Four-tier Monolithic 3D ICs: Tier Partitioning Methodology and Power Benefit Study

Kwang Min Kim[1], Saurabh Sinha[2], Brian Cline[2], Greg Yeric[2], and Sung Kyu Lim[1]
[1]School of ECE, Georgia Institute of Technology, Atlanta, GA, USA
[2]ARM Inc., Austin, TX, USA
kwangmin@gatech.edu, limsk@ece.gatech.edu

ABSTRACT

Monolithic 3D IC is an emerging technology to continuously satisfy demands for power reduction under challenges posed by traditional device scaling. In this paper, for the first time, we study power benefits of 4-tier monolithic 3D ICs compared with 2-tier monolithic 3D and 2D ICs. We present a tier partitioning methodology that significantly extends the capability of a state-of-the-art flow built for 2-tier monolithic 3D ICs. We develop two complete RTL-to-GDSII design flows to achieve this goal and offer quantitative comparisons. In addition, we study impacts of inter-tier via usage on 2-tier and 4-tier monolithic 3D ICs. Our experiments show that poorly controlled inter-tier via usage results in up to 6.05% degradation in total power savings. Thus, we develop an effective strategy to achieve inter-tier via configurations to optimize power metrics. Experiments show that 4-tier monolithic 3D ICs outperform 2-tier and 2D IC by 15% and 50% in terms of power and 25% and 75% in area under the same performance.

CCS Concepts

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

Keywords

Monolithic 3D; Partitioning; Placement

1. INTRODUCTION

In modern VLSI, power consumption, performance, and area (PPA) have been main targets for the next generation technology. In the past, these efforts had been achieved by device scaling and innovations in technology such as such as high-k metal gates, FD-SOI, and FinFET. However, as device scaling reaches its physical limits, power management becomes more difficult to resolve. In addition, interconnect scaling faces increased RC product problems [6]. On the other hand, necessity for low power devices has been rising due to increasing demands for mobile devices and emerging IoT era.

Three-dimensional integrated circuit (3D IC) is an emerging technology to continuously achieve power reduction. Instead of fabricating a large chip, 3D ICs integrate dies vertically on the top of others. If dies are fabricated separately and bonded together, through-silicon-vias (TSVs) are used as 3D connections between dies. In TSV-based technologies, there are challenges in bonding such as die alignment and precision issues. On the other hand, monolithic 3D ICs (M3D ICs) have been emerging as another solution. In M3D ICs, each die grows on the top of others, thereby eliminating bonding challenges posed by TSV-based 3D ICs. Another significant advantage of M3D ICs is nano-scale monolithic inter-tier vias (MIVs) [1]. Unlike micron-scale TSVs, MIVs introduce negligible area overhead and allow ultra high density integration.

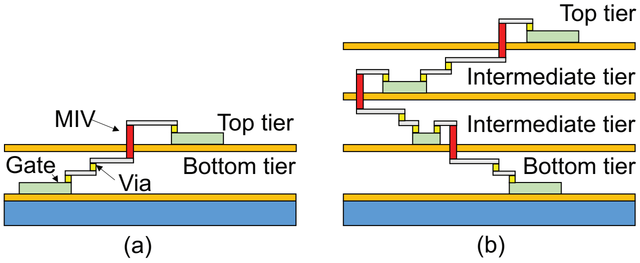


Figure 1: Side-views of (a) 2-tier and (b) 4-tier monolithic 3D ICs

Previous studies have shown that 3D ICs effectively reduce power consumption. In [4], authors achieved 16.08% power reduction on a CPU core using 2-tier M3D IC design. In [3], an average power reduction of 16.8% for high-performance devices and 14.3% for low standby power devices is achieved using 2-tier M3D designs. While efforts to explore power benefits of M3D ICs continue, recent studies have been limited to only 2-tier M3D IC designs. Then, how much more power benefits can we achieve from multi-tier M3D IC designs? Motivated by this question and advantages of M3D technology, we examine the opportunities to further extend existing M3D methodology and its advantages using multiple tiers. Figure 1 demonstrates how MIVs are used to connect multiple tiers. Between two consecutive tiers, MIVs connect the topmost routing layers of bottom die and the bottom-most routing layers of top die.

