Heterogeneous Monolithic 3-D IC Designs: Challenges, EDA Solutions, and Power, Performance, Cost Tradeoffs

Sai Pentapati and Sung-Kyu Lim, *Fellow, IEEE*

***Abstract*— Monolithic 3-D (M3D) integrated circuits (ICs) introduce a new aspect to the physical design problem by stacking dies vertically. We introduce a novel heterogeneous design for M3D ICs, which utilizes different technology processes for each die. We also propose enhancements to the design flow and improved partitioning methods to support our suggested heterogeneous 3-D IC design. We perform 3-D partitioning with low-cost, low-power, and low-performance cells on the bottom die, and vice versa on the top die. This arrangement allows us to create a high-performance 3-D IC that is more cost-effective and power-efficient than both the 3-D and 2-D homogeneous implementations. We demonstrate these advantages using four different netlists, showing up to a 23% improvement in performance per cost (PPC), and a 16% improvement in power delay product (PDP) when comparing our heterogeneous M3D design to the best 2-D designs.**

***Index Terms*— Heterogeneous integrated circuit (IC), monolithic 3-D IC, power performance area cost (PPAC) improvement.**

I. Introduction

As MOORE'S law scaling becomes increasingly harder to achieve, international roadmap for devices and systems (IRDS) proposed several alternatives that can provide additional benefits and bolster the impact of Moore's law. One such methodology is the 3-D packaging/fabrication. Generally, in 3-D packaging techniques, several predesigned chiplets/dies are attached together using microbumping or hybrid bonding to provide 3-D connectivity between the dies. While such packaging methodologies provide some benefits compared to a 2-D design, they are limited by the microbumping/hybrid-bonding pitch. Sequential fabrication of dies is a 3-D manufacturing technique that can provide a high 3-D bandwidth.

Several papers have tackled the problem of a 3-D integrated circuit (IC) design from various angles. The most common are: 3-D placers [1], [2], [3]. Placement is only part of the equation, and pseudo-3-D flows [4] aim to improve the place and route (PnR) flow in a holistic manner using the existing tools. This is achieved using several techniques such as better modeling of the 3-D ICs within 2-D EDA tools, enhanced 3-D optimization [5], improved design-aware partitioning solutions [6], or targeting 3-D structures that can be translated to 2-D space with minimal quality loss [7]. These flows rely on the improved placement or routing of the 3-D IC designs to derive the PPA benefits. The heterogeneous solution proposed in this work leverages process independence between the dies in 3-D IC to achieve better power performance area cost (PPAC). Compared to the heterogeneity in 2.5-D ICs and memory on logic designs, where the technology difference is limited to the block level of just the memory macros, we suggest highly integrated heterogeneous 3-D ICs and the benefits of such designs.

The dense connectivity possible with the monolithic integration supports heterogeneity at a gate level, with the capability of splitting a single path across different dies. However, doing so requires greater control of partitioning the cells across the two tiers as a simple min-cut partitioning can create unintended consequences if timing-critical cells are assigned to a slower tier. None of the current EDA tools support optimization across different technology nodes or even different tiers. So, partitioning needs to be aware of the process node differences. This is in contrast to Arka et al. [8] proposed a heterogeneous monolithic 3-D (M3D) with low cut size using architectural partitioning and analyzed the benefits of such implementation using a high-level simulator.

In this article, we use Pin-3-D [5] flow to implement our heterogeneous 3-D designs. Additional stages and functionality are added to Pin-3-D to support our multitechnology setup. The PPAC impact is studied using four different register transfer levels (RTLs): Advanced encryption standard (AES) (cell dominant), Netcard (large, wire dominant design), low density parity check (LDPC) encoder and decoder (wire dominant), and Cortex A7 CPU (general-purpose design). These are designed in five configurations (two types of 2-D designs, two 3-D designs, and the heterogeneous 3-D design) as in Fig. 1.

II. Technology Setup

Depending on the choice of technology nodes, a heterogeneous 3-D IC can be used to optimize different aspects of the 3-D IC. For example, Gomes et al. [9] design an SoC in heterogeneous fashion by using the low-power die for periphery logic blocks on the bottom die with 22 nm, and all the timing-critical blocks including the cores and the cache are in the 10 nm technology node on the top die. The dies are packaged together using a microbumping approach along with a low-cost boundary framework that reduces the impact of die-to-die communication on latency and power.

Manuscript received 3 June 2023; revised 11 September 2023 and 26 November 2023; accepted 2 December 2023. Date of publication 4 January 2024; date of current version 27 February 2024. This work was supported in part by the Ministry of Trade, Industry and 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 DOE Office of Science Research Program for Microelectronics Co-design (Abisko). *(Corresponding author: Sai Pentapati.)*

The authors are with the Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: sai.pentapati@gatech.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TVLSI.2023.3347372.

Digital Object Identifier 10.1109/TVLSI.2023.3347372

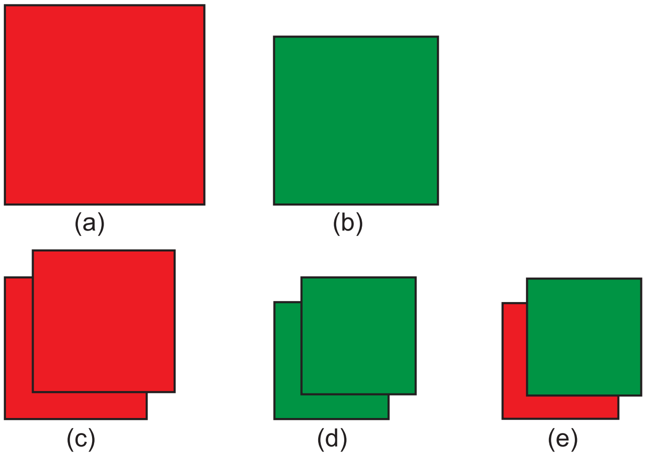

Fig. 1. Five different configurations of 2-D and 3-D using 9- and 12-track cells: (a) 12-track 2-D, (b) 9-track 2-D, (c) 12-track 3-D, (d) 12-track 3-D, and (e) 9 + 12-track heterogeneous 3-D.

Overall chip cost is an important metric for any chip, and Ku et al. [10] showed that the additional integration cost of 3-D ICs makes it more expensive. However, we show that by choosing suitable technologies for top and bottom dies, we can target cost reduction in 3-D without impacting the performance.

