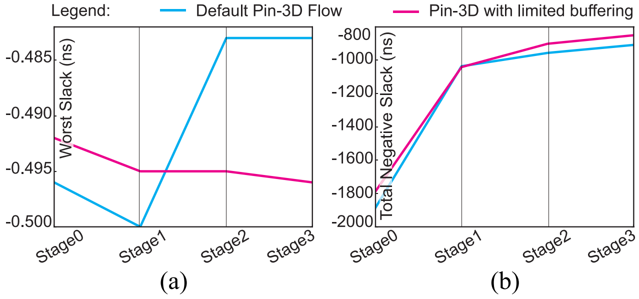
Due to these net modifications of the transparent/fixed die, timing optimization is much faster than otherwise. Fig. 6 shows the three different types of buffer insertion during an optimization stage in Pin-3D.

To demonstrate, we compare two processor designs optimized two different ways with Pin-3D: 1) including and 2) excluding the capability of buffer insertion on fixed die nets. During top-die optimization, the nets that only connect to the cells in the bottom die are marked as do not touch by the tool. This lets the logical structure of the net to remain unchanged while allowing any modifications to the physical layout of net during global and/or detail routing. Similarly, constraints are also applied vice versa during bottom die optimization. Table V shows the PPA impact in Pin-3D without the additional type-3 buffer insertion. We see that it leads to a worse-total slack and therefore a smaller effective frequency. The total number of buffers added during Pin-3D optimization is not very significant as it is at the post-route stage. But of the 2200 buffers inserted, 300 of them were type-3 that are inserted by breaking the net(s) in the transparent tier. While the PPA impact is negligible, the main benefit of this type-3 insertion is the number of optimization stages needed for timing closure.

Fig. 7 shows the evolution of the worst and total negative slacks in the 3-D OpenPiton design. Stage0 refers to the post-route stage, where no 3-D optimization is performed. Stages 1–3 are the three stages of optimization (top-tier, bottom tier, top-tier) of Pin-3D. First, we see that the worst-negative slack by excluding type-3 buffers (pink) is not able to match the default (blue) WNS value in 3-D. This is where the type-3 buffers show their importance. Since we only have a few buffer insertions and even fewer type-3 buffers inserted in a design, the total negative slack is not affected significantly.

TABLE V
EFFICIENCY OF THE PIN-3D OPTIMIZATION IN TIMING CLOSURE. CRITICAL PARAMETERS FOR CORTEX-A7 ARE NORMALIZED W.R.T. THE 2-D DESIGN

| | | OpenPiton | | Cortex-A7 | |
|---|---|---|---|---|---|
| | Units | Default | Exclude3 | Default | Exclude3 |
| Footprint | mm$^2$ | 0.6032 | 0.6032 | $\alpha$ | $1.0\,\alpha$ |
| Std. Cell Area | µm$^2$ | 0.3242 | 0.3237 | $\beta$ | $1.0\,\beta$ |
| Total Buffers | – | 14051 | 13474 | $\gamma$ | $0.999\,\gamma$ |
| Worst Slack | ns | -0.483 | -0.496 | -0.280 | -0.293 |
| Effective Freq. | GHz | 0.870 | 0.860 | $\delta$ | $0.987\,\delta$ |
| Number of Buffers Added During Pin-3D | | | | | |
| Total | – | 2232 | 1620 | 616 | 602 |
| Type1 | – | 843 | 783 | 213 | 278 |
| Type2 | – | 1046 | 837 | 329 | 324 |
| Type3 | – | 343 | 0 | 74 | 0 |



(a) (b)

Fig. 7. (a) Worst-negative slack and (b) total negative slack trends during the three stages of Pin-3D optimization. Limited buffering specifies Pin-3D optimization with no type-3 buffers (refer to Fig. 6).

