Hetero-3D: Maximizing Performance and Power Delivery Benefits of Heterogeneous 3D ICs

Lingjun Zhu
Georgia Institute of Technology
Atlanta, GA, USA
lingjun@gatech.edu

Jiawei Hu
Georgia Institute of Technology
Atlanta, GA, USA
jiaweihu@gatech.edu

Gauthaman Murali
Georgia Institute of Technology
Atlanta, GA, USA
gauthaman@gatech.edu

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

ABSTRACT

Heterogeneous 3D integration, blending multiple technology nodes, emerges as a promising strategy for enhancing performance and maintaining low power consumption in next-generation computing systems. This paper presents Hetero-3D, an RTL-to-GDS design flow tailored specifically for heterogeneous 3D ICs. Hetero-3D integrates an area-unbalanced 3D floorplanner with an ML-based power delivery and signal router, working in tandem for rigorous PPA (Power, Performance, and Area) optimization. Using two CPU benchmarks, we showcase a remarkable 15% increase in maximum frequency and a substantial 50% decrease in voltage drop compared to homogeneous 3D baselines. Moreover, Hetero-3D effectively addresses voltage drop issues in the power delivery network while delivering an additional 5% frequency boost. This study emphasizes the EDA solutions that unlock the potential of mixed-node stacking as a crucial enabler for performance scaling in future ICs.

CCS CONCEPTS

• **Hardware → 3D integrated circuits**.

KEYWORDS

Heterogeneous Integration, Physical Design, Power Delivery

ACM Reference Format:
Lingjun Zhu, Jiawei Hu, Gauthaman Murali, and Sung Kyu Lim. 2024. Hetero-3D: Maximizing Performance and Power Delivery Benefits of Heterogeneous 3D ICs. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED '24), August 5–7, 2024, Newport Beach, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3665314.3670850

1 INTRODUCTION

As performance and power scaling trends slow down in advanced technology nodes, heterogeneous 3D integration emerges as a robust solution poised to tackle these challenges. This methodology facilitates the vertical stacking of various technology nodes or components to form a unified 3D integrated circuit (IC). Such stacking capabilities are instrumental in surmounting performance or power limitations inherent in a single technology node while enabling the integration of diverse functionalities within a System-on-Chip (SoC). Notable commercial examples, such as Intel's Lakefield CPU [3] and AMD's 3D V-Cache [9], underscore the substantial improvements in Power, Performance, and Area (PPA) metrics attained through heterogeneous 3D integration.

Table 1: Qualitative PPA and power integrity comparison among various 3D configurations. A quantitative comparison is presented in Table 3.

	28nm 3D	16nm 3D	Hetero-3D
Power	High	Low	Medium
Performance	Low	High	Medium
Area	Large	Small	Medium
Cost	Low	High	Medium
IR drop	Low	High	Medium

The design and optimization of heterogeneous 3D ICs necessitate customized physical design flows and Electrical Design Automation (EDA) tools. Existing EDA software primarily targets 2D or homogeneous 3D ICs, lacking direct support for multiple technology nodes and high-density 3D interconnects. In addition, power delivery is becoming an escalating concern in 3D ICs due to their high power density and long power delivery path. Prior studies, such as [6], have explored the potential EDA solutions for heterogeneous 3D, emphasizing the distinctions from conventional 2D ICs.

In [5], the authors presented heuristic algorithms for 3D IC Placement and Routing (PNR) based on analytical models, but these approaches have not addressed the complexities associated with heterogeneous integration across different process nodes. Similarly, in [1, 8], the authors expanded the capabilities of commercial EDA tools to enable Memory-on-Logic (MoL) 3D ICs, while still constrained by uniform technology node and die area for both tiers. Consequently, there exists a pressing need to develop advanced EDA flows capable of accommodating more flexible heterogeneous 3D IC configurations, spanning various technologies, floorplan arrangements, and considerations for power integrity.

In this paper, we introduce a tailored physical design flow, termed as Heterogeneous 3D (Hetero-3D), specially engineered to capitalize on the advantages offered by fine-pitch hybrid bonding, integration of multiple technology nodes, and asymmetrical floorplan. Fig. 1 shows the cross-section view of the proposed Hetero-3D IC structure. A general intuition about heterogeneous integration is that it can strike a balance between PPA, cost, and power integrity by combining the benefits of multiple technology nodes, as summarized in Table 1.