A. Cost Trends and Picking the Libraries

Cost per mm^2 of the wafer and the cost per transistor (or cost per element, in general) are two important metrics used to analyze the chip cost. These are related as

$$\text{Cost per Element} = \text{Wafer Cost per mm}^2 \times \text{Area of Element}.$$

The wafer cost per cm^2 trend from Intel for 1μm–65 nm is presented in [11]. The trend follows a saw-tooth-type pattern due to the following effects: as technology advances, the manufacturing complexity and cost initially increase. This creates the rise in cost per mm^2. Then, with the introduction of larger wafers and yield improvements, the cost per cm^2 decreases and creates the saw-tooth trend. As technology nodes advanced, the wafer sizes became stagnant at around 300 mm diameter leading to a sustained wafer cost per cm^2 increase starting from the 130 nm node.

As technology advances, we typically see a rise in wafer costs. However, this is usually offset by an increase in transistor density, which results in a lower cost per component in advanced process nodes. But the cost per component seemed to plateau with recent nodes and then began to increase from the 7 to 5 nm node (refer to [12, Table 9]). This indicates that while the cost of manufacturing a standard IC used to drop with each new generation due to the reduced cost per component, this trend is not as noticeable in recent nodes. In the 28-nm technology node that we currently utilize, we continue to operate under the assumption that newer nodes are less expensive than their predecessors.

In [9], power reduction with older technology nodes is achieved using specific process development kits (PDKs), such as low-power 22 nm and high-performance 10 nm nodes. Because newer technology nodes often come with high initial costs and limited capacity, moving some blocks to a different node can help keep the overall cost of the chip down. However, in this work, we assume that the technology nodes used are mature and have no associated setup costs. Therefore, our solution does not consider the manufacturing costs of newer nodes due to maturity.

TABLE I
"Qualitative" Comparisons of Expected Ideal PPAC Behavior of the Five Technology and Design Variations at Their Expected Maximum Frequencies. 1 Means the Worst, and 5 the Best

	9 Track (slow & small)		12 Track (fast & large)		9+12 Tracks (combined)	
	2D	3D	2D	3D	2D	3D
Frequency	1	2	3	5	-	4
Power	4	5	1	2	-	3
Power/Freq	3	4	1	2	-	5
Footprint	4	5	1	2	-	3
Si Area	5	5	1	1	-	3
Die Cost	5	4	2	1	-	3

Without the specific libraries optimized for power and performance, the improvements in the PPAC Metrics of heterogeneous 3-D ICs cannot be realized. So, to illustrate the benefits of heterogeneity, we use multitrack variants of a single technology node as substitutes for the different nodes. These multitrack cell libraries, which vary due to process and voltage changes, have different physical, electrical, and transistor properties on two dies. This makes them a suitable substitute for heterogeneous technologies. Another advantage of using multitrack cells is that they share the same back end of the line (BEOL) technology, simplifying the interface between the tiers.

As mentioned in [13], cells with fewer horizontal M1 tracks (shorter cell height) have a smaller area without additional wafer costs, as the transistor technology node remains similar. In our research, we demonstrate heterogeneity by choosing a 12-track (highest available) and a 9-track (smallest available) library from a foundry 28 nm node. The 12-track cells are faster, larger, consume more power and have a higher die-cost than the 9-track cells. The heterogeneous IC targets a mix of the two to create high-frequency, low-power, and low-cost designs. We have outlined the expected behavior of the five different technology and design configurations in Table I. In Section II-B, we will explore the impact of the heterogeneous cells in the full chip context.

B. Quirks of Heterogeneity

As discussed in the above sections, the technology processes on the two tiers need to be selected carefully to effectively use the heterogeneous integration. In later sections, we will see the effect of this behavior. Compared to heterogeneous integration of hybrid bonding or microbumping designs, the density of monolithic technology requires additional considerations as the partitioning splits timing paths across the two tiers with different voltage levels.

While level shifters are designed to handle the effects of heterogeneous voltages, we will see in Section III-B that the cells have significant drawbacks in monolithic heterogeneous designs. As such, the maximum voltage difference between the two libraries for a heterogeneous design is kept to a small value: $V_{\text{DDH}} - V_{\text{DDL}} < 0.3 \times V_{\text{DDH}}$ (where V_{DDH} is higher, and V_{DDL} is the lower voltage level of the two). If the difference between the two voltages ($V_{\text{DDH}} - V_{\text{DDL}}$) is greater

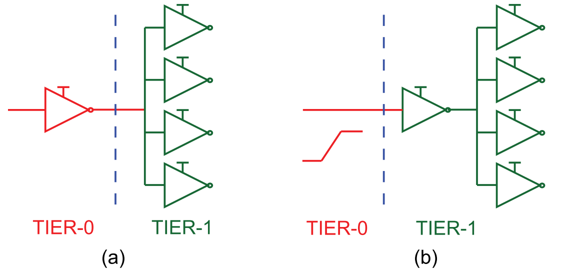

Fig. 2. Two types of boundary conditions due to heterogeneity in an FO-4 inverter. (a) Heterogeneity at the driver output. (b) Heterogeneity at the driver input.

TABLE II

IMPACT OF HETEROGENEOUS TECHNOLOGY WHEN THE DRIVER AND THE LOAD ARE IN DIFFERENT TIERS [SEE FIG. 2(A)]. TIME IS IN ns, POWER IS IN MICROWATT

	Case-I	Case-II	Δ%	Case-III	Case-IV	Δ%
Tier-0	fast	fast	–	slow	slow	–
Tier-1	fast	slow	–	slow	fast	–
Rise Slew	15.6	14.6	-6.7	14.6	16.6	14.2
Fall Slew	18.2	15.2	-16.9	19.1	20.7	8.1
Rise Del.	12.5	10.9	-13.1	23.6	25.2	6.4
Fall Del.	16.4	13.4	-18.1	26.2	32.0	22.3
Lkg. Pow.	0.093	0.092	-0.3	0.003	0.003	-1.3
Total Pow.	3.86	3.69	-4.3	2.00	2.18	9.0

than the threshold voltage, the signal high and signal low inputs will not be registered properly. Among the libraries considered, the lowest V_{thp} is $>0.3 \times V_{\text{DDH}}$ which in turn is $>V_{\text{DDH}} - V_{\text{DDL}}$ giving us some voltage margin between V_{thp} and $V_{\text{DDH}} - V_{\text{DDL}}$ and ensuring proper functionality without the need for voltage shifters.

However, correct functionality is not sufficient to ensure high performance. The timing of these cells on the boundary of tiers should not be degraded. To study the various effects of the varying voltage levels, an FO-4 inverter is studied with the two partitioning boundary configurations as shown in Fig. 2. In case 1, the load cells on Tier-1 receive a voltage level corresponding to Tier-0. And in case 2, the driver on Tier-1 receives a different voltage level than its characterized value.