TABLE VI
3-D ECO OPTIMIZATION RESULT ON REGISTER-TO-REGISTER PATHS USING PIN-3D ECO. WE USE CORTEX-A7 IN 28 NM

| Cortex-A7 | 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 |

## E. ECO for 3-D

The PPA results of the ECO algorithm introduced in Section IV-E are shown in Table VI. By identifying registers to both up-size and down-size, we see that the overall increase in the flip-flop area and power is negligible. The number of violating paths (defined as unique begin-end point pairs) have reduced from 450 without ECO to 270, which in turn leads to a 33% reduction in the Total Negative Slack. It is important to note that the worst-negative slack does not improve by much with the ECO method suggested. This is because the total negative slack of a design is a compound of all the violating paths, but the worst slack is only from a single path. By adjusting the slack of each path by a small amount, we can reduce the total slack significantly. The worst path is not

TABLE VII
TSMC 28-nm Benchmark PPA Comparisons Among Commercial 2-D, Compact-2D [3], and Pin-3D Optimized Designs

| | Cortex-A7 | | | Cortex-A53 | | | 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 | 0.956 | 1 | 0.882 | 0.933 | 77.75 | 61.04 | 61.95 | 73.06 | 68.02 | 67.16 |
| Gate Count | 1 | 0.975 | 0.986 | 1 | 0.952 | 0.964 | 55,858 | 48,612 | 49,009 | 240,218 | 229,217 | 232,366 |
| Routing Statistics | | | | | | | | | | | | |
| Total WL (m) | 1 | 0.812 | 0.750 | 1 | 0.786 | 0.759 | 2.290 | 1.526 | 1.418 | 9.981 | 7.308 | 7.040 |
| MIV Count | – | 50,474 | 118,517 | – | 168,759 | 363,339 | – | 14,574 | 28,567 | – | 66,218 | 166,767 |
| Power Breakdown by Activity Type | | | | | | | | | | | | |
| Internal Power | 1 | 0.944 | 0.961 | 1 | 0.915 | 0.925 | 90.99 | 62.41 | 66.67 | 73.92 | 71.29 | 70.94 |
| Switching Power | 1 | 0.863 | 0.854 | 1 | 0.823 | 0.814 | 165.26 | 112.29 | 112.6 | 79.39 | 62.41 | 61.81 |
| Leakage Power | 1 | 0.739 | 0.826 | 1 | 0.743 | 0.752 | 0.19 | 0.12 | 0.133 | 0.298 | 0.207 | 0.199 |
| Power Breakdown by Cell Type | | | | | | | | | | | | |
| Sequential Power | 1 | 0.983 | 0.983 | 1 | 0.972 | 0.991 | 15.61 | 15.76 | 15.80 | 78.03 | 71.48 | 74.21 |
| Macro Power | 1 | 0.999 | 0.999 | 1 | 0.994 | 0.995 | | — | | | — | |
| Combinational Power | 1 | 0.832 | 0.848 | 1 | 0.810 | 0.806 | 236.50 | 155.40 | 159.90 | 54.71 | 52.44 | 49.16 |
| Clock Power | 1 | 0.974 | 0.933 | 1 | 0.916 | 0.885 | 4.30 | 3.68 | 3.62 | 10.87 | 9.99 | 9.57 |
| Memory Net Statistics | | | | | | | | | | | | |
| 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 Pow | 1 | 1.003 | 0.834 | 1 | 0.752 | 0.751 | | — | | | — | |
| Overall Power and Timing of the Designs | | | | | | | | | | | | |
| Total Power (mW) | 1 | 0.905 | 0.911 | 1 | 0.871 | 0.871 | 256.44 | 174.82 | 179.4 | 153.61 | 133.91 | 132.95 |
| Total Negative Slack (ns) | 1 | 10.786 | 0.999 | 1 | 7.919 | 0.832 | -20.23 | -141.21 | -32.13 | -2.145 | -783.74 | -1.221 |
| Avg. Negative Slack (ns) | 1 | 1.268 | 0.625 | 1 | 1.761 | 0.714 | -0.011 | -0.067 | -0.018 | -0.012 | -0.032 | -0.010 |
| Total Positive Slack (ns) | 1 | 0.587 | 1.512 | 1 | 0.508 | 1.347 | 662.0 | 562.8 | 672.2 | 5545.3 | 3955.3 | 6471.9 |
| Effective Timing and Energy Statistics | | | | | | | | | | | | |
| Effective Freq. (GHz) | 1 | 0.843 | 1.058 | 1 | 0.844 | 1.054 | 1.429 | 1.221 | 1.415 | 1.160 | 0.969 | 1.193 |
| Energy (pJ) | 1 | 1.074 | 0.861 | 1 | 1.031 | 0.827 | 179.50 | 143.18 | 126.8 | 132.41 | 138.19 | 111.44 |
| Energy × Delay (pJ ∗ ns) | 1 | 1.275 | 0.814 | 1 | 1.220 | 0.785 | 125.65 | 117.26 | 89.59 | 114.14 | 142.62 | 93.41 |
| Energy × Delay old [5] | 1 | 1.275 | 0.879 | 1 | 1.220 | 0.859 | 125.65 | 117.26 | 97.91 | 114.14 | 142.62 | 95.02 |