ISLPED '24, August 5–7, 2024, Newport Beach, CA, USA

ACM ISBN 979-8-4007-0688-2/24/08…$15.00
https://doi.org/10.1145/3665314.3670850

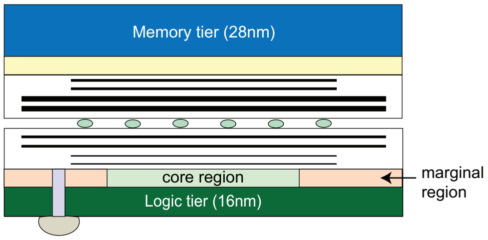

Figure 1: Face-to-face heterogeneous 3D IC used in this paper.

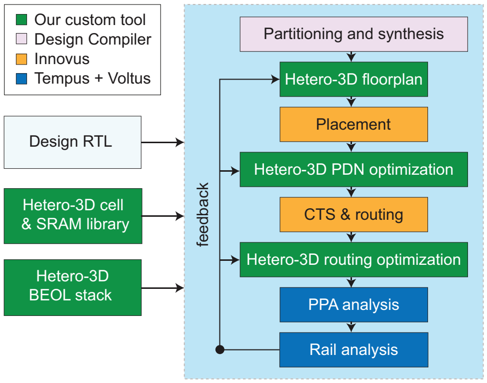

Figure 2: Overview of our Hetero-3D design flow that combines commercial tools and our custom plug-ins.

To scrutinize these intuitions, we quantify the power integrity benefits of Hetero-3D ICs and propose physical design methods aimed at translating these advantages into performance enhancements. Specifically, we implement dedicated EDA approaches to set up Hetero-3D technology, rapidly generate unbalanced floorplans, and perform Power Delivery Network (PDN)/routing optimization for Hetero-3D ICs. We demonstrate the PPA advantages alongside the power integrity benefits of Hetero-3D ICs based on two CPU processors and two commercial technology nodes. Furthermore, we analyze the root causes of these impacts and then demonstrate additional performance gain with heterogeneous 3D ICs.

2 HETERO-3D METHODOLOGY

2.1 Overview of the Approach

The proposed Hetero-3D design flow is specifically tailored for integrating different technology nodes and MoL stacking, which also allows an unbalanced floorplan on each tier. Fig. 2 illustrates the overview of our RTL-to-GDS flow.

In this work, we consider Hetero-3D ICs with two tiers: a memory tier (on the top) and a logic tier (on the bottom). Only SRAM blocks are allowed to be placed on the memory tier, while standard cells and small SRAM blocks can be placed on the logic tier. The flow allows the logic tier to have a smaller core region (meaning the region with most of the logic cells) than the memory tier, making the tier partitioning and floorplan more flexible. The marginal region is filled with decoupling capacitors (decaps) and signal buffers.

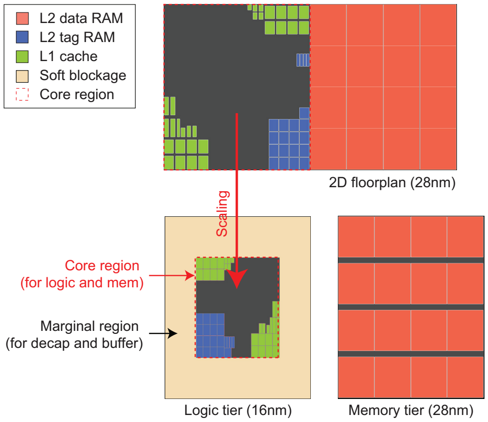

Figure 3: Tier partitioning and floorplanning for memory macros in a Hetero-3D IC.

Moreover, we exploit the heterogeneity of different technology nodes: a 28nm technology node for the memory tier and a more advanced 16nm technology node for the logic tier. Based on commercial 28nm and 16nm 2D Process Design Kits (PDKs), we create a Hetero-3D Back End Of Line (BEOL) stack by stacking 28nm metal layers on the top of the 16nm layers. These 16nm metal layers are thin and highly resistive, while the 28nm metal layers are wide and less resistive. For standard cell and SRAM libraries, we first obtain the original 2D version from the commercial 28nm and 16nm PDKs. We provide a supply voltage $V_{DD} = 0.8V$ to both tiers and generate the cell and SRAM libraries based on the voltage corner. Based on the technology setup, we build two commercial-grade CPUs as the benchmark designs for this study. We configure the CPUs with a large L2 cache capacity to enable the MoL stacking.