Such cells that have inputs and outputs on different tiers are referred to as “boundary cells.” These cells need to be carefully considered to make sure the heterogeneous technologies can work well together. This means, in Case 1 (*heterogeneity at driver output*) since the driver output slew depends on cell load from a different technology, the output slew can be quite different than intended for the technology. Similarly, this slew would then be the input slew for the load cells and can be quite different from the expected input slew in this technology. For accurate timing within the EDA tools, this slew should be within the characterized input slew range of the load cells. To avoid such inaccuracies, only libraries with significant overlap in characterized slew ranges are considered for heterogeneous integration. As reported in Table II, SPICE analysis shows that when the different tier from the load, the slew changes only by at most ±15%. Slew characterization in libraries is done over a wide range spanning two–three orders of magnitude. So, small variations of ±15% are easily captured by the tool.

The next arrangement of boundary cells is called *heterogeneity at input*. Here, the driver and the load are physically located in the same tier, but the input signal to the driver comes from a different tier. The impact of such heterogeneity is reported in Table III. It is important to note that the delays here are just stage delays, and the opposite signs imply that overall path delay would not differ significantly for a given register-to-register timing path. This is because even if a path goes through the top and bottom dies multiple times, the delay inaccuracy from going to tier 1 → 2 and the inaccuracy from going to tier 2 → 1 would be in opposing directions. Even on a relatively short path of ten logic cells deep, having a timing inaccuracy of 5% at one stage will not have a significant impact on the overall path timing.

TABLE III

IMPACT OF HETEROGENEOUS TECHNOLOGY WHEN THE INPUT TO THE DRIVER OF AN FO4 IS FROM DIFFERENT TIER [SEE FIG. 2(B)]. TIME IS IN ns, POWER IS IN MICROWATT

	Case-I	Case-II	Δ%	Case-I	Case-II	Δ%
Tier-0	fast	slow	–	slow	fast	–
Tier-1	fast	fast	–	slow	slow	–
Driver V_G	0.9V	0.81V	10.0	0.81V	0.9V	11.1
Rise Slew	15.6	16.9	8.1	14.6	13.1	-9.9
Fall Slew	18.2	19.4	6.6	19.1	17.6	-8.1
Rise Del.	12.5	13.0	3.4	23.6	22.4	-5.3
Fall Del.	16.4	17.1	4.1	26.2	24.8	-5.1
Lkg. Pow.	0.093	0.330	250	0.003	0.002	-44.9
Total Pow.	3.86	4.21	9.2	2.00	1.99	-0.6
% Input Pins	31	15	–	33	21	–

While the timing discrepancy is relatively small, we see that the discrepancy in leakage power can be quite large and asymmetric between the fast → slow and slow → fast boundary types. While the deltas here again look pretty large, the leakage power is only a small portion of the total power, and this discrepancy does not impact the overall power significantly. One of the reasons for the extreme discrepancy is the SPICE setup used. We are assuming ideal wires in the SPICE netlists of the FO4 inverter. Since the PDK used is from foundry technology, we do not have access to the actual cell SPICE netlists of the inverters, and only the encrypted wire models were available for physical design. If we included wire models, the base case would show much higher leakage due to the exponential relationship between leakage power and node voltages. This would also reduce voltage variation sensitivity. If we had a well-defined cell model, this difference would be much smaller than what we currently observe.

C. Cost Model

To accurately assess the cost and benefits of 3-D integration, we need a well-defined cost model. In this study, we adapt the model from a previous study to better suit our needs. This model is outlined in Table IV. We consider a die with one front end of the line (FEOL) layer and eight BEOL layers as the standard. We assume that FEOL contributes 30% of the cost and that the metals in BEOL have a consistent cost per layer. We add a 5% wafer cost as a penalty for 3-D integration. All these assumptions are based on the valid ranges and assumptions from [10]. The 2-D wafer cost includes the cost of the FEOL layer and six metal layers, while the 3-D wafer cost includes two FEOL layers, two tiers of six metal layers each, and the added 3-D integration cost. We derive the die cost from the wafer cost using (1)–(5) in Table IV. We also include an additional penalty for 3-D integration in the “3-D Yield Degradation.”

TABLE IV

COST MODEL ASSUMPTIONS [10]

Baseline wafer cost (FEOL+8 metals)	C'
Wafer FEOL cost	$0.3 \times C'$
Wafer BEOL cost (up-to 6 metals)	$0.66 \times C'$
3D integration cost (α)	$0.05 \times C'$
Wafer Diameter	300 mm
Defect Density (D_w)	$0.2\,\text{mm}^{-2}$
Wafer yield (κ)	0.95
3D Yield Degradation (β)	0.95
2D Wafer Cost (C_{2D})	$0.96 \times C'$
3D Wafer Cost (C_{3D})	$1.97 \times C'$

$$\#\text{ Dies } (DPW) = \frac{\text{Wafer Area } (= A_w)}{\text{Die Area } (= A_d)} - \sqrt{2\pi \frac{A_w}{A_d}} \quad (1)$$

$$\text{2D Die yield } (Y_{2D}) = \kappa \times (1 + \frac{A_d D_w}{2})^{-2} \quad (2)$$

$$\text{3D Die yield } (Y_{3D}) = \kappa \times \beta \times (1 + \frac{A_d D_w}{2})^{-2} \quad (3)$$

$$\#\text{ Good Dies}(N_{GD}) = DPW \times Y_{[23]D} \quad (4)$$

$$\text{Die Cost} = \frac{C_{[23]D}}{N_{GD} \times Y_{[23]D}} \quad (5)$$

III. HETEROGENEOUS 3-D IC DESIGN FLOW

A. Enhancements to the Pin-3-D Flow

The current state-of-the-art electronic design automation (EDA) only supports the PnR of 2-D ICs. So, here we use Pin-3-D [5] flow which has been proposed as a way to optimize M3D IC designs. Pin-3-D can optimize the 3-D dies with the help of overall chip timing which is extremely important in a heterogeneous context. But, as mentioned in the article [5], this flow does not have a 3-D clock tree stage.

In heterogeneous 3-D IC design, it is crucial to reoptimize the clock tree at the 3-D stage considering the process node differences. To rectify some of the shortcomings of Pin-3-D flow and to make it more suitable for heterogeneous 3-D ICs, we enhance the Pin-3-D flow in a few major areas. In this work, we introduce timing-based partitioning, 3-D clock tree design, and the repartitioning of tiers to achieve timing closure. The result of the optimizations is shown in Table V.

1) Timing-Based Partitioning: Pin-3-D flow and other existing 3-D flows such as Compact-2-D and Shrunk-2-D can be separated into two stages: The pseudo-3-D stage and the 3-D stage.

During the pseudo-3-D stage, the entire design is implemented in a 2-D fashion, and so it is limited to a single technology process and a die. Once this is optimized, the cells are partitioned into two tiers. The partitioning of the netlist into two tiers is based on a placement-driven FM min-cut algorithm with area balancing.