Due to the current unavailability of 3D EDA tools, we develop a methodology to build commercial quality M3D designs to design and investigate benefits of multi-tier M3D ICs. In recent studies, efforts have been made to extend the capability of commercial 2D tools to build M3D designs. The authors of [2] have built a folding-based technique while the authors of [4] have taken a scaling-based approach. However, these approaches are limited to only 2-tier M3D ICs. Therefore, to analyze benefits of multi-tier M3D ICs, a new comprehensive design methodology must be developed. In this work, our contributions are as follows: (1) We develop a re-

ISLPED '16, August 08-10, 2016, San Francisco Airport, CA, USA
© 2016 ACM. ISBN 978-1-4503-4185-1/16/08. . . $15.00
DOI: http://dx.doi.org/10.1145/2934583.2934623

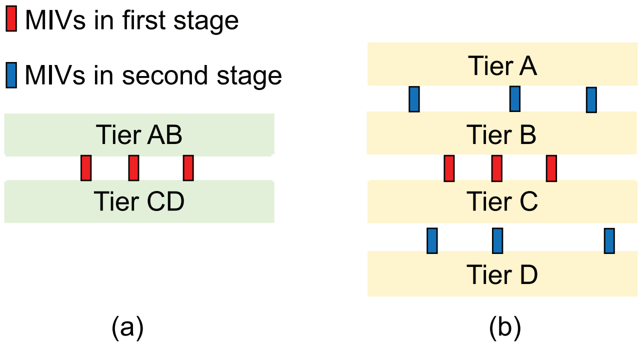


Figure 2: Recursive monolithic inter-tier via (MIV) insertion: (a) 2-tier stage, and (b) 4-tier stage

liable design methodology to build 4-tier M3D ICs using existing commercial 2D EDA tools. To the best of authors' knowledge, this work is the first attempt to build a complete RTL-to-GDSII flow for 4-tier M3D ICs. (2) We explore comprehensive benefits introduced by 4-tier M3D ICs compared to 2-tier M3D and 2D counterparts. Our study shows that 4-tier M3D designs enhance benefits offered by 2-tier M3D ICs by at least 50%. (3) We study impacts of MIV usage on M3D benefits and devise an effective strategy to obtain power-optimized 4-tier M3D ICs. Our findings show that if MIVs are poorly managed, overall M3D benefits degrade significantly.

2. DESIGN METHODOLOGY

We develop two RTL-to-GDSII 4-tier M3D methodologies inspired by shrunk2D flow [4]. While we preserve the general approach taken by shrunk2D in 2-tier M3D design, we heavily modify and extend it to build 4-tier M3D designs effectively. While our two methodologies show differences in their strategies, both adopt the shrunk2D flow in a recursive fashion. Figure 2 illustrates how this works. In both methods, we first create a 2-tier M3D from an initial 2D design. For simplicity, we designate top tier as "Tier AB" and bottom tier as "Tier CD." Then we partition Tier AB into 2 tiers, forming top two tiers of 4-tier M3D design. We name these tiers as "Tier A" and "Tier B." Likewise, we partition Tier CD to form "Tier C" and "Tier D." Hence, we have a total of four tiers: Tiers A, B, C, and D. Note that the MIVs placed during the first stage (Tier AB – Tier CD) become MIVs between Tiers B and C.

As introduced in original shrunk2D flow, we use Cadence Encounter to insert MIVs between consecutive tiers. However, since Encounter only supports a maximum of 15 metal layers for routing, stacking all four tiers together and determining MIVs in a single run is not possible. Thus, we insert MIVs connecting intermediate tiers in the first partitioning stage. While preserving these MIVs, we insert MIVs connecting top two tiers and bottom two tiers when we perform recursive partitioning in the second stage.