2.2 Hetero-3D Partitioning and Floorplanning

We adopt the constraint-graph-based macro placement method from [7] to generate the initial floorplan. To summarize, the general idea is to partition larger SRAM blocks to the memory tier with an older technology node to exploit the performance and power integrity advantages. Compared with the previous work, the key difference of our Hetero-3D flow is that we allow unbalanced floorplans. With technology scaling, the 16nm standard cells occupy a much smaller area (-70%) than the 28nm ones. SRAM area also scales with a lower ratio. We propose to use the empty space to improve the power integrity of 3D ICs.

Fig. 3 shows an example of tier partitioning and floorplanning in the Cortex-A53 Hetero-3D IC. Starting with a 28nm 2D floorplan, we select 16 L2 data RAMs to be partitioned into the memory tiers, and the remaining SRAM blocks and standard cells are to be placed on the logic tier. Then, the SRAM blocks on each tier are clustered and placed based on the constraint graphs. As highlighted with the red rectangle in the figure, the logic tier has a smaller core region than the memory tier, as the cells and SRAMs are scaled from 28nm to 16nm on the logic tier. We put the core region in the center of the floorplan for shorter interconnects and better clock quality, and the surrounding area will be filled with decaps and buffers.

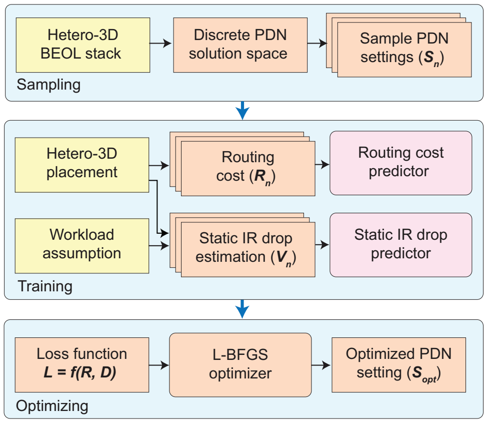

Figure 4: Our proposed PDN optimization flow.

Table 2: PDN parameters we optimize for Hetero-3D IC (shown in green). u_1 and u_5 are pre-fixed by the cell library and memory compiler.

	Layer	Direction	Width	Pitch	Util.
Logic tier	M1	H	w_1	p_1	u_1
	M5	V	w_5	p_5	u_5
	M6	H	w_6	p_6	u_6
	M7	V	w_7	p_7	u_7
	M8	H	w_8	p_8	u_8
Memory tier (MT)	M8$_{MT}$	V	w_9	p_9	u_9
	M7$_{MT}$	H	w_{10}	p_{10}	u_{10}
	M6$_{MT}$	V	w_{11}	p_{11}	u_{11}

2.3 Hetero-3D PDN Design and Optimization

We define the PDN with a set of variables: stripe width (w_i), stripe pitch (VDD-to-VSS pitch, p_i), and track utilization (u_i) for each layer i. Therefore, the goal of PDN optimization is to find the best PDN settings that balance both performance and power integrity. We propose a PDN design and optimization flow for Hetero-3D ICs based on Machine Learning (ML) models, as shown in Fig. 4.

2.3.1 *Solution Space Creation:* To reduce computational costs, we create a discrete PDN solution space based on the Hetero-3D BEOL stack. Similar to conventional PDN design in 2D ICs, we enable PDN mesh on upper metal layers (M6-M8) on each tier, with additional vertical PDN stripes enabled on M5 of the logic tier to connect to the horizontal standard cell power rails on M1. An example PDN setting for the Hetero-3D IC is shown in Table 2.

On M1 and M5 of the logic tier, the power stripes are directly connected to the cell and SRAM power or ground pins, and the stripe width and pitch are fixed based on the library requirements. For other layers, we fix the stripe pitch (p_i) to the maximum metal width. Then, we calculate the irredundant stripe width based on the minimum routing width and spacing of each layer. The irredundant stripe width is a set of discrete width values that minimizes the number of routing tracks occupied by the PDN stripes given a certain stripe pitch and PDN usage [2]. The 28nm metal layers (M6$_{MT}$-M8$_{MT}$) offer a larger solution space for stripe width because these layers allow wider wires. As a result, Hetero-3D provides a larger room for PDN optimization than homogeneous 3D. Also, each stripe width is one-to-one corresponding to a utilization value ($u_i = w_i/p_i$), and thus, we can use the utilization as the main variable to represent the PDN setting in the optimization process.