In heterogeneous 3-D, partitioning also needs to consider the difference in timing between the dies. This is because the cells get assigned to the slower die will result in degraded delay and have a negative impact on overall chip PPA. If additional cell sizing and buffer insertion had to be performed by the tool to overcome this PPA degradation, power and area would be negatively impacted. In cases where cell sizing cannot overcome the timing degradation due to the cells placed on a slower die, we would not be able to achieve the target frequency for a 3-D IC. Therefore, a timing-based partitioning is introduced here.

TABLE V

IMPROVEMENTS OBTAINED WITH OUR HETEROGENEOUS VERSION OF PIN-3-D FLOW [5]

		Pin-3D [5]	Hetero-Pin-3D
Frequency	GHz	1.200	1.200
WL	mm	3.22	3.23
WNS	ns	-0.489	-0.060
Total Power	mW	224.1	198.8

A similar problem of timing-based partitioning has been tackled in [14] to address the process variation and/or degradation between the two tiers of a 3-D IC. Samal et al. [14] used path-based timing analysis to find critical cells and assign them to the faster die. However, this approach does not achieve a large enough cell coverage to work in our case. In this work, we use a cell-based approach to achieve complete coverage. This is used because missing even a small fraction of critical cells can lead to a large timing degradation. So, instead of path-based slack measurement, we visit the cells individually and find the worst slack among the paths going through the cell. Each cell is assigned timing criticality based on this number. This is then used for finding and fixing these critical cells on the faster die to avoid timing issues.

In general, the timing-critical cells in an ASIC usually come from a specific RTL block, like a multiplier block, a float-processing unit, and so on. As such, these cells would be placed physically close to each other to avoid long routing within critical blocks. The drawback here is that this can lead the timing-based partitioning to assign dense physical clusters onto a single die. This creates overlaps between cells after partitioning, which need to be legalized in 3-D. This additional legalization creates a mismatch between placement in the pseudo-3-D stage and the final 3-D stage. To avoid this mismatch and inaccuracies, we limit the timing-based partitioning to only 20%–30% of the total cell area. To keep cut size at a manageable limit, we still perform bin-based FM min-cut partitioning on the remaining cells in the design.

2) Supporting Heterogeneous Clock Tree: The Pin-3-D flow on which we base our design methodology does not support a 3-D clock tree optimization. At any point in the 3-D stage of the Pin-3-D flow, only the cells from a single die are treated as traditional standard cells while leaving others to be treated as transparent cells that do not occupy silicon area. This was done to remove overlap between the cells from different dies. However, this also makes the tool treat these cells as macros, breaking up the clock tree and unable to optimize the clock tree under the 3-D setting.

We use a different approach here by representing the cells of other die(s) as "COVER" cells in the design. Traditionally, such cells are used to define things such as bump pads which do not have an active area but are instantiated as cells in the design. Using this, when one die is being optimized in Pin-3-D, all other cells are treated as "COVER" cells. This approach allows for a simpler clock tree representation that allows for complete 3-D clock optimization with effective timing and power.

B. Drawbacks of Level Shifters

As discussed in Section II-B, the difference in voltage levels between the dies in a heterogeneous design creates a problem for the signals crossing the boundary dies. While this can be remedied by adding voltage shifters at the boundary crossings, the dense interconnectivity of a monolithic design makes it costly in terms of the overall chip PPA. In the M3D implementation of the processor design, on average, we observe ≈15% of the nets are connecting the two tiers, and so would require level shifters if the voltage difference is too disparate. And since the nets break up at path level, this addition of voltage shifters would increase the timing delays on a large number of paths degrading the timing. Without the use of high interconnection density that blocks the use of voltage shifters, we would not be able to implement a dense heterogeneous design which is the main motivation of this work. So, by limiting the voltage difference between the two dies to 10% and limiting subsequent timing and power discrepancies, we allow for a heterogeneous design without needing level shifters.

C. Repartitioning Using ECO

The last portion of the heterogeneous flow update is the introduction of repartitioning for 3-D ICs. In homogeneous 3-D flows, tier-partitioning remains unchanged after the initial min-cut step. However, in the case of heterogeneous 3-D ICs, the partitioning needs to adapt based on the accurate timing difference between the dies. This is more pronounced as the pseudo-3-D is only based on a single technology node and cannot be optimized for the heterogeneous design. While timing-based partitioning helps to properly assign the cells, the timing data is obtained at a very initial stage and cannot be fully accurate. So, we propose a repartitioning technique to identify cells that are too slow/fast for its tier, and repartition to its proper tier. The repartitioning algorithm is presented in 1.

Algorithm 1 Repartitioning Algorithm

```
$unbalance \leftarrow delta_{area}/total_{area}$
$d_k \leftarrow d_0$          ▷ used for delay threshold for critical cells
$n_p \leftarrow n_0$          ▷ number of paths considered per loop
while $unbalance > unbalance_{th}$ do
    $d_{th} \leftarrow d_k \times$ (avg. cell delay of $n_p$ critical paths)
    for all $cell \in$ the $n_p$ critical paths do
        $d_c \leftarrow$ delay of $cell$ on its critical path
        if $d_c > d_{th}$ then
            $all\_crit += 1$
            if $cell \in$ slow die then
                $slow\_crit += 1$
                append $cell$ to $move\_list$
            end if
        end if
    end for
    if $slow\_crit/all\_crit < crit_{th}$ then
        break                              ▷ Stop re-partitioning
    end if
    Move all cells in $move\_list$ to the fast die with ECO
    update timing
    if $\text{Delta}_{\text{WNS}} < W_{th}$ || $\text{Delta}_{\text{TNS}} < T_{th}$ then
        Undo the cell moves done
        $d_k * = \alpha$                   ▷ $\alpha < 1$
    end if
    update $unbalance$
end while
```

IV. Experimental Results

A. Setup and Methodology

1) Setup: To evaluate the heterogeneous 3-D IC, we use a commercial 28 nm library with a 9-track cell version at 0.81 V on the top tier and the 12-track cells operating at 0.90 V on the bottom tier. There are six signal routing layers per tier and are the same layers used for signal routing in the 2-D design. The physical and electrical properties of the routing layers on each of the two tiers are the same as those of the first six routing layers of the 2-D BEOL.

2) Methodology: The overall flow for the heterogeneous 3-D IC design is discussed in Section III. Here, we discuss the methodology and criteria used for comparing the various design implementations. To start, we first have five different options in which each netlist is implemented: 9-track implementations (2-D and 3-D), 12-track implementations (2-D and 3-D), and heterogeneous (3-D). For the 9- and 12-track implementations, the netlists are synthesized in the respective technology nodes, for better PPA. Based on this netlist, the floorplan area is fixed using target cell utilization. The faster 12-track 2-D implementations are swept across a range of frequencies to find the maximum achievable target. The consideration for timing failure here is a worst negative slack of ≈5%–7% of the clock period. Allowing for a small negative slack shows that the achieved frequency is the max possible. When the timing target is not set tightly, the tool starts optimizing for power sacrificing any positive slack and makes it harder to compare the maximum achievable frequency. For each netlist, the 12-track 2-D maximum frequency is used as the target for other implementations of the same design.

