Road to High-Performance 3D ICs: Performance Optimization Methodologies for Monolithic 3D ICs

Kyungwook Chang, Sai Pentapati, Da Eun Shim, and Sung Kyu Lim
School of ECE, Georgia Institute of Technology, Atlanta, GA
k.chang@gatech.edu, limsk@ece.gatech.edu

ABSTRACT

As we approach the limits of 2D device scaling, monolithic 3D IC (M3D) has emerged as a potential solution offering performance and power benefits. Although various studies have been done to increase power savings of M3D designs, efforts to improve their performance are rarely made. In this paper, we, for the first time, perform in-depth analysis of the factors that affect the performance of M3D, and present methodologies to improve the performance. Our methodologies outperform the state-of-the-art M3D design flow by offering 15.6% performance improvement and 16.2% energy-delay product (EDP) benefit over 2D designs.

CCS CONCEPTS

• **Hardware → 3D integrated circuits**; **Physical design (EDA)**;

KEYWORDS

Monolithic 3D ICs, High-performance, Optimization

1 INTRODUCTION

With extensive data and memory requirements, high performance has become a game changing design factor in the modern world. With benefits coming from 2D device scaling slowly saturating, monolithic 3D IC (M3D) provides significant power savings and has been in the spotlight in recent years. M3D makes use of nano-sized monolithic inter-tier vias (MIVs) to connect sequentially fabricated top and bottom tiers, which allows for fine-grained vertical integration that yields performance and power benefits.

Since current commercial EDA tools do not support 3D placement of cells, several studies have devised M3D design flows using 2D commercial EDA tools [3, 8, 9].

Previous studies have explored and shown the performance optimization for 3D ICs [5, 7]. However, these studies have focused primarily on investigating through silicon via (TSV)-based 3D ICs, which inevitably show much lower vertical integration densities due to the larger micron-scale size and the keep-out zone (KOZ) of TSVs. TSV-based 3D ICs fail to fully benefit from 3D IC stacking due to the influence that course granularity of vertical integration has on wire-length and latency. There has been active research on power benefits of M3D designs over 2D counterparts with the M3D design methodologies, offering quantifiable power savings [4, 9]. However, none of these work attempted performance optimization for M3D but instead focused only on power benefits. The authors of [2] proposed methodologies to improve the performance of M3D designs by vertically stacking diffusion area or cells, but their methods involve custom cell designs, which require excessive effort and time.

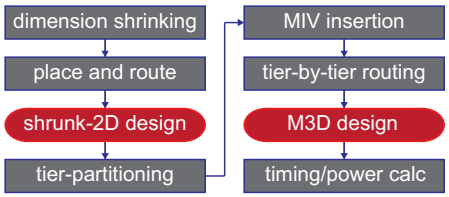

Figure 1: The M3D design flow [9] used as the baseline of this work

In this paper, we perform comprehensive studies on optimizing the performance of M3D designs. The main contributions of this paper are as follows: (1) We quantify the performance benefit of M3D designs over 2D designs. (2) To the best of our knowledge, this is the first work which presents methodologies to maximize the performance benefits of M3D designs. (3) We perform in-depth analysis on the key factors that affect the performance.

2 M3D FULL-CHIP DESIGN FLOW

Due to the lack of ability of current EDA tools to place cells in 3D space, the shrunk-2D design flow [9] uses dimensional shrinking techniques along with 2D commercial EDA tools to implement M3D designs. Figure 1 summarizes the design flow. An M3D design is implemented in a footprint half the size of its 2D counterpart by utilizing two tiers. In order to make use of 2D commercial EDA tools to place all cells and wires on the halved footprint, the width and height of standard cells are first scaled down by $1/\sqrt{2}$. The width and pitch of metal layers are scaled down as well to compensate for the shrunk footprint.

The shrunk cells and wires are used throughout all design stages (including placement, clock-tree-synthesis (CTS), and routing) to implement a *shrunk-2D design*, which provides the optimized x-y placement of cells. The cells in the *shrunk-2D design* are then expanded back to their original size, resulting in overlapping instances. In order to remove the overlaps and place cells on two tiers, the design is first divided into multiple partitioning bins on the x-y-plane. Then, for each bin, the cells are partitioned into two tiers using an area-balanced min-cut partitioning algorithm (i.e., the z-location of cells is determined). This algorithm minimizes the

ISLPED '18, July 23–25, 2018, Seattle, WA, USA

ACM ISBN 978-1-4503-5704-3/18/07…$15.00
https://doi.org/10.1145/3218603.3218636

number of vertical connections between two tiers while balancing the cell area of the top and bottom tier. The cells are then legalized by moving along the x-y-plane to remove any overlaps remaining after tier partitioning.