2.3.2 *Routing Cost and IR Drop Prediction:* We develop a routing cost model with PDN usage ($\boldsymbol{u}$) as the input feature and wire resistance ($\boldsymbol{R}$) as the output feature. The model is similar to what has been proposed in [2], but significant modifications are made for Hetero-3D. To avoid overfitting in the ML model, care must be taken to select the objective and the features used for prediction. First, we adopt the total wire resistance as the objective function for routing cost estimation, incorporating the various resistivity contributed by the Hetero-3D BEOL stack instead of the total wirelength as in the original work. Moreover, we implement a set of Gaussian Process Regression (GPR) models to predict the wire resistance on each layer separately instead of using one model for the total resistance. This provides more fine-grained control of the regression model and allows us to incorporate prior knowledge about the BEOL stack. For example, We adopt an anisotropic Radial Basis Function (RBF) kernel for GPR model on layer i, with various length scale $l_{i,k}$ for input feature (u_k):

$$l_{i,k} = 10^{|i-k|} \tag{1}$$

This kernel function allows the model to have a smaller length scale for the input feature on the nearby layers and a larger length scale for the input feature on distant layers. After fitting the GPR model for each layer, we can predict the wire resistance on each layer separately. Then, we sum up the wire resistance on all layers to obtain the total wire resistance.

Similarly, we implement another GPR model to estimate the worst IR drop. The model is trained with the PDN usage ($\boldsymbol{u}$) as the input feature and IR drop ($\boldsymbol{V}$) as the output feature. The training data is generated with what-if IR drop simulation from Voltus by scaling the PDN resistance based on the stripe width. This approach is much faster than rebuilding the PDN for each sample.

2.3.3 *PDN Optimization.* We formulate the PDN optimization problem as a constrained optimization problem. The goal is to find the best PDN setting that minimizes the total wire resistance while satisfying the maximum IR drop constraint ($V_{thresh} = 0.10 \times V_{DD}$). The optimization problem can be formulated as:

$$\begin{aligned} \underset{\boldsymbol{u}}{\text{minimize}} \quad & \boldsymbol{R}(\boldsymbol{u}) \\ \text{subject to} \quad & \boldsymbol{V}(\boldsymbol{u}) \le V_{thresh} \end{aligned} \tag{2}$$

Therefore, the loss function of the optimization problem can be converted into a Lagrangian function:

$$\mathcal{L}(\boldsymbol{u}, \lambda) = \boldsymbol{R}(\boldsymbol{u}) + \lambda \times (\boldsymbol{V}(\boldsymbol{u}) - V_{thresh}) \tag{3}$$

where λ is the Lagrangian multiplier. Using the trained GPR models, we can quickly evaluate the loss function for each PDN setting. Therefore, we employ the Limited-memory Broyden, Fletcher, Goldfarb, Shanno (L-BGFS) algorithm [4] to solve the optimization problem.

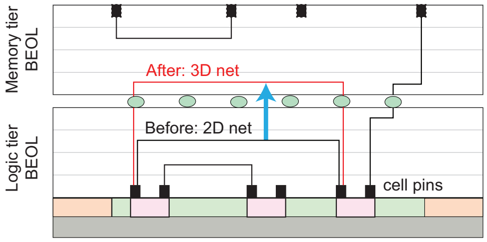

Figure 5: Hetero-3D routing optimization based on metal layer sharing.

2.4 Hetero-3D Routing Optimization

In our Hetero-3D flow, we use the standard 2D EDA tool (Innovus) to place the cells on the logic tier directly, with regard to the Hetero-3D stack, projected memory pins, and placement blockages. Similarly, we utilize Innovus to perform signal routing in Hetero-3D ICs. In our Hetero-3D BEOL stack, the metal track density and resistivity change abruptly at the hybrid bonding layer, and the 16nm metal layers have significantly higher resistivity and track density than the 28nm ones. As a result, the conventional router may not be able to fully utilize the routing resources on the two tiers and thus lead to PPA degradation.