While the synthesis stage is relatively straightforward for homogeneous 2-D and 3-D designs, it is not so for the heterogeneous flow. For heterogeneous 3-D, the "pseudo-3-D stage" only supports a single technology node. So, the synthesis, area estimations, and optimization for pseudo-3-D are based on the 12-track technology. After the pseudo-3-D stage, the timing-based partitioning followed by FM min-cut creates our 3-D design. The partitioning distributes the cells across the two tiers: 12- and 9-track. By mapping some of the 12-track cells in pseudo-3-D to the 9-track tier, the cell area decreases due to the smaller cell size of the 9-track version. With half the cell area now based on the 25% smaller 9-track cells, the overall area reduces by 12.5%. The footprint is reduced accordingly to maintain the chip utilization at this stage. Once the 3-D database is set up, we use the proposed flow in Section III to create the heterogeneous 3-D GDS.

B. Full-Chip PPAC

1) Heterogeneous 3-D IC Results: Based on the RTL-to-GDS methodology discussed, we provide the power, performance, area, and cost (PPAC) results of AES, LDPC, Netcard, and a commercial CPU designed in heterogeneous 3-D with the 9 + 12-track cell combination in Table VI. To simplify

TABLE VI

PPAC Results of Our 3-D Heterogeneous Designs (Raw Data Based on a Commercial Foundry 28 nm Technology)

	Units	netcard	aes	ldpc	cpu
Frequency	GHz	1.750	3.000	1.125	1.200
Area	mm^2	0.384	0.126	0.216	0.390
Chip Width	µm	438	251	329	442
Density	%	82	86	64	88
WL	m	6.560	1.022	5.500	3.073
# MIVs	×1000	153	62	83	98
Total Power	mW	550	138	339	188
WNS	ns	-0.037	-0.028	-0.026	-0.055
TNS	ns	-0.41	-4.827	-0.902	-15.54
Effective Delay	ns	0.608	0.361	0.597	0.888
PDP	pJ	334.5	49.8	310.1	167
Die Cost	$10^{-6}C'$	6.16	1.97	3.41	6.26
PPC	$\frac{\mathrm{GHz}}{\mathrm{mW}\times 10^{-6}C'}$	0.517	11.06	0.946	1.02

the comparisons, the absolute values are only provided for the heterogeneous designs. For all the homogeneous implementations, only the percent delta is reported in Table VII, except for a few key metrics. Going through the metrics in Table VI, here, the area corresponds to the total silicon area (which is the same in 2-D and its homogeneous 3-D counterpart in our methodology). Power analysis is done using fixed input activity factors, and statistical switching propagation in Innovus. Sign-off timing is performed using the in-design Tempus timing engine of Innovus. For the power delay product (PDP) calculation, effective delay (=clock period–worst slack) is used to include the small deviations in the timing slack. Cost per cm^2 is defined as (Die Cost/Total Si Area) and gives an estimate of the cost of heterogeneous 3-D IC in general. Finally, the performance per cost (PPC) is calculated based on achieved frequency, power consumption at this frequency, and die cost. Intuitively, it shows the achievable performance per unit of power and cost.

The cell area of the four netlists shows that Netcard and CPU designs are the largest netlists considered. Netcard is a simple logic RTL with 250k cells, and the CPU has 150k cells along with a large area dedicated to the cache that contributes to 40% of the footprint. The overall density of the two tiers is reported for 3-D designs, and the low utilization of LDPC is notable as it is an extremely wire-dominant circuit. The routing is extremely congested with a large number of global connections, so a tighter integration would lead to a worse PPA for LDPC. In all the designs presented in Table VI, we see that the worst slack is slightly negative showing that the optimization is almost at the limit and the timing met condition (worst negative slack $<\approx 7\%$ of clock period) is satisfied. The frequency target used here is the maximum achieved frequency of the 12-track 2-D counterparts as discussed in Section IV-A2 and is used for future iso-performance comparisons in Table VII.

In the following sections, we provide the percent benefit of 3-D heterogeneous designs with respect to different homogeneous combinations. The placement, routing, and zoomed-in layouts of the CPU design under the two 2-D implementations and heterogeneous 3-D is shown in Fig. 3. From Fig. 3(c), we can see the different cell heights across the two dies. The window size and magnification are kept constant across the two tiers in 3-D for a good visual comparison of the cell heights.

2) Heterogeneous 3-D Versus 9-Track 2-D: The first four data columns in Table VII show the percent deltas of the 9-track 2-D implementations compared with heterogeneous 3-D. Except for PPC, a −ve shows that the heterogeneous 3-D implementation improves on the specified metric. Cost per cm^2 shows that heterogeneous 3-D is more expensive per unit area due to the added 3-D integration cost and yield degradation. The cost per cm^2 differs based on the block since the yield is dependent on the die area. This is why AES and LDPC have similar costs per cm^2, and Netcard and CPU are similar.

Overall, 9-track 2-D designs are the worst performers compared to heterogeneous 3-D. The densities are more or less similar to the heterogeneous 3-D IC for three out of the four netlists considered. LDPC is a very wire-dominated design, and the routing feasibility drives the optimization. For designs other than LDPC, we see that the total area of the chip is larger than heterogeneous implementation. While we would expect the smaller cells of the 9-track design to have a smaller cell area, the high target operating frequency creates an over-correction in the synthesis stage by adding more buffers and larger cell sizes to account for the slower cells in the 9-track version. This leads to the worse chip area as the target utilization is kept constant across all implementations. Due to the larger area and more cells, the 9-track 2-D design is comparatively worse across most of the metrics. We note across the 9-track designs, the simpler AES and LDPC designs show a good overall slack (−0.03 ns in both cases), but as the design complexity increases the timing deviates further from the target.