In order to determine the location of MIVs, the metal stack is duplicated to account for metal layers in each tier. The cells are annotated with their respective tiers, and their pins are assigned to metal layers on their corresponding metal stack. The top and bottom tiers are merged into a single design and routed using a 2D commercial EDA tool. The locations of MIVs are determined by the location of vias between the top metal layer of the bottom tier and the bottom metal layer of the top tier. The top and bottom tier designs are then created by performing tier-by-tier routing with the cell placement of each tier and MIV location information. Lastly, timing and power analysis are done on the resulting tier designs along with MIV parasitics.

3 M3D PERFORMANCE IMPROVEMENT

In order to quantify the performance of a design, we employ the slack which describes the timing closure of a timing path in a design, and worst negative slack (WNS) which determines the maximum performance of the design. A negative WNS indicates that the design fails to close timing, and a slower clock frequency is required The slack of a timing path is determined by Equation (1).

$$\text{slack} = T_{CP} + T_{skew} - T_{setup} - D_{path}, \tag{1}$$

where T_{CP} and T_{skew} indicate the clock period and clock skew between the start and end point of a timing path, respectively. T_{setup} is the setup time of the end point, and D_{path} represents the total path delay of the combinational logic in the timing path.

Performance optimization for M3D designs is performed with three techniques: parasitic adjustment during implementation of *shrunk-2D designs*, path-based tier partitioning and clock buffer tier partitioning while transforming *shrunk-2D designs* into M3D designs.

3.1 Parasitic Adjustment

A *shrunk-2D design* is an emulated 2D version of the corresponding M3D design. When placing the cells, their drive-strength, x-y location, and buffer insertion are determined during implementation of the *shrunk-2D design* which will not be changed afterwards. Since cell placement and optimizations are based on wire parasitics, it is crucial for the *shrunk-2D design* to estimate wire parasitics of the final M3D design correctly as described in Equation (2).

$$\begin{aligned} R_{S2D} &= R_{M3D}, \quad C_{S2D,GND} = C_{M3D,GND} \\ C_{S2D,X,h} &= C_{M3D,X,h}, \quad C_{S2D,X,v} = C_{M3D,X,v}, \end{aligned} \tag{2}$$

where R and C_{GND} represent the resistance and ground capacitance of a wire. $C_{X,h}$ and $C_{X,v}$ are the horizontal (i.e., x-y-plane) and vertical (i.e., z-axis) coupling capacitance of the wires, respectively.

Current EDA tools utilize RC parasitic look-up tables (i.e., qrcTech-File or nxtgrd) to determine the resistance and capacitance of wires based on geometric and material parameters of wires. Therefore, the dimensional shrinking of wires makes the *shrunk-2D design* suffer from inaccurate RC parasitic estimation of the final M3D design.

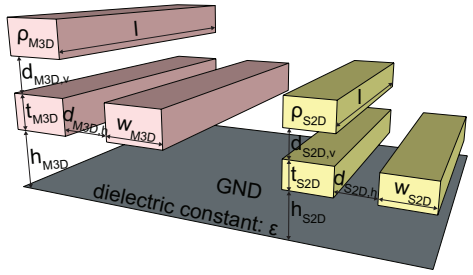

Figure 2: Comparison of wire geometries in M3D (pink) vs. *shrunk-2D design* (yellow)

The resistance, and the ground and coupling capacitances of the wires are described by the following equations.

$$\begin{aligned} R = \rho \cdot l/(t \cdot w), \quad C_{GND} = \epsilon \cdot (w \cdot l)/h \\ C_{X,h} = \epsilon \cdot (t \cdot l)/d_h, \quad C_{X,v} = \epsilon \cdot (w \cdot l)/d_v, \end{aligned} \tag{3}$$

where ρ and ϵ indicate the resistivity of the wire and the dielectric constant of the dielectric layer. w, h, and t represent the width, height from the ground, and thickness of the wire, whereas d_h and d_v are the horizontal and vertical spacing of the wires, respectively.