To overcome this issue, we propose a routing optimization method that leverages the benefits of heterogeneous stacking. The main idea is to utilize the less resistive and less congestive metal layers on the memory tier to route logic-tier nets, which can reduce the routing congestion and wire delay on the logic tier, as shown in Fig. 5. This concept has been described in a previous work [1] as *metal layer sharing* for signal routing. However, it has not been implemented based on a Hetero-3D IC or incorporated with PDN designs. Our approach is to enable and adjust metal sharing for signal routing during post-route optimization. We select a set of nets on the top timing critical paths and then apply additional routing rules to encourage the router to use the memory tier for routing when available. By doing this iteratively, we can further improve the utilization of the routing resources on the two tiers and thus improve the PPA of the Hetero-3D design.

2.5 Hetero-3D Decap Insertion

After routing optimization, we perform decap insertion in the Hetero-3D design to mitigate the dynamic voltage drop. The decap insertion is performed in two steps: *decap insertion in the marginal region* and *decap insertion in the core region*. In the first step, we insert as many decaps as possible into the marginal region reserved during floorplanning. As this region is mostly empty (except for a small number of signal buffers), we choose the decap cell with the highest capacitance-to-area ratio from the cell library and place it densely in the marginal region. Then, for the core region, we use the conventional decap insertion method to insert decaps in the empty space between standard cells and SRAM blocks.

Then, we perform rail analysis to evaluate the static IR drop and dynamic voltage drop of the Hetero-3D ICs. We utilize the *max power* program (provided by the IP vendor) as the workload for the rail analysis. During static analysis, the average switching activity is calculated for each net based on the workloads, and the IR drop is evaluated. During dynamic analysis, the tool goes through the entire time window (10 ns) of the workloads and captures the worst-case voltage drop (including the $L \cdot \mathrm{d}I/\mathrm{d}t$) drop. The resulting static IR drop map and dynamic voltage drop map are used for the power integrity and PDN design quality for the Hetero-3D ICs.

Table 3: Quantitative comparison among various memory-on-logic 3D configurations. Hetero-3D A53 with 16nmm logic and 28nm memory shows 16nm-quality performance (fast) with 28nm-grade IR-drop (low noise).

	28nm	16nm	Hetero
Power (mW)	343.8	242.3	241.2
Frequency (MHz)	750.0	880.0	861.9 (almost 16nm)
Area (mm^2)	2.18	1.13	2.11
Estimated cost	1.00	1.73	1.55
IR drop (mV)	55.2	67.2	59.7 (almost 28nm)

3 EXPERIMENTAL RESULTS

3.1 Quantitative Comparison Summary

Using the state-of-the-art Macro-3D flow [1], we build the two baseline homogeneous 3D ICs with 28nm and 16nm technology nodes, respectively. Table 3 summarizes the results in the Cortex-A53 benchmark. The trends of PPA metrics match our expectation, with Hetero-3D ICs achieving a balance in performance and IR drop. We perform more experiments to analyze the root causes of the benefits and maximize the performance of Hetero-3D ICs.

3.2 Heterogeneous 3D vs. Homogeneous 3D

First, the Hetero-3D ICs are designed with the same target frequency as the homogeneous 3D ICs and a default PDN setting (PDN utilization is equal to 60% for all configurable PDN layers). The total runtime of the design flow is 8 hours and 40 mins for Cortex-A53 and 6 hours and 57 mins for Cortex-A7 with 8 Intel XEON CPUs.

Unlike the Hetero-3D IC, the homogeneous 3D baselines, (a) 28nm 3D and (b) 16nm 3D, have a balanced floorplan. Fig. 6 shows the placement layouts of Cortex-A53 Hetero-3D ICs, together with the homogeneous 3D baselines. The decap and the PDN stripes are not shown in the layouts for clarity. As shown in both Hetero-3D ICs, the core region of the logic tier significantly shrinks as technology scales down from 28nm to 16nm.