3) Heterogeneous 3-D Versus 12-Track 2-D: Next, we compare the 12-track 2-D designs (which were our baseline for timing comparisons) with the heterogeneous 3-D. Note that similar to 9-track 2-D, the cost per cm^2 is smaller for 12-track 2-D due to the lack of additional costs specific to 3-D. We first see that the 2-D designs require a larger area to meet the target utilization. Compared to LDPC, which is a wire-dominant design, AES is at the opposite extreme as it is extremely cell-dominant. So, the faster cells in AES allow for better optimization during PnR, resulting in a 15% lower density in the 2-D implementation. Additionally, AES is not well suited to our heterogeneous methodology, as its design is small/simpler. Timing paths in AES are more similar across the entire design which makes it harder to optimize with heterogeneous 3-D IC. The AES design used here is a 128-bit encryption circuit, and all the 128-bits have a very similar functional path, making the design very symmetric. Because of these reasons, the wirelength reduction with heterogeneous 3-D is the least for AES at 17% compared to 25%–33% for the other three netlists. The effective delay is also smaller with heterogeneous 3-D showing that we can meet better timing targets than the 12-track 3-D designs. Once again, AES proves as an exception as the effective delay is slightly larger with heterogeneous 3-D. Except for AES, there is a significant power reduction across all three designs due to the strategic use of the slower 9-track cells for noncritical

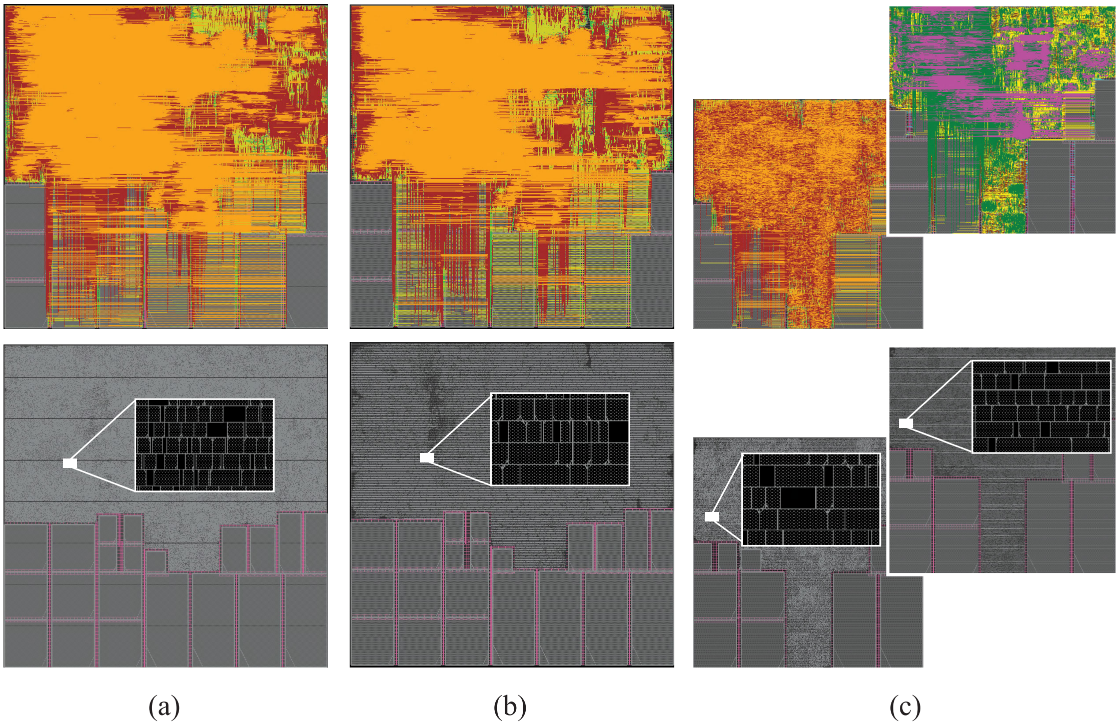

Fig. 3. Routing, placement, and zoomed-in layouts. (a) Two-dimensional 9-track processor, (b) 2-D 12-track processor, and (c) 3-D heterogeneous processor.

TABLE VII

PPAC PERCENTAGE DELTA (=(3-D HETERO - CONFIG)/CONFIG × 100) OF 3-D HETEROGENEOUS DESIGN WITH RESPECT TO DIFFERENT HOMOGENEOUS CONFIGURATIONS. A −VE (+VE FOR PPC) VALUE IMPLIES THAT HETEROGENEOUS IMPLEMENTATION OUTPERFORMS THE PARTICULAR CONFIGURATION

CONFIG →	2D 9-Track				2D 12-Track				M3D 9-Track				M3D 12-Track			
	netcard	aes	ldpc	cpu	netcard	aes	ldpc	cpu	netcard	aes	ldpc	cpu	netcard	aes	ldpc	cpu
Si Area	-28.6	-27.6	-13.9	-23.1	-12.3	-13.7	-13.9	-7.8	-28.6	-27.6	-13.6	-23.2	-12.3	-14.9	-13.6	-8.0
Density	3.2	2.9	-12.4	4.1	-1.8	14.5	2.2	5.4	8.7	6.9	-10.0	4.1	4.2	14.8	27.9	5.4
WL	-33.0	-30.9	-14.4	-34.8	-25.3	-17.0	-31.1	-33.5	7.8	-10.4	18.6	-8.9	-2.2	0.0	-8.7	-2.1
Total Power	-12.7	-15.0	-8.1	-21.1	-10.5	-8.0	-29.7	-14.6	-4.2	-8.6	5.4	-15.9	-6.0	-3.4	-5.6	-10.1
Eff. Delay	-14.7	-1.6	-0.8	-12.9	1.0	4.0	-2.6	-6.2	-19.3	0.0	1.2	-12.9	6.1	6.2	2.2	4.2
PDP	-25.6	-16.3	-8.8	-31.2	-9.6	-4.2	-31.5	-9.3	-22.6	-8.5	6.6	-26.7	-0.2	2.7	-3.5	-6.3
Die Cost	-27.9	-23.3	-9.5	-21.8	-9.7	-8.4	-9.5	-4.7	-29.7	-27.8	-13.9	-24.1	-12.7	-15.1	-13.9	-8.3
Cost per cm^2	0.98	5.9	5.1	1.7	3.0	6.1	5.1	3.4	-1.5	-0.3	-0.3	-1.2	-0.5	-0.2	-0.3	-0.3
PPC	59.1	53.5	20.3	61.9	23.7	18.6	57.2	23.0	48.6	51.6	10.2	56.7	21.9	21.9	23.0	21.4
Width (μm)	733	417	501	712	662	382	501	650	518	295	353	504	468	272	353	460
WNS (ns)	-0.14	-0.03	-0.03	-0.19	-0.03	-0.01	-0.05	-0.01	-0.18	-0.03	-0.02	-0.19	-0.00	-0.01	-0.01	-0.02
TNS (ns)	-450	-3.61	-5.13	-357	-2.14	-0.44	-7.60	-0.03	-832	-1.95	-0.04	-316	-0.00	-0.03	-0.01	-0.59

cells. This is significant as it shows the capability of both heterogeneous design along with the timing-based partitioning and repartitioning flows. An important note here is that the memories in the CPU design are of the same size in both technology variants.