Figure 2 compares wires in an M3D design, and the corresponding wires in the *shrunk-2D design* with the same length, l. The following constraints are applied on the width and horizontal spacing of the wires in the *shrunk-2D design* due to the dimensional shrinking on the x-y-plane.

$$w_{S2D} = w_{M3D}/\sqrt{2}, \qquad d_{S2D,h} = d_{M3D,h}/\sqrt{2} \tag{4}$$

From Equation (3) and Equation (4), we derive Equation (5), which describes the RC parasitics estimated in the baseline shrunk-2D design flow as it assumes all parameters (except the width and horizontal spacing) are the same in the *shrunk-2D* and the corresponding M3D design.

$$\begin{aligned} R_{S2D-base} &= \sqrt{2} \cdot R_{M3D} \\ C_{S2D-base,GND} &= C_{M3D,GND}/\sqrt{2} \\ C_{S2D-base,X,h} &= \sqrt{2} \cdot C_{M3D,X,h} \\ C_{S2D-base,X,v} &= C_{M3D,X,v}/\sqrt{2} \end{aligned} \tag{5}$$

The above equation indicates that the *shrunk-2D design* implemented with the baseline shrunk-2D flow estimates the parasitic of the M3D design incorrectly.

In order for the *shrunk-2D design* to accurately estimate the RC parasitics of the wires in the M3D design while using the dimensional scaling technique, the RC parasitic look-up tables should be updated by modifying the geometric and material parameters (i.e., the terms in the right-hand side in Equation (3)), so that the look-up tables reflect the shrunk wires to determine wire RC parasitics. One of the solutions for Equation (2) with the constraints in Equation (4) is described in the following equations.

$$\begin{aligned} \rho_{S2D} = \rho_{M3D}/2, \quad \epsilon_{S2D} = \epsilon_{M3D}, \\ h_{S2D} = h_{M3D}/\sqrt{2}, \quad t_{S2D} = t_{M3D}/\sqrt{2}, \\ d_{S2D,v} = d_{M3D,v}/\sqrt{2} \end{aligned} \tag{6}$$

In our new Parasitic Adjustment (PA) methodology, we use Equations (6) to generate an RC parasitic look-up table, which is fed to

Table 1: Parasitic error rate comparison between the baseline shrunk-2D vs. our parasitic adjustment (PA) in LDPC designs. Capacitance values are in pF and resistance in $M\Omega$.

parameters		shrunk-2D	M3D	ERR%
w/o PA	pin cap	137.6	137.6	0.0 %
	wire cap	250.6	202.3	23.9 %
	total cap	**388.2**	**339.9**	**14.2 %**
	resistance	**8.663**	**7.470**	**16.0 %**
w/ PA	pin cap	118.8	118.8	0.0 %
	wire cap	227.5	191.4	18.9 %
	total cap	**346.3**	**310.2**	**11.6 %**
	resistance	**5.039**	**4.833**	**4.3 %**

commercial EDA tools to calculate the RC parasitics of the wires in a design. With the adjusted RC parasitic look-up table, we implement a *shrunk-2D design*. It is important to note that the adjusted RC parasitic look-up table is used only for the *shrunk-2D design*, and the original RC parasitic look-up table is used to calculate wire parasitics of the final M3D design.

Table 1 compares the error rate of the parasitics of the *shrunk-2D design* with and without parasitic adjustment from the parasitics of the final M3D design. The table clearly shows that the parasitic adjustment decreases the error rate of the parasitics of the *shrunk-2D design* to 11.6% and 4.3% for the total capacitance and resistance of the design, respectively. Although the parasitic adjustment technique reduces the error rate of the parasitics compared to the baseline shrunk-2D flow, errors still remain because the cells in a *shrunk-2D design* are partitioned into the top and bottom tiers after implementing the *shrunk-2D design*, and the wires are re-routed with the partitioned design.

With the optimized *shrunk-2D design* implemented with the parasitic adjustment technique, the next two sub-sections describe the performance optimization methodologies while transforming the *shrunk-2D design* into the M3D design.

3.2 Path-based Tier Partitioning

As cells are not assigned to tiers in *shrunk-2D designs*, the timing closure in the *shrunk-2D design* does not take into account the routing resources necessary to connect cells in different tiers through MIVs (i.e., inter-tier routing overhead). The inter-tier routing overhead in the M3D designs makes the delay of the timing paths crossing the tiers longer than the *shrunk-2D design*. As the area-balanced min-cut partitioning algorithm in the baseline shrunk-2D design flow is timing-agnostic, dividing cells in critical timing paths into two tiers can degrade the performance of the M3D designs.