We observe that the 16nm homogeneous 3D IC (b) achieves the highest effective frequency and lowest power consumption among the three configurations, while the 28nm homogeneous 3D IC (a) shows the lowest effective frequency and highest total power. The effective frequency is the maximum achievable frequency estimated based on the target frequency and worst slack. This is because the 16nm standard cells and SRAMs provide higher performance and lower power compared with the 28nm ones. On the other hand, the Hetero-3D IC (c) reaches a middle ground of the two homogeneous 3D ICs in terms of both frequency and power consumption, as it combines the two technology nodes and thus opens up new opportunities for performance and power optimization.

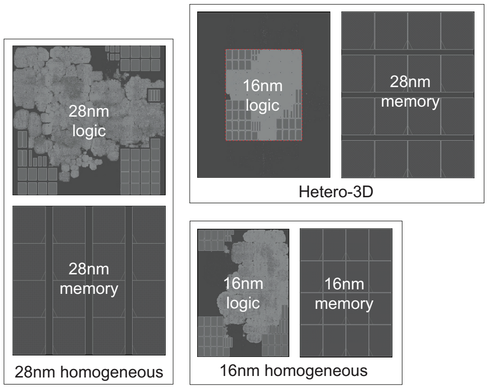

Figure 6: Three flavors of placement for Cortex-A53 3D ICs.

The layout on Fig. 6 shows that the placer honors the placement blockage and only places the logic cells in the core region. This leads to an unbalanced Hetero-3D floorplan and dense cell placement on the logic tier, which benefits signal routing. Also, the 3D ICs utilize a large number of hybrid bonding pads and signal routing wires on the memory tier. The reason is that we enable Hetero-3D routing to share metal resources across the tiers. As a result, the Hetero-3D ICs achieve a much shorter wirelength (up to -38%) compared with the 28nm 3D IC (a).

3.3 Power Integrity Benefits

The Hetero-3D IC (c) achieves 30% and 26% power reduction compared with the 28nm 3D IC (a) in Cortex-A53 and Cortex-A7 design, respectively. The power reduction in the Hetero-3D IC mainly stems from internal power and sequential power saving. Compared with the 28nm 3D IC, the Hetero-3D IC utilizes 16nm technology for the standard cells, which offers lower cell power at the same speed. As a result, (c) achieves 41% lower internal power compared with (a).

Based on the power analysis results, we generate the 3D power density maps in Voltus as shown in Fig. 7. The power consumption from both tiers is projected to a 2D plane, and the average power density is annotated. Among the three configurations, the 16nm 3D IC (b) suffers from the highest power density due to the small cell size and footprint area. Additional power hotspots are created where the memory-tier SRAM and logic-tier cells overlap, which propose challenges for vertical power delivery. The Hetero-3D IC (c) has a similar maximum power density as (b), but it demonstrates the lowest average power density thanks to the large footprint area and marginal area of the logic tier. Therefore, the Hetero-3D IC offers a larger room to distribute the power consumption and tackle the power delivery issues.

The proposed Hetero-3D structure also provides significant power delivery benefits. We perform static IR and dynamic rail analysis to quantify the impacts of heterogeneous integration on power delivery. Table 4 summarizes the power integrity metrics. The least resistive path (LRP) represents the least resistive path from the power and ground sources to the IR drop hotspots. These results show that the 16nm homogeneous 3D IC (b) suffers from high static IR and dynamic voltage drop due to its high power density and metal resistivity, while the 28nm homogeneous 3D IC (a) has the best power integrity in terms of voltage drop.

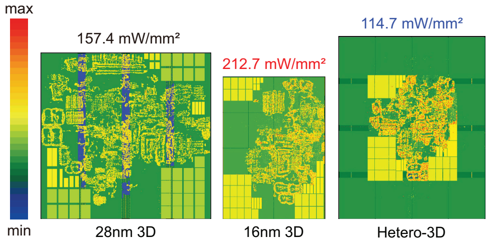

Figure 7: Power density maps and the average values of memory-on-logic A53 3D ICs.

Table 4: Power integrity metrics in the Cortex-A53 3D ICs. RLRP means the resistance of the least resistance path.