specifically targeted with ECO and hence it is not improved. We believe this is one of the first works showing applicability of ECO in monolithic 3-D designs. And by tailoring the ECO algorithm we can easily target various metrics of the design.

## VII. PPA Results and Analysis

The final GDSII layouts of the 2-D and Pin-3D implementations of the two processor designs are shown in Fig. 8, and the final PPA results of the 2-D, C2D, and Pin-3D based implementations of the four test-bench circuits are given in Table VII. While we have not focused on making the designs DRC clean, and routing is similar across 2-D and 3-D designs. The last row shows the EDP achieved using our preliminary flow [5], and we see that except for netcard (which is heavily cell dominant), we see 8%–10% reduction in EDP with the improvements to Pin-3D.

*a) Cortex-A7* As mentioned in Section V, the results for both Cortex-A7 and A53 are normalized w.r.t. the 2-D designs. Pin-3D shows better results than 2-D for all the key design metrics, such as cell count, wirelength, power, and timing slacks. The 3-D design shows ∼9% power benefit at the same target frequency of 2-D. Most of the power benefit in 3-D comes from the wirelength reduction (here, 25% smaller compared to 2-D) which leads to smaller wire load and therefore reduced switching power.

Switching power is a function of the sum of wire cap and input gate capacitance of the cells. So the 25% wirelength reduction translated to 15% switching power reduction in the design. Internal and leakage power are more dependent on the cell itself and are a function of cell area and cell types present in the final design. With reduced wire load, the cell strength and buffer strength can be reduced which explains the 4.4% drop in cell utilization and the smaller 1.4% drop in the cell count. This results in the 4% drop in internal power.

The reduction in leakage power drop is more drastic at a 16.4% reduction compared to 2-D. Leakage is significantly dependant on the threshold voltage type of the cell. The cell distribution based on the $V$th and the contribution of each cell type to leakage power is shown in Table VIII. In 2-D, we see that more than 50% of the cells are from the lowest-$V$th domain, and these contribute to around 80% of the total leakage power consumption. In Pin-3D the number of lowest-$V$th cells drop from 51% to 43% which is the main source of leakage power improvement.

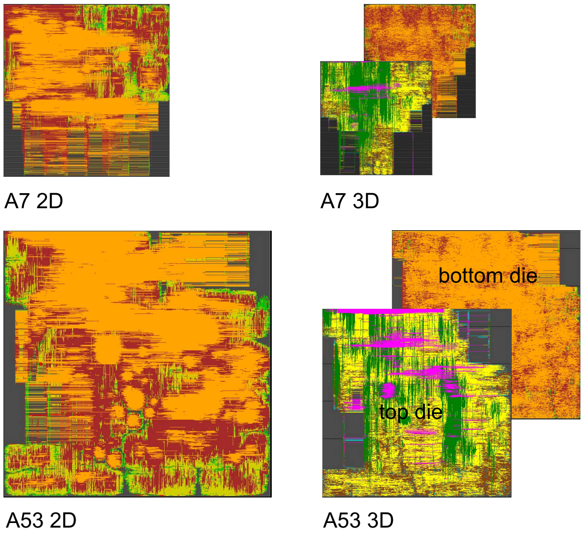


Fig. 8. GDS layouts of our Cortex-A7 and Cortex-A53 designs. For 3-D designs, we show the placement for the top die, and the routing for the bottom die. We use a TSMC 28-nm technology in all designs.