We employ a path-based tier partitioning algorithm [10] to prevent increasing the delay due to MIVs on the critical paths. This algorithm first pre-places all cells in a number of critical timing paths on a single tier to avoid the inter-tier routing overhead, and then performs the area-balanced min-cut partitioning with the remaining cells. The algorithm was originally devised to overcome inter-tier degradation of M3D designs, but also improves the performance of M3D designs by minimizing inter-tier routing overhead.

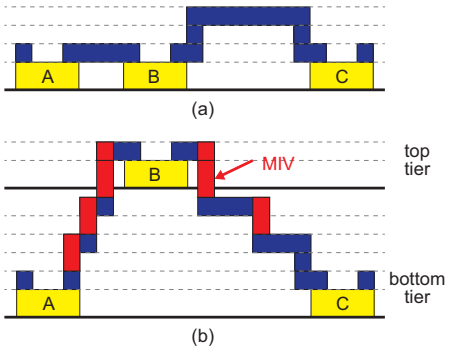

Figure 3: Inter-tier routing overhead (red) resulting from assigning cell B on the top tier while transforming (a) a *shrunk-2D design* to (b) the M3D design.

3.3 Clock Buffer Tier Partitioning

While the aforementioned two methodologies focus on reducing D_{path} in Equation (1), the slack of a timing path can also be improved by preventing negative T_{skew}. The following describes the clock skew of a timing path.

$$T_{skew} = L_{end} - L_{start}, \tag{7}$$

where L_{end} and L_{start} are the clock latency at the clock pin of the end and start point of the timing path. Increasing L_{end} may improve slack for a timing path, but it imposes a tighter timing constraint on the subsequent path, which uses the end point of the current timing path as its start point. Therefore, it is important to obtain almost uniform clock latencies, making T_{skew} nearly zero for all timing paths.

The inter-tier routing overhead induced by MIVs also negatively affects the quality of the clock tree of a design. Figure 4 illustrates the degraded clock skew in the M3D design due to a MIV insertion. The clock signal at the start point (cell C) and the end point (cell D) of the timing path are fed from clock buffers, cell A and cell B, respectively. In the *shrunk-2D design*, commercial EDA tools minimize T_{skew} in Figure 4 (a), maximizing the timing budget (i.e., first three terms in Equation (1)) of the timing paths. However, if cell A is placed on the top tier by the tier partitioner, inter-tier routing overheads are introduced to the clock path to cell C. This, in turn increases the clock latency at the clock pin of the start point (cell C) by α (as shown in Figure 4 (b)), reducing the timing budget of the timing path.

To prevent the clock skew degradation, we employ the methodology proposed in [1], which was originally used to minimize the clock power consumption of M3D designs. The methodology partitions clock cells onto a single tier, which helps minimize MIV utilization in the clock paths. However, placing all clock cells including flip-flops may cause significant area skew between tiers, increasing wire congestion in the tier with a higher cell area. The wire congestion increases the wire-length of the M3D designs, resulting in performance degradation. Therefore, we only fix clock buffers in a single tier, and let flip-flops be free to be partitioned in either tier, so that every clock path in the clock tree utilizes at most one MIV, minimizing the inter-tier routing overhead.

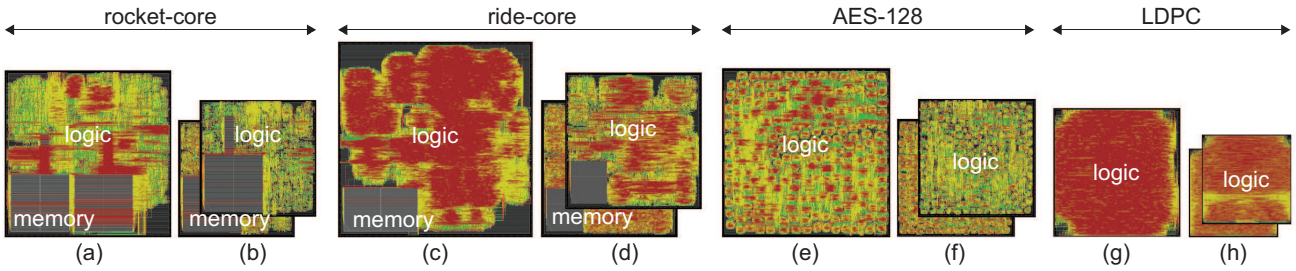

Figure 5: GDS layouts of (a) rocket-core 2D, (b) rocket-core M3D, (c) ride-core 2D, (d) ride-core M3D, (e) AES-128 2D, (f) AES-128 M3D, (g) LDPC 2D, and (h) LDPC M3D at their maximum achievable clock frequencies using NanGate 45nm PDK.