4) Heterogeneous 3-D Versus 9-Track 3-D: The 3-D 9-track implementations suffer from the same complications as the 2-D 9-track designs due to the slow cells. Only for LDPC does the 3-D 9-track design compare to the heterogeneous implementation in terms of timing and power. This is because of the smaller wirelength that can be achieved in this design with 3-D implementation. This leads to better wirelength in the design which significantly impacts the overall QoR in this circuit. Even with this improvement, the overall PPAC is still better with heterogeneous designs by a noticeable margin.

5) Heterogeneous 3-D Versus 12-Track 3-D: Finally, we analyze the 12-track homogeneous 3-D implementations. These are expected to have the best timing and power across the homogeneous implementations as they have the faster 12-track cells and also the placement and routing benefits of 3-D design. This is clearly seen as it outperforms heterogeneous 3-D in terms of delay in all circuits. But this comes at a cost to power consumption, as the slower 9-track cells of heterogeneous 3-D consume less power than the 12-track 3-D designs. Together, the PDP comes out to be similar across the two implementation types with a slight advantage to the heterogeneous 3-D. However, the added die cost in 12-track 3-D further improves the PPAC in heterogeneous 3-D to outperform 12-track 3-D by ≈20%. Cost per cm^2 between all 3-D options are only within 1% of each other and differ due to the yield differences from the die area.

C. In-Depth Analysis of Clock, Critical Path, and Memory Connections

From the above comparisons, we see that the overall PPAC is significantly better with heterogeneous 3-D ICs compared

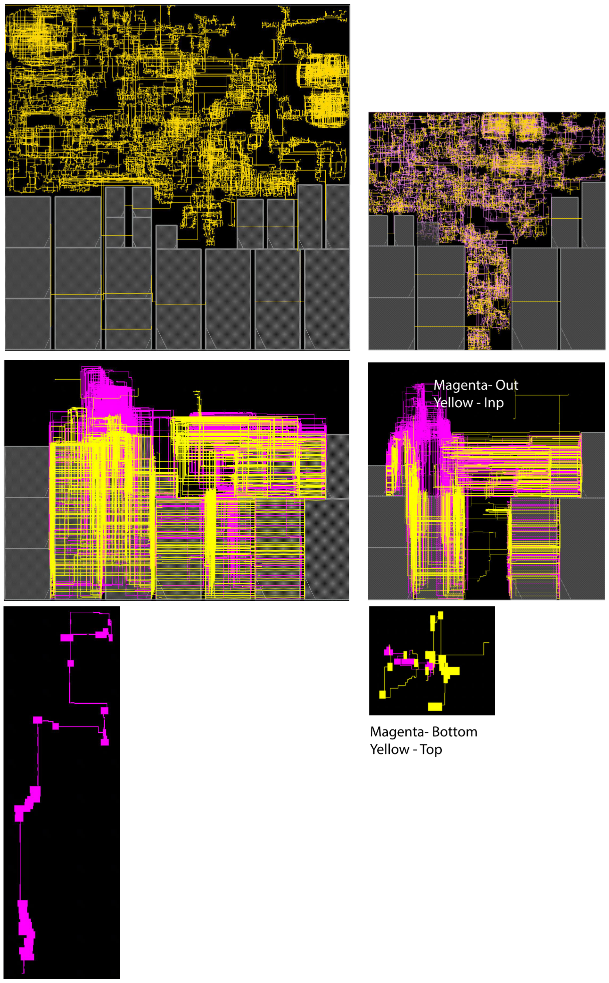

Fig. 4. Detailed layouts of 2-D and heterogeneous-3-D designs. (a) Clock tree. (b) Memory nets. (c) Critical path.

TABLE VIII
CLOCK NETWORK, CRITICAL PATH, AND MEMORY INTERCONNECT ANALYSES

	Units	9-track 2D	12-track 3D	Hetero 3D
Memory Interconnects				
Input Net Latency	ps	25.0	16.1	15.5
Output Net Latency	ps	37.5	33.2	28.7
Net Switching Power	ps	5.47	4.02	3.41
Clock Network				
Buffer Count		1593	1502	1330
Top Buffer Count		–	935	1045
Bottom Buffer Count		–	563	285
Buffer Area	μm^2	1277	1175	982
Wirelength	mm	0.114	0.108	0.107
Max Latency	ns	0.234	0.292	0.713
Max Skew	ns	0.058	0.142	0.344
100 Path Avg. Skew	ns	-0.008	0.000	-0.011
Critical Path				
Clock Period	ns	0.833	0.833	0.833
Slack	ns	-0.003	-0.012	-0.055
Clock Skew	ns	-0.014	0.013	0.045
Setup Time	ns	0.004	0.008	0.001
Path Delay	ns	0.845	0.831	0.845
Wire Delay	ns	0.014	0.009	0.021
Wirelength	µm	273.0	156.9	231.0
Top Wirelength	µm	–	79.2	59.6
Bottom Wirelength	µm	–	77.9	171.4
Cell Delay	ns	0.830	0.822	0.824
Total Cells		45	43	33
# MIVs		–	9	6
Top Cells		–	23	8
Top Cell Delay	ns	–	0.447	0.343
Avg. Top Delay	ns	–	0.019	0.043
Bottom Cells		–	20	25
Bottom Cell Delay	ns	–	0.375	0.482
Avg. Bottom Delay	ns	–	0.019	0.019

to the best 2-D or homogeneous 3-D counterparts. This is especially true for complex IPs which have various functional blocks where the timing criticality of cells is more diverse than simple designs such as LDPC or AES. In LDPC, there is a high degree of interconnectivity and the timing paths span the entire chip. The AES circuit is a 128-bit encryption system, and the functional path for each bit is very similar to each other. Due to this, the timing paths are closely matched between the different bits and do not provide good timing criticality separation for the timing-based partitioning in heterogeneous 3-D ICs. In addition to the overall PPAC, we want to better understand the detailed design metrics and how they are affected by our heterogeneous methodology.

To do this, we look at the clock tree, critical paths, and the memory interconnect metrics in the commercial CPU core. These metrics are compared across three implementations: best 2-D implementation (12-track), best homogeneous 3-D (12-track), and heterogeneous 3-D. These are shown in Table VIII.

1) Memory Interconnects: The connections to and from memories are generally the largest sources of timing criticality in complex IPs with memory macros. We take a closer look at these paths here. Fig. 4(b) highlights the connections to and from memory macros on the 2-D and heterogeneous 3-D layouts. The connections in yellow are the connections to memory macros, and the nets coming out of the macros are shown in magenta. Visually, we can see the benefit of 3-D in simplifying these nets due to its multitiered placement. The average (root mean square) latencies of these nets are compared across the designs in Table VIII. We see that the latency decreases significantly in 3-D due to a smaller footprint. This further improves with heterogeneous 3-D due to its smaller footprint and closer placement of the 9-track cells. The smaller footprint and reduced wirelength also allow for the reduced switching power on these nets in 3-D and more specifically heterogeneous 3-D.