TABLE VIII
CELL $V$TH DISTRIBUTION IN CORTEX-A7 2-D AND PIN-3D DESIGNS. LOWEST–HIGHEST $V$TH ARE LABELED 1–4

| | 2D | | Pin-3D | |
|---|---|---|---|---|
| $V_{th}$ | % of Cells | % of Lkg. | % of Cells | % of Lkg. |
| Type1 | 51.4 | 82.7 | 43.4 | 78.2 |
| Type2 | 31.7 | 10.6 | 35.0 | 13.5 |
| Type3 | 9.0 | 0.7 | 10.1 | 1.0 |
| Type4 | 7.9 | 0.2 | 11.5 | 0.3 |

Instead of looking at the total power by internal, switching, and leakage, we can also split the power consumption based on the cell type. Sequential and macro power in 3-D have barely any improvements. For macros, the power is mostly dependent on the memory macro design which remain the same between 2-D and 3-D. Sequential power comprises the total internal and switching power of the sequential cells (flip-flops) that are connected to the clock network. As clock has a very high-activity factor compared to many other combinational logic, sequential power is dominated by the internal power which is not directly benefited from a 3-D placement. Moreover, as the sequential cells cannot be removed or added during the physical design stage, there is only limited optimization available during PnR.

The effective frequency (= [1/clock period − worst slack]) of the 3-D design is 5.8% higher than 2-D design due to the 3-D placement and the optimization methodology used in Pin-3D. Combined with the power savings from 3-D, the power delay product is 14% smaller than 2-D, and the energy delay product is 19% smaller.

*b) Cortex-A53* As both Cortex-A53 and Cortex-A7 are processor designs with L1 cache only and have a similar architecture and hierarchy, the benefits are similar between the two. The Cortex-A53 design is much larger than Cortex-A7 and we see a better reduction in overall power, cell count, and cell area in 3-D. With larger cell area reduction, the switching, internal, and leakage power reductions reach up to 18.6%, 7.5%, and 24.8%, respectively. Overall this contributes to the 12.9% total static power reduction. The total negative slack also improves compared to 2-D as it is 16.8% smaller. Effective frequency improvement is similar at around 5% and the overall EDP benefit compared to 2-D was 21.5%. The larger size and higher-combinational cell count of Cortex-A53 was useful to extract the higher power and timing savings. Wirelength reduction depends mostly from the routing complexity, wirelength distribution of the nets, and other metrics. It is not significantly dependent on the design size and so we see a reduction of 24.1% similar to Cortex-A7.

*c) LDPC and netcard* The two open-source RTLs are used to show the raw design metrics with all the three implementations. The time related metrics are reported in ns, and power metrics are in mW for the two designs. All other units are reported in the table.

## A. *Memory Net Analysis*

Memory macro design and its placement is one of the main timing bottlenecks of any physical design implementation. In cases with a high-memory workload, the nets connected to the memories become a crucial part of power savings. In such cases, 3-D implementation is especially beneficial to reduce the latency and wire load of memory nets. By placing cells on top of macros, rather than around the macros, the connections to memories are simplified. For example, we see long nets over macros in 2-D designs of Cortex-A53 and Cortex-A7 in Fig. 8. By placing macros on top of each other, we create easier access between macros and logic area creating a reduction in routing over large macro cells.

The impact is clearly seen in the memory net statistics of the processor circuits in Table VII. We see that due to the larger size and complexity of the Cortex-A53 design, the impact of 3-D placement is also higher. Here, the Root Mean Square input and output net latency reduce by more than 40%. Note that using Root Mean Square places higher weightage to larger latency values which are crucial in determining any timing bottlenecks. The net switching power of the memory macros are reduced by ∼25% in Cortex-A53 and 17% in Cortex-A7.

## B. *Timing Path Analysis*

The top-100 critical register-to-register paths are analyzed for a better understanding of the path-level trends. By analyzing a single path group allows for better-analysis and cleaner data instead of mixing than different paths, such as memory-to-register or register-to-output. For example, on memory paths a significant portion of the path delay is the internal macro delay that is only dependant on the macro design. On register-to-output paths, only a portion of the combinational logic is present in the design.