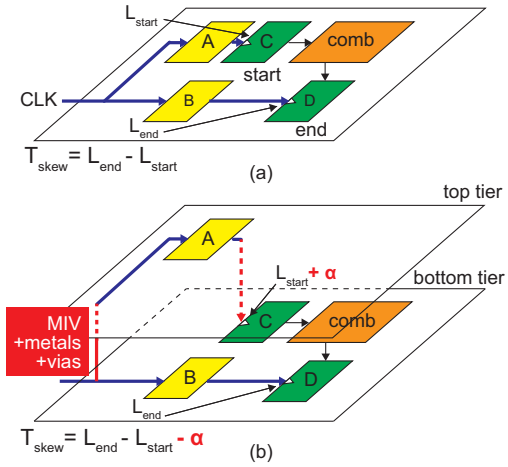

Figure 4: Impact of inter-tier connection on clock skew. (a) Clock tree in *shrunk-2D design*, (b) skew degradation due to inter-tier routing overhead (= MIV + metals + vias).

4 EXPERIMENTAL RESULTS

4.1 Experimental Setup

The NanGate FreePDK45 Open Cell Libraryis used for 2D and M3D implementations. Rocket-core and ride-core , which are based on the RISC-V instruction set architecture, and AES-128 and LDPC from OpenCores.org are used as test vehicles. The 2D and M3D designs of the four benchmarks are implemented sweeping the target frequencies in order to obtain their maximum frequencies. The footprint and floorplan of each design are customized to maximize performance without design rule check (DRC) violations, and kept constant during frequency sweeps. Figure 5 shows the die images of the 2D and M3D designs of the four benchmarks at their maximum frequencies.

4.2 M3D Performance Improvement Results

Table 2 compares the 2D and M3D designs with their maximum achievable clock frequencies. We observe the following:

- Footprint: As M3D designs place cells on two tiers, the footprint of the M3D designs are ~50% smaller than their 2D counterparts.
- Wire-length and std. cell area: The wire-length reduction of the critical timing path reaches 53.4%, which results from a smaller footprint, and hence, shorter distances between cells in the M3D designs. The reduced wire-length helps reduce wire-load of the timing path, which in turn decreases standard cell area, showing up to 35.1% reduction.
- Capacitance and resistance: The reduced wire-length decreases the wire capacitance and resistance of the critical timing path, while the reduced pin capacitance is attributed to the standard cell area savings of the timing path.
- Delay: The reduced wire parasitics help lower the RC delay of the nets in the critical timing path. Since the wire-loads at the output of the cells are reduced, the cell delays are also decreased, offering up to a 24.6% total delay reduction.
- Effective frequency: As the delay of the timing path is decreased, the effective frequency, the highest clock frequency at which the design can operate, is increased up to 15.6%.
- Total power: As M3D designs operate at higher effective frequencies, the total power consumption of the M3D designs is essentially higher than the 2D designs.
- Energy-delay product (EDP): Due to the improved performance, M3D designs deliver up to 16.2% EDP improvement compared to their 2D counterparts.

It is important to note that the slack of a timing path is determined not only by D_{path} but also by T_{CP}, which is smaller in the M3D designs. Therefore, a worse WNS in the M3D design does not necessarily mean that the timing closure is worse than its 2D counterpart.

Table 2 also shows that the performance improvement of ride-core is lower than other benchmarks, showing only 5.0% improvement. This is mainly attributed to the complexity of the timing paths in the design. Figure 6 shows the timing path wire-length distribution of the four benchmarks. Ride-core has a higher number of timing paths with large wire-length, which makes it difficult for EDA tools to close timing. In those designs, cells tend to be clustered to each other as shown in Figure 5 (c), providing less room for reducing wire-length and standard cell area of the critical timing path in the M3D designs.

Figure 7 shows the benefits of M3D designs on timing closure, comparing the slack distribution of timing paths in the LDPC 2D and M3D designs. The 2D design suffers from a large number of timing paths with near-zero and negative slack, whereas the timing paths of the M3D design show a larger number of higher positive slack.