2) Clock Network: A proper clock tree design is paramount to achieving high-frequency targets. If the clock tree is not well optimized, it can have adverse effects on a large number of paths that might not be fully recoverable with data path optimization. Even when the timing degradation due to a bad clock tree is recovered by cell sizing on data paths, the power impact will be significantly larger. The clock tree has different characteristics across the three implementations of the CPU design in Table VIII. The number of clock buffers decreases slightly in the 12-track 3-D design compared to 2-D due to the smaller footprint requiring a smaller clock tree. However, with heterogeneous 3-D, the clock tree is significantly unbalanced with more than 75% of the clock on the top-die. This arises due to the clock tree methodology and the widely varied timing between the tiers. Most of the clock tree is inserted on the top tier with only a small portion of clock cells on the bottom tier. This, in turn, leads to a smaller cell area of the clock tree, which results in lower clock power. The latency and skew in 3-D are not as good as in 2-D, due to the lack of robust native 3-D clock algorithms. These are significantly worse in heterogeneous 3-D due to the portion of the clock on the slower tier. More importantly, when we look at the skew on the top 100 critical paths, we see that even in heterogeneous 3-D, the skews are well managed due to the clock methodology used in our flow.

3) Critical Path: Finally, we look at several breakdowns of the timing critical path in the three implementations. This helps us analyze the path breakdowns and how critical paths evolve across the three flavors of implementation. We first see that the path delays are pretty similar across the three implementations since the target period is the same (0.833 ns). In homogeneous 3-D, the critical path is 43 cells long with a similar split across the two dies. Moreover, the average cell delay is pretty similar at ≈19 ps in 2-D, and all the 12-track tiers of homogeneous and heterogeneous 3-D ICs. The average cell delay on the 9-track tier of heterogeneous 3-D, on the other hand, is more than 2× at 45 ps. We also see that the logic depth of the critical path in heterogeneous 3-D is the smallest at just 33 cells. 25 out of these 33 cells on the critical path are located on the faster bottom tier, with only eight cells on the top tier. However, due to the slower cells on the top tier, the 8 cells on this slower tier contribute to a large portion of the cell delay. Here, we see that long critical paths do not utilize the slower tier showing the impact of our timing-based partitioning. Additionally, we use a slower tier for placing some of the cells on otherwise noncritical paths can help with reducing the overall power consumption.

D. Guidelines for Designing Heterogeneous 3-D ICs

From Section II-B, we saw that the voltage difference should be kept to a minimal across the process nodes for heterogeneous M3D IC with high interconnect bandwidth. Various modifications to the basic 3-D flow were required to extract the timing, power, and cost benefits from the heterogeneous 3-D ICs. The heterogeneous 3-D methodology works best with complex IPs, where the functionality and timing criticality of the blocks vary. To achieve this, metal layers at either side of the MIV/3-D interface should not vary significantly in pitch to use the commercial routers for finding the best 3-D via locations. If not, the router will place the vias nonoptimally. The clock tree needs to be properly monitored and optimized so that the critical paths do not have large skews. Multitrack variations of a given technology node are usually the best and simplest option to create heterogeneous 3-D ICs as they satisfy both the voltage and BEOL constraints.

V. Conclusion

We have presented a new framework for utilizing the sequentially fabricated 3-D ICs with large 3-D interconnect density. We identified the potential pitfalls of such integration in Section II-B and presented necessary modifications in Section III to extract most out of our proposed heterogeneous 3-D IC. Choosing the right mix of technologies is key for heterogeneous 3-D IC and is currently done manually as metal track variants only, and more exploration is beneficial. Additionally, the current research is done with ideal power delivery, and a thorough study of the power delivery networks for heterogeneous 3-D ICs is required for a complete understanding of heterogeneous 3-D ICs.

Aspects such as clock and routing are modeled accurately, and we found that heterogeneous 3-D ICs significantly improve timing closure with respect to low-power processes and conserve power and chip area with respect to high-performance variants. Combined, the heterogeneous 3-D ICs show a PPAC benefit ranging from 10% to 50% compared to 3-D designs, and the benefit increases to about 18%–57% compared to 2-D.

References

[1] 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.

[2] G. Luo, Y. Shi, and J. Cong, "An analytical placement framework for 3-D ICs and its extension on thermal awareness," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 4, pp. 510–523, Apr. 2013.

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

[4] H. Park, B. W. Ku, K. Chang, D. E. Shim, and S. K. Lim, "Pseudo-3D approaches for commercial-grade RTL-to-GDS tool flow targeting monolithic 3D ICs," in *Proc. Int. Symp. Phys. Design*, Mar. 2020.

[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 3D ICs," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Nov. 2020, pp. 1–9.

[6] K. Chang et al., "Cascade2D: A design-aware partitioning approach to monolithic 3D IC with 2D commercial tools," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2016, pp. 1–8.

[7] 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 3D ICs," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2020, pp. 37–42.

[8] A. I. Arka, B. K. Joardar, R. G. Kim, D. H. Kim, J. R. Doppa, and P. P. Pande, "HeM3D: He terogeneous manycore architecture based on m onolithic 3D vertical integration," *ACM Trans. Design Autom. Electron. Syst.*, vol. 26, no. 2, pp. 1–21, Mar. 2021.

[9] W. Gomes et al., "8.1 Lakefield and mobility compute: A 3D stacked 10nm and 22FFL hybrid processor system in 12×12 mm^2, 1 mm package-on-package," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 144–146.

[10] B. W. Ku, P. Debacker, D. Milojevic, P. Raghavan, and S. K. Lim, "How much cost reduction justifies the adoption of monolithic 3D ICs at 7 nm node?" in *Proc. Int. Conf. Comput.-Aided Design*, 2016.

[11] K. Flamm, "Measuring Moore's law: Evidence from price, cost, and quality indexes," Nat. Bureau Econ. Res., Tech. Rep., Apr. 2018. [Online]. Available: https://www.nber.org/system/files/working_papers/w24553/w24553.pdf

[12] S. Khan and A. Mann. (Apr. 2020). *AI Chips: What They are and Why They Matter*.[Online]. Available: https://www.cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter/

[13] P. Debacker et al., "Low track height standard-cells enable high-placement density and low-BEOL cost (conference presentation)," *Proc. SPIE*, vol. 10148, May 2017, Art. no. 1014803.

[14] S. K. Samal, D. Nayak, M. Ichihashi, S. Banna, and S. K. Lim, "Tier partitioning strategy to mitigate BEOL degradation and cost issues in monolithic 3D ICs," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2016, pp. 1–7.