Table IX presents the averaged path statistics for the four circuits considered here. Since the Cortex-A benchmarks are processor designs, register-to-register paths are not usually the overall critical paths of the design. We see that in Cortex-A7 Pin-3D, the register-to-register paths have a worse-average negative slack than the 2-D design. This does not mean that the path is longer as we see that the path delay (from launch

TABLE IX
TOP 100 CRITICAL PATH AVERAGES OF REGISTER-TO-REGISTER PATH GROUP. THE CORTEX-A METRICS ARE NORMALIZED W.R.T. THE CLOCK PERIOD

| | Cortex-A7 | | | Cortex-A53 | | | 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 | -0.098 | -0.157 | -0.260 | -0.102 | -0.025 | -0.123 | -0.034 | -0.019 | -0.140 | -0.012 |
| Clock Skew | 0.050 | 0.025 | 0.175 | 0.135 | 0.117 | 0.076 | 0.005 | 0.055 | 0.021 | -0.026 | 0.017 | -0.015 |
| Setup Time | 0.007 | 0.024 | 0.011 | 0.004 | 0.010 | 0.006 | 0.019 | 0.017 | 0.017 | 0.023 | 0.100 | 0.033 |
| Path Delay | 0.984 | 1.236 | 0.912 | 1.108 | 1.133 | 1.020 | 0.668 | 0.717 | 0.663 | 0.821 | 0.824 | 0.794 |
| Cell Delay | 0.891 | 1.188 | 0.834 | 0.912 | 1.066 | 0.966 | 0.578 | 0.663 | 0.611 | 0.597 | 0.688 | 0.627 |
| Wire Delay | 0.093 | 0.048 | 0.078 | 0.105 | 0.067 | 0.054 | 0.090 | 0.054 | 0.052 | 0.224 | 0.136 | 0.167 |

to capture registers, inclusive) is smaller in Pin-3D compared to 2-D. But the path delay is higher in C2D due to the post-partitioning timing deviations as shown in Fig. 5. We note that clock skew on the Pin-3D paths is higher than the 2-D designs and is the cause for the worse slack in Pin-3D. The reason behind this is that, in Pin-3D implementations of Cortex-A7, the overall critical path is connected to a memory, and longer paths have a higher priority toward useful clock skewing to improve the timing slack of these paths. As a result, the clock skews for other noncritical paths become bad, resulting in the trends seen here. As the path criticality changes in Cortex-A53, we see that the Pin-3D clock skews are much smaller to not cause additional timing bottlenecks for the already long paths (path delay = 1.02 × clock period). In both these designs, we see that the wire delay contribution is much smaller in 3-D compared to 2-D.

For LDPC and Netcard, all the values are in ns, and the 100 paths considered are also the top-100 critical paths of the overall design. Clock tree is designed to help timing of critical paths, and we see that the average clock skews are similar within 15 ps of the 2-D skew. LDPC has a very dense path connections that are similar to each other in terms of logical complexity. This makes it difficult to make use of clock skew on critical paths without sacrificing timing of other paths. With more larger circuits like netcard, we do see the negative skew that helps the critical paths. Both of these circuits also have large wirelengths and so a higher portion of the path delay comes from wire delays.

### C. Routing Analysis and Metal Layer Savings

As observed in Fig. 8, the usage of top-most metal layer in the top-die (shown in pink) is very low in M3-D Cortex-A processor designs. Less than 2% of the total wirelength is routed on the top layer in 3-D, compared to 10% in the 2-D implementation. This is due to the difference in the routing stacks of 2-D and M3-D designs. In M3-D, wirelength is shorter, and this implementation method makes better use of other metal layers as discussed in Section IV-C. Due to the reduced usage on the top-most metal layers, routing in 3-D can effectively be completed with fewer metal layers in the design. This is seen in Table X, where removing a metal layer from the 3-D BEOL, only shows < 1% impact on total power and delay of 3-D Cortex-A7.

TABLE X
IMPACT ON PPA WITH ONE METAL LAYER REMOVED

| Cortex-A7 | Pin-3D | -1 Metal |
|---|---|---|
| Frequency | 1 | 1 |
| Total Metal Layer Count | 12 | 11 |
| Wire Length | 1 | 1.002 |
| MIV Count | 104,219 | 104,621 |
| Total Power | 1 | 1.001 |
| TNS | 1 | 0.958 |
| Eff. freq. | 1 | 0.999 |

### D. Heterogeneous 3-D IC Optimization

The Pin-3D optimization methodology provides a versatile and robust way to incorporate cells from different technology nodes into a single circuit at a path level. As discussed in Section V-B, for the first-time we design a 3-D IC with gate-level process node heterogeneity (15-nm top-die, 45-nm bottom die). Pin-3D flow requires an input pseudo-3D stage to proceed with incremental 3-D placement, clock tree optimization, routing, and timing optimization as shown in Fig. 2. Pseudo-3D flow only supports a single technology node and so we start with the 15-nm process node to synthesize and obtain the input design.