Table 2: Maximum performance comparison of 2D vs. M3D. We use the performance improvement methodologies presented in Section 3. The percentage values in the M3D designs are w.r.t. their 2D counterparts. Area values are in in um^2, frequency in MHz, power in mW, EDP in $pJ \cdot ns$, length in um, capacitance in pF, resistance in Ω, and time in ns.

parameters	rocket-core			ride-core			AES-128			LDPC		
	2D	M3D		2D	M3D		2D	M3D		2D	M3D	
full-chip PPA												
footprint	0.384	0.188	(-51.0%)	0.523	0.256	(-52.4%)	0.390	0.191	(-51.0%)	0.228	0.112	(-51.0%)
eff freq	**783.3**	**905**	**(15.5%)**	**524**	**550.0**	**(5.0%)**	**1,388**	**1,585**	**(14.2%)**	**716**	**828**	**(15.6%)**
total power	156.7	175.2	(11.8%)	147.8	149.8	(1.4%)	177.0	207.9	(17.5%)	219.4	252.1	(14.9%)
EDP	**255.4**	**214.0**	**(-16.2%)**	**539.3**	**495.9**	**(-8.1%)**	**91.9**	**82.8**	**(-10.0%)**	**428.0**	**368.2**	**(-14.0%)**
critical timing path												
target freq	813	938	(15.4%)	530	562	(5.9%)	1,375	1,750	(27.3%)	750	875	(16.7%)
wire-length	900.4	419.3	(-53.4%)	865.1	816.9	(-5.6%)	91.9	95.5	(3.9%)	1,519.9	987.3	(-35.0%)
std. cell area	53.7	34.8	(-35.1%)	76.1	72.6	(-4.5%)	18.9	17.4	(-7.8%)	46.3	40.5	(-12.6%)
wire cap	0.110	0.063	(-42.9%)	0.167	0.138	(-16.9%)	0.021	0.022	(2.3%)	0.236	0.191	(-19.2%)
pin cap	0.250	0.098	(-60.7%)	0.263	0.283	(7.5%)	0.031	0.041	(32.7%)	0.169	0.125	(-26.1%)
resistance	4,234	3,242	(-23.4%)	5,287	5,438	(2.9%)	1,109	1,090	(-1.7%)	5,581	3,647	(-34.7%)
setup time	0.016	0.031	(92.6%)	0.015	0.019	(30.4%)	0.037	0.037	(0.5%)	0.000	0.028	(-)
clk skew	-0.011	-0.131	(1,104.6%)	0.071	-0.068	(-195.9%)	-0.039	-0.045	(14.8%)	-0.214	-0.068	(-68.1%)
delay	1.249	0.943	(-24.6%)	1.966	1.732	(-11.9%)	0.644	0.548	(-14.9%)	1.183	1.113	(-5.9%)
WNS	-0.046	-0.039	(-15.7%)	-0.024	-0.038	(58.3%)	0.007	-0.059	(-1,004.8%)	-0.063	-0.066	(3.6%)

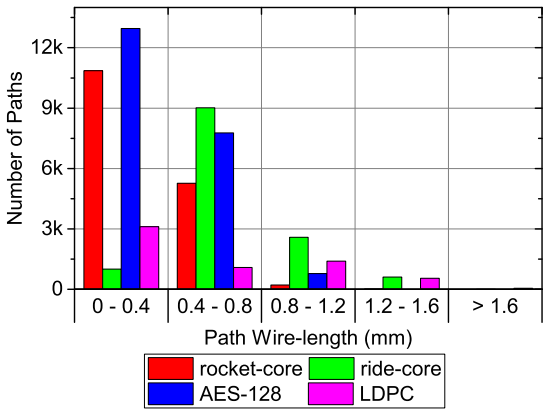

Figure 6: Timing path wire-length distribution of the benchmarks based on 2D designs

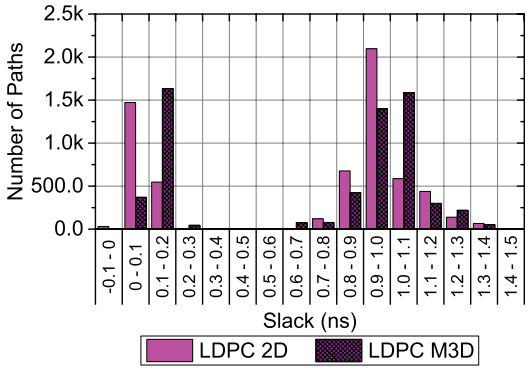

Figure 7: Slack distributions of 2D vs. M3D timing paths in LDPC designs at the maximum target frequency

The figure indicates that the reduced wire-length and standard cell area of the timing paths of M3D designs help ease timing closure, hence, offering performance improvement.