Configuration	28nm 3D	16nm 3D	Hetero-3D
Static IR analysis			
Static drop (mV)	7.16	22.27	11.13
Static RLRP (Ω)	47.78	423.66	205.84
Dynamic analysis			
Dynamic drop (mV)	55.22	67.19	59.65
Dynamic RLRP (Ω)	64.11	839.58	459.23
Marginal decap (nF)	0.00	0.00	8.70
Core decap (nF)	0.25	0.61	0.15
Intrinsic decap (nF)	2.21	2.23	2.25
Total decap (nF)	2.46	2.84	11.10

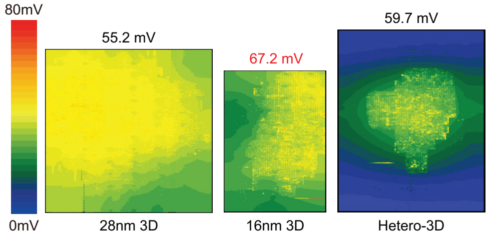

Figure 8: Dynamic IR drop maps and the maximum values of memory-on-logic A53 3D ICs.

On the other hand, the Hetero-3D IC (c) mitigates the power delivery issues in 16nm technology by leveraging the 28nm memory tier. The Hetero-3D IC achieves 50% and 11% lower static and dynamic voltage drop compared with (b), respectively. The reason is that the LRP in Hetero-3D bypasses the resistive 16nm metal layers with the 28nm stripes, which leads to a lower resistance (0.5X). Also, the Hetero-3D IC has a lower power density (0.5X), and more decaps (3.9X) inserted.

3.4 Boosting Performance Further

In this section, we deploy the customized PDN and routing optimization approaches to redesign the PDN in Hetero-3D ICs in order to achieve higher performance. We first train the ML models with $N = 100$ samples for each design to predict the routing cost (total

Table 5: Max-performance PPA and voltage drop comparisons. Δ_{28} and Δ_{16} show the differences w.r.t 28nm and 16nm 3D.

Design	Cortex-A53					Cortex-A7				
Configuration	28nm 3D	16nm 3D	Hetero-3D	Δ_{28}	Δ_{16}	28nm 3D	16nm 3D	Hetero-3D	Δ_{28}	Δ_{16}
Core area (mm^2)	2.2	1.1	0.8	-65.2%	-33.0%	1.1	0.8	0.4	-62.1%	-45.5%
Memory area (mm^2)	2.2	1.1	2.1	-3.3%	86.3%	1.1	0.8	1.1	0.0%	43.9%
Wirelength (m)	21.13	12.81	12.83	-39.3%	0.2%	12.25	8.08	8.56	-30.1%	5.9%
Bonding pad #	67853	17199	17093	-74.8%	-0.6%	18599	11428	14254	-23.4%	24.7%
Total power (mW)	508.89	337.87	336.45	-33.9%	-0.4%	388.93	267.50	277.69	-28.6%	3.8%
Effective freq. (MHz)	762.7	1012.0	1052.4	38.0%	4.0%	1027.2	1035.3	1084.4	5.6%	4.8%
Static drop (mV)	7.3	23.3	12.1	65.7%	-48.0%	14.8	25.0	17.6	19.0%	-29.5%
Dynamic drop (mV)	80.9	80.1	65.7	-18.8%	-18.0%	73.8	76.6	74.9	1.4%	-2.3%

Table 6: Optimized PDN parameters for Hetero-3D ICs.

Layer	Variables	Cortex-A53	Cortex-A7
M6	u_6	0.38	0.57
M7	u_7	0.38	0.45
M8	u_8	0.38	0.46
M8$_{MT}$	u_9	0.62	0.46
M7$_{MT}$	u_{10}	0.38	0.47
M6$_{MT}$	u_{11}	0.62	0.65

wire resistance) and static IR drop based on our iso-performance Hetero-3D ICs. Then, we use the trained models to perform PDN optimization, with the objective of minimizing the total wire resistance and the constraint of static IR drop (less than 80mV).

In the optimized settings, PDN usages are significantly reduced on the logic tier (u_6 - u_8) as the optimizer tries to minimize the wire resistance contribution from the 16nm metal layers. On the other hand, the PDN usage on the memory tier is increased (u_9 - u_{11}) because the optimizer tries to leverage the 28nm metal layers to reduce the wire resistance. For example, in Cortex-A53, the low IR drop value from the default PDN design suggests a large headroom for PDN downsizing, and thus it arrives at a solution with lower PDN utilization than the default setting (60% utilization). For Cortex-A7, the default PDN design already has a high IR drop, and thus the optimizer reaches a more conservative solution.