For a heterogeneous 3-D IC, the partitioning step assigns the cells to different technology nodes of top and bottom tiers. Partitioning should be able to create an area-balanced solution with no large clusters as discussed in [9]. But due to the heterogeneous nature of the design, this creates new constraints specific to heterogeneous 3-D ICs. Monolithic fabrication assumption that we use for the die fabrication means that the shape and size of the dies on the top and bottom tiers are the same. So when there is a significant skew in final cell areas across the two tiers, one of the tiers will be under-utilized and the other will be congested. Area balancing is important for efficient usage of the die area, and this is addressed by the use of a global scaling factor $\alpha$ (= average area ratio to 15-nm $\longrightarrow$ 45-nm cells) while partitioning the 15-nm design.

During the min-cut partitioning, when a move changes the cell tier from 15 nm to 45 nm, its area is scaled by a factor $\alpha$. This allows us to maintain area balance with heterogeneous cells. Usage of such a global scaling factor will not always be accurate since different cells have different area scaling factors

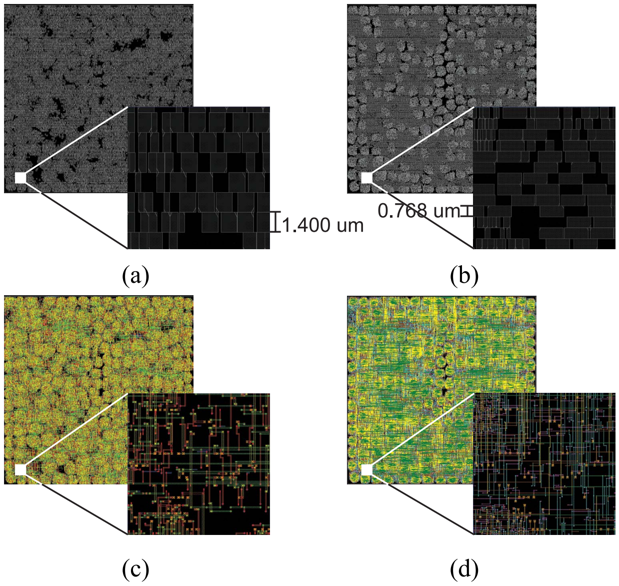


Fig. 9. Layout of our 45-nm+15-nm heterogeneous 3-D IC design of 128-bit AES benchmark using Pin-3D. (a) and (b) Full placement in bottom and top dies, respectively, along with standard row height. (c) and (d) Full routing of the bottom and top dies, respectively, with zoom-in windows for each.

TABLE XI
PPA Results of Our 45-nm+15-nm Heterogeneous 3-D IC Design of 128-Bit AES Benchmark Using Pin-3D. We Use 2 GHz as the Target Frequency of the Whole Design

| Design Metric | Total | Top-Die | Bottom-Die |
|---|---|---|---|
| Technology Node | Hybrid | 15 nm | 45 nm |
| Number of Cell Rows | – | 285 | 156 |
| Cell Area ($\mu m^2$) | 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 ($\mu m^2$) | 48,246 | | |
| Optimization Statistics | | | |
| 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 | | |

between technologies. For example, complex cells, such as flip-flops, would scale differently than a simple inverter of buffer and if there's a large concentration of flip-flops or other complex cells, the usage of single global scaling factor might not be very useful to reduce local area skew between the dies after partitioning. For a better partitioning, we would need a cell based scaling solution so that areas are accurately calculated between the dies.

To freely move the cells between the two different dies during partitioning, every cell in the 15-nm node should have a counterpart of the same logic type, driving strength, threshold voltage type, and pin list. These constraints allows cells from the 15-nm pseudo-3D design stage to be freely assigned to the 45-nm node during partitioning. There should also be a one-to-one mapping from the 15-nm ⟶ 45-nm nodes so that pins are not added or removed during partitioning. Considering such constraints for the cell and library design, we limit the usage of the cells to the basic logic gates, such as buffers, inverters, NAND, NOR, XOR, and flip-flops.