4.3 On Performance Improvement Methods

The impact of the performance improvement methodologies presented in Section 3 is shown in Table 3. In all benchmarks, the M3D designs implemented with parasitic adjustment outperform M3D designs using the baseline shrunk-2D design flow, showing up to 8.0% performance improvement. The baseline shrunk-2D fails to utilize enough number of buffers and drive-strength cells on the critical timing paths due to inaccurate RC parasitic estimation, reducing the performance improvement. The benefit of the adjusted parasitics is greater in wire-dominated circuits (e.g., LDPC) as wire parasitic estimation significantly affects the cell placement.

Table 3: Impact of the parasitic adjustment (PA), path-based tier partitioning (PTP), and clock buffer tier partitioning (CLK-FIX) on the performance of the M3D designs. Values are the performance improvement w.r.t. their 2D designs.

techniques	rocket-core	ride-core	AES-128	LDPC
base	10.2%	0.5%	11.5%	7.6%
+ PA	12.6%	3.6%	13.5%	**15.6%**
+ PTP	14.9%	**5.0%**	**14.2%**	10.7%
+ CLK-FIX	**15.5%**	4.4%	11.9%	10.1%

In the path-based tier partitioning method, pre-placing cells in a number of critical timing paths improves the performance of the M3D designs by reducing inter-tier routing overhead along their critical timing paths. However, in benchmarks with a high number of timing paths with near-zero slack like LDPC (as shown in Figure 7), pre-placing does not improve the performance. In those

benchmarks, the partitioning method improves the timing closure of the selected timing paths. However, the method makes it more difficult for the area-balanced min-cut partitioner to minimize the number of MIVs. This increases the delay of non-selected timing paths with near-zero slack, creating new critical timing paths.

The clock buffer tier partitioning technique offers a performance benefit to the rocket-core M3D design. The sequential cells occupy 19.1% of the total cells in the 2D rocket-core, which is a higher percentage compared to other benchmarks (ride-core: 9.3%, AES-128: 8.1%, and LDPC: 3.0%). As the clock tree is more complex in sequential cell dominated designs, their M3D designs benefit more from the methodology, decreasing inter-tier routing overhead. On the other hand, it negatively affects the performance of combinational cell dominated circuits since it is prone to increase MIVs, worsening the delay of the timing paths.

4.4 Impact on Energy

Energy consumed by a design increases as the performance of the design increases. Because the energy metric does not consider energy and performance of a design simultaneously, the EDP metric is used instead. [6]. The improved performance of M3D designs benefits the EDP of the designs offering up to 16.2% improvement as shown in Table 2. Moreover, as presented in [4], the wire-length reduction of M3D designs due to reduced footprints help reduce the standard cell area, which reduces wire capacitance and pin capacitance. This in turn, decreases power consumption of the design, which helps improve the EDP metric.

The parasitic adjustment technique further decreases the power consumption of M3D designs. Table 4 shows the iso-performance power comparison of the M3D designs implemented using the baseline shrunk-2D design flow (M3D w/ S2D-base) and using the shrunk-2D flow with parasitic adjustment (M3D w/ S2D-PA). Because the RC parasitics of the wires in the *shrunk-2D designs* more accurately depict the final M3D designs with parasitic adjustment, lower number of buffers and drive-strength cells are utilized, resulting in lower internal and pin capacitance switching power.

5 OBSERVATIONS

Our observations and guidelines to improve the performance of M3D designs are summarized as follows:

- M3D offers performance improvement over 2D designs, showing up to 15.6% higher maximum clock frequency.
- The parasitic adjustment method helps estimate the RC parasitics of the M3D designs more accurately and insert minimal buffers in the right place. The method presents benefits not only in performance, but also power savings of M3D designs.
- Pre-placing cells in a number of critical timing paths in a single tier helps improve the performance of M3D designs by reducing inter-tier routing overhead in critical timing paths.
- Partitioning clock buffers into a single tier minimizes the number of MIVs in the clock tree. This reduces the clock skew of timing paths, offering increased timing budget in sequential cell dominated circuits.
- M3D designs offer EDP improvements over 2D designs because of the increased performance as well as the power benefit of M3D designs, which comes from the parasitic adjustment technique which further reduces standard cell area.