After applying the optimized PDN settings to Hetero-3D ICs, we push the clock frequency to the limit for all the designs. The maximum achievable frequency is defined as the frequency point where the 3D ICs can satisfy both the timing and dynamic IR drop (less than 80 mV) constraints. As shown in Table 5, the Hetero-3D ICs achieve 4.0% and 4.8% higher frequency compared with the 16nm 3D ICs in Cortex-A53 and Cortex-A7, respectively.

The performance improvements of Hetero-3D originate from lower path delays and larger voltage drop headroom in both Cortex-A7 and A53. We select Cortex-A53 for a detailed timing analysis. The timing critical path of Cortex-A53 is a register-to-register path. Compared with the 28nm homogeneous baselines, the Hetero-3D IC uses the high-performance 16nm cells on the logic tier for buffering and similar 28nm metal wires on the memory tier for signal routing. By combining the advantages of two process nodes, the Hetero-3D IC achieves a lower path delay (-8.0%) and clock latency (-6.2%) compared with the 28nm homogeneous 3D.

The Hetero-3D overcomes the frequency throttling issue caused by high voltage drop. As shown in Table 5, the dynamic voltage drop of the 16nm homogeneous 3D IC is at the threshold, which prevents further frequency increase. However, the power delivery benefits of Hetero-3D enable more voltage drop headroom with the dedicated PDN design and decap placement. In other words, it allows us to trade the power integrity improvements into up to 4.8% frequency boost. With these experiments, we demonstrate that the Hetero-3D ICs can achieve higher performance with manageable power integrity, compared with homogeneous 3D ICs.

4 CONCLUSIONS

In this paper, we present the Hetero-3D flow to enable the 3D integration of two technology nodes with an unbalanced floorplan. It fully utilizes the silicon and metal resources of heterogeneous 3D stacking to optimize the PPA. We identify two key benefits: (1) Hetero-3D ICs leverage the power and performance benefits offered by multiple technology nodes. (2) The Hetero-3D floorplan and PDN optimization help overcome the power delivery bottlenecks in 3D ICs and achieve a balance between timing and power integrity.

ACKNOWLEDGMENTS

This work was funded by the SRC under Task 3136.002.

REFERENCES

[1] Lennart Bamberg et al. 2020. Macro-3D: A Physical Design Methodology for Face-to-Face-Stacked Heterogeneous 3D ICs. In *2020 Des. Autom. Test Eur. Conf. Exhib. DATE.* (Mar. 2020), 37–42.

[2] Wen-Hsiang Chang et al. 2016. Generating Routing-Driven Power Distribution Networks with Machine-Learning Technique. In *Proc. 2016 Int. Symp. Phys. Des.* (ISPD '16). (Apr. 3, 2016), 145–152.

[3] Wilfred Gomes et al. 2020. 8.1 Lakefield and Mobility Compute: A 3D Stacked 10nm and 22FFL Hybrid Processor System in 12×12mm2, 1mm Package-on-Package. In *2020 IEEE Int. Solid-State Cir. Conf. - ISSCC.* (Feb. 2020), 144–146.

[4] Dong C. Liu et al. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45, 1, (Aug. 1, 1989), 503–528.

[5] Jingwei Lu et al. 2016. ePlace-3D: Electrostatics based Placement for 3D-ICs. In *Proc. 2016 Int. Symp. on Phys. Des.* (ISPD '16). (Apr. 3, 2016), 11–18.

[6] Dragomir Milojevic et al. 2013. Design issues in heterogeneous 3D/2.5D integration. In *Asia and South Pacif. Des. Autom. Conf. (ASP-DAC)*, 403–410.

[7] H. Murata et al. 1996. VLSI module placement based on rectangle-packing by the sequence-pair. *IEEE Tran. on Comp.-Aided Des. of Int. Cir. and Sys.*, 15, 12, (Dec. 1996), 1518–1524.

[8] Sebastien Thuries et al. 2020. M3D-ADTCO: Monolithic 3D Architecture, Design and Technology Co-Optimization for High Energy Efficient 3D IC. In *2020 Des., Auto. & Test in Euro. Conf. & Exhibition (DATE).* (Mar. 2020), 1740–1745.

[9] John Wuu et al. 2022. 3D V-Cache: the Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU. In *2022 IEEE Int. Solid- State Circuits Conf. ISSCC.* Vol. 65. (Feb. 2022), 428–429.