Starting from the partitioning stage, the core of the Pin-3D flow remains similar for heterogeneous 3-D IC. Fig. 9 shows the placement and routing layouts of the 3-D heterogeneous aes-128 design. Fig. 9(a) and (c) show the bottom 45-nm tier, where the zoomed-in placement show the tall 1.4 μm cell rows. Similarly Fig. 9(b) and (d) corresponding to the top-tier's smaller and shorter 0.768 -μm height cells in the 15-nm node. The routing in this tier also shows a much thinner and closely spaced wires. While there are more cells in the 15-nm node compared to the 45-nm node, the reduced pitch of the 15-nm node creates space for more routing tracks compared to the 45-nm die. Because of this, we do not see any additional congestion in either the 15-nm or the 45-nm dies.

Note that the one-shot routing introduced for homogeneous 3-D is still applicable for the heterogeneous design due to the way we define the 3-D stack. Referring to Fig. 1(b), in heterogeneous 3-D design, the routing tracks are defined such that the metal layers of each technology will use their corresponding technology rules. So the commercial routers follow the track definitions defined in the design allowing for a single-shot routing without causing additional issues. The MIV layer is the trickiest, since it has to interface between the two technologies. Here, we simply define the MIV to follow the minimum width and spacing rules of the 15-nm node since it has smaller min width and spacing rules between the top and bottom layer of the MIV. MIVs land only on the intersection between the tracks of top and bottom metal layers and so they do not impact the metal layer routing. To extend the analysis of heterogeneous 3-D, it is important to carefully consider the impact of MIV rules in routing of heterogeneous 3-D ICs.

The final PPA of the heterogeneous design are given in Table XI, and the values support the observations from Fig. 9. Cell Area is similar between the two tiers with the help of partitioning. The 15-nm die has ~70% of the total cell count and 68% of the routed wirelength due to the smaller cell areas of the FEOL and thinner routing pitch (more routing tracks) of BEOL. As clock tree is important for the overall timing control of the design, it is fixed on to the 15-nm die as this is the pseudo-3D input tech node. We see this in Table XI where there are no clock buffers and wires on the bottom die. Sequential cells also contribute to significant cell delays and are fixed to the faster top-die.

The combination of large cell count, and the power-hungry clock-network and sequential cells on the top-die shows a high-power consumption on this die. A significant benefit of this skewed power distribution is with thermal configuration. A heat-sink placed in contact with the top-die can absorb most of the thermal power. The power density is smaller on the bottom die and is further away from the heat sink. Thermal cooling has always been a major issue for 3-D ICs, and heterogeneous 3-D ICs create an in-built power skew than can be leveraged to tackle this.

As mentioned earlier, the heterogeneity shown here is at a gate-level, where each path has logic gates from different tech nodes. This is seen on the critical path, where 7 of the 18 logical cells on this paths are on the 15-nm die and contribute just 0.051 ns to the overall delay. The other 11 cells on the 45-nm die are particularly slow and contribute to most of the path delay. Before optimization, the paths are distributed randomly between the two tiers, and simply changing the cell node from 15 nm $\longrightarrow$ 45 nm without optimization for 31 000 of 107 000 total cells degrades the timing of the paths passing through the 45-nm die. This is seen in the pre optimization worst and total negative slacks of the design which are huge even in the small test-bench used. Optimization with Pin-3D has helped reclaim almost all the negative timing slacks giving a good timing closure to heterogeneous 3-D designs.

## VIII. Conclusion

In this article we proposed our Pin-3D methodology for incremental placement optimization, clock tree optimization, routing, timing optimization, and ECO for 3-D ICs using the commercially available PnR tools. Compared to the current state-of-the-art 3-D flows, we showed that adding our Pin-3D optimization improves every aspect of the design from placement, routing, and PPA. Especially, we see more than a 10× smaller total negative slack for the Cortex-A7 design and similarly high reductions in other benchmarks as well. Compared to 2-D designs, we were able to see ∼20% reduction in EDP for the Cortex-A series benchmarks, and ∼30% EDP improvement for the LDPC benchmark. 3-D routing with Pin-3D also allowed for savings in the BEOL cost without any meaningful effect to the important PPA metrics. We also saw how the buffering insertion with the Pin-3D methodology is superior in terms of critical path timing compared to a fairly limited die-by-die optimization. And finally, a proof-of-concept design of a heterogeneous 3-D IC shows the versatility of Pin-3D as well as exploring more complex structures that are possible with 3-D IC design.