Table 4: Iso-performance power comparison of the M3D designs implemented with baseline shrunk-2D design flow (S2D-base) and parasitic adjusted shrunk-2D design flow (S2D-PA). The percentage values in the M3D designs are w.r.t. their 2D counterparts. Power values are in mW.

designs	paramaters	M3D /w S2D-base		M3D /w S2D-PA	
rocket -core	sw. pwr	41.3	(-9.1%)	40.1	(-11.8%)
	int. pwr	112.4	(-1.4%)	110.4	(-3.1%)
	lkg. pwr	3.0	(-2.7%)	2.9	(-6.8%)
	tot. pwr	**156.7**	**(-3.5%)**	**153.4**	**(-5.6%)**
ride -core	sw. pwr	53.6	(-16.7%)	49.6	(-23.0%)
	int. pwr	78.1	(-3.4%)	72.8	(-10.0%)
	lkg. pwr	4.1	(-4.8%)	3.7	(-14.2%)
	tot. pwr	**135.9**	**(-9.2%)**	**126.1**	**(-15.7%)**
AES -128	sw. pwr	64.5	(-7.6%)	61.2	(-12.2%)
	int. pwr	100.2	(-1.7%)	95.3	(-6.6%)
	lkg. pwr	3.6	(-3.0%)	3.2	(-13.5%)
	tot. pwr	**168.3**	**(-4.1%)**	**159.7**	**(-9.0%)**
LDPC	sw. pwr	79.0	(-41.9%)	71.4	(-47.5%)
	int. pwr	71.6	(-21.4%)	63.5	(-30.2%)
	lkg. pwr	2.1	(-22.2%)	1.8	(-32.7%)
	tot. pwr	**152.6**	**(-33.5%)**	**136.7**	**(-40.5%)**

6 CONCLUSION

In this paper, we, for the first time, performed comprehensive studies on the factors that impact the performance benefit of M3D designs, and presented methodologies to further improve the performance. We found that the presented methodologies help raise the maximum achievable clock frequencies of M3D designs and supported our findings with an in-depth analysis. This work demonstrates the road to high-performance 3D ICs, showing the potential of M3D as a solution for 2D device scaling challenges.

REFERENCES

[1] K. Acharya et al. 2016. Monolithic 3D IC design: Power, performance, and area impact at 7nm. In *Proc. Int. Symp. on Quality Electronic Design.*

[2] S. Bobba et al. 2011. CELONCEL: Effective design technique for 3-D monolithic integration targeting high performance integrated circuits. In *Proc. Asia and South Pacific Design Automation Conf.*

[3] K. Chang et al. 2016. Cascade2D: A design-aware partitioning approach to monolithic 3D IC with 2D commercial tools. In *Proc. IEEE Int. Conf. on Computer-Aided Design.*

[4] K. Chang et al. 2016. Match-making for Monolithic 3D IC: Finding the Right Technology Node. In *Proc. ACM Design Automation Conf.*

[5] J. Cong et al. 2007. Thermal-Aware 3D IC Placement Via Transformation. In *Proc. Asia and South Pacific Design Automation Conf.*

[6] R. Gonzalez and M. Horowitz. 1996. Energy Dissipation in General Purpose Microprocessors. *IEEE Journal of Solid-State Circuits* (1996).

[7] G. Katti et al. 2010. Through-Silicon-Via Capacitance Reduction Technique to Benefit 3-D IC Performance. *IEEE Electron Device Letters* (2010).

[8] B.W. Ku et al. 2005. Physical Design Solutions to Tackle FEOL/BEOL Degradation in Gate-level Monolithic 3D ICs. In *Proc. Int. Symp. on Low Power Electronics and Design.*

[9] S. Panth et al. 2014. Design and CAD Methodologies for Low Power Gate-level Monolithic 3D ICs. In *Proc. Int. Symp. on Low Power Electronics and Design.*

[10] S. Samal et al. 2016. Tier Partitioning Strategy to Mitigate BEOL Degradation and Cost Issues in Monolithic 3D ICs. In *Proc. IEEE Int. Conf. on Computer-Aided Design.*