## References

[1] "International roadmap for devices and systems," 2018. [Online]. Available: https://irds.ieee.org/images/files/pdf/2018/2018IRDS_ES.pdf

[2] 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 Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 36, no. 10, pp. 1716–1724, Oct. 2017.

[3] B. W. Ku, K. Chang, and S. K. Lim, "Compact-2D: A physical design methodology to build two-tier gate-level 3-D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 6, pp. 1151–1164, Jun. 2020.

[4] L. Bamberg, A. García-Ortiz, L. Zhu, S. Pentapati, D. E. Shim, and S. K. Lim, "Macro-3D: A physical design methodology for face-to-face-stacked heterogeneous 3-D ICs," in *Proc. DATE*, 2020, pp. 37–42.

[5] S. S. K. Pentapati, K. Chang, V. Gerousis, R. Sengupta, and S. K. Lim, "Pin-3D: A physical synthesis and post-layout optimization flow for heterogeneous monolithic 3-D ICs," in *Proc. IEEE/ACM Int. Conf. On Comput. Aided Design (ICCAD)*, 2020, pp. 1–9.

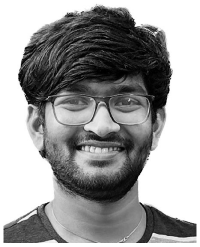
[6] J. Lu, H. Zhuang, I. Kang, P. Chen, and C.-K. Cheng, "EPlace-3D: Electrostatics based placement for 3D-ICs," in *Proc. 2016 Int. Symp. Phys. Des.*, 2016, pp. 11–18.

[7] M.-K. Hsu, V. Balabanov, and Y.-W. Chang, "TSV-aware analytical placement for 3-D IC designs based on a novel weighted-average wirelength model," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 4, pp. 497–509, Apr. 2013.

[8] B. Goplen and S. Sapatnekar, "Efficient thermal placement of standard cells in 3-D ICs using a force directed approach," in *Proc. Int. Conf. Comput.-Aided Design*, 2003, pp. 86–89.

[9] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Placement-driven partitioning for congestion mitigation in monolithic 3-D IC designs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 4, pp. 540–553, Apr. 2015.

[10] J. Balkind et al., "OpenPiton: An open source manycore research framework," in *Proc. 21st Int. Conf. Archit. Support Program. Lang. Operat. Syst.*, 2016, pp. 217–232.



**Sai Pentapati** received the B.Tech. degree from the Indian Institute of Technology Kharagpur, Kharagpur, India, in 2017, and the M.Sc. and Ph.D. degrees from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2020 and 2022, respectively.

His main research interests include physical design of 3-D ICs and design technology co-optimization.

**Kyungwook Chang** received the B.S. and M.S. degrees in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2007 and 2010, respectively, and the Ph.D. degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2019.

He was with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea, in 2021, where he serves as an Assistant Professor. His current research interests include computer-aided design solutions for 3-D ICs, design-technology co-optimization, physical/logic design and analysis, power delivery network, and parallel computing/memory architecture.

**Sung Kyu Lim** (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees from the Computer Science Department, University of California, Los Angeles, Los Angeles, CA, USA, in 1994, 1997, and 2000, respectively.

He is a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2001. He is the author of *Practical Problems in VLSI Physical Design Automation* (Springer, 2008) and *Design for High Performance, Low Power, and Reliable 3-D Integrated Circuits* (Springer, 2013). He has published more than 400 papers on 2.5-D and 3-D ICs. His research focus is on the architecture, design, test, and electronic design automation (EDA) solutions for 2.5-D and 3-D ICs. His research is featured as Research Highlight in the Communication of the ACM in January 2014.

Dr. Lim received the National Science Foundation Faculty Early Career Development Award in 2006. He received the ACM SIGDA Distinguished Service Award in 2008. He received the Best Paper Award from the IEEE Transactions on Electromagnetic Compatibility and the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems in 2022. He received the Best Paper Award from several conferences in EDA, including ACM Design Automation Conference in 2023. He joined the Defense Advanced Research Projects Agency in 2022 as a Program Manager in the Microsystems Technology Office.