2.1 Method 1

The first method builds a 2-tier design initially and recursively build 2-tier designs again to create final 4-tier design. As shown in Figure 3 (a) and 4, this flow is divided into two stages: (1) initial 2-tier M3D design and (2) recursive 4-tier M3D design. In (1), we first build a scaled 2D design on a footprint that is 50% less than the 2D footprint by using standard cells and metal layers whose X-Y dimensions are scaled down by a factor of $1/\sqrt{2}$ (Step a-1). This scaling factor reflects theoretical 29.3% reduction in half-perimeter wire length (HPWL) and 50% reduction in area for 2-tier M3D compared with 2D design. At this point, we run all required de-

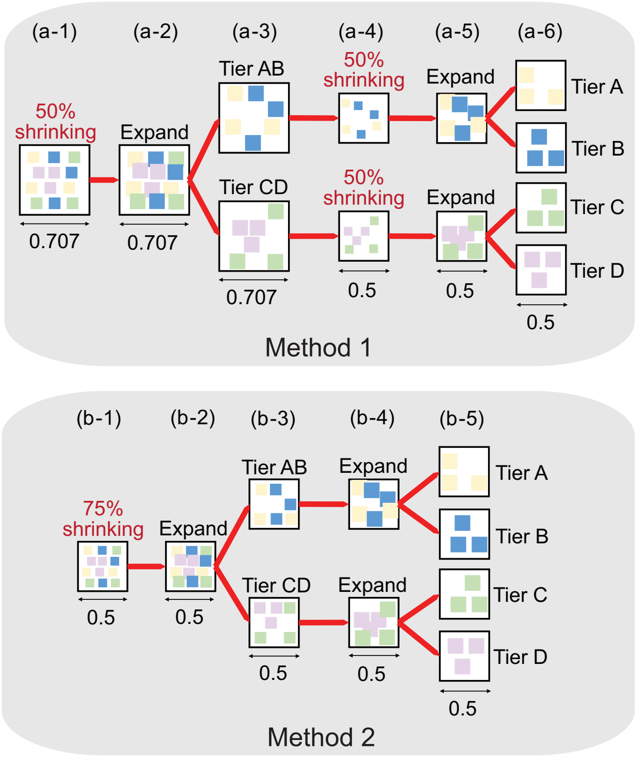


Figure 3: Proposed 4-tier design methodologies: (a) Method 1 based on 50% shrinking, and (b) Method 2 based on 75% shrinking. Scaling factors are with respect to 2D (1.0).

sign stages such as placement, post-placement optimization, clock tree synthesis (CTS), routing, and post-routing optimization. Our design at this stage takes a holistic view of 2-tier M3D design and inserts a minimum number of buffers required to meet timing. Once this scaled 2D design is built, cells in this scaled down 2D design are expanded back to original sizes, causing overlaps (a-2). Then, cells are split into Tier AB and Tier CD using a partitioner (a-3).

Once partitioned, cell placements are legalized to remove overlaps. Next step is to insert MIVs between Tier AB and Tier CD. Initially, these MIVs connect the bottom routing layer of Tier AB and top routing layer of Tier CD. When Tier AB and Tier CD are split in later stage, these MIVs propagate to Tier B and C and connect the bottom routing layer of Tier B and top routing layer of Tier C as shown in Figure 2. We pre-insert MIVs connecting intermediate tiers due to limited routing capability of current EDA tools.

After MIVs are inserted, we perform tier-wise trial routing to extract estimated parasitics of each tier. Then, timing constraints for each tier are extracted from Synopsys PrimeTime using parasitics, tier-wise netlists, and top-level netlist. Next, we perform final tier-wise timing-driven routing using these timing constraints and extract sign-off parasitics. Before proceeding into the second stage, we extract tier-wise timing constraints from final routing results from previous stage using PrimeTime.

In (2), we perform scaled 2D designs for Tier AB and Tier CD separately using footprint that is 50% less than the current footprint using tier-wise netlists and updated timing constraints. This new footprint is 75% less than the 2D footprint and used for final 4-tier M3D designs. As in the previous stage, standard cells, metal layers, pins, and MIVs are all scaled down again by a factor of $1/\sqrt{2}$ (a-4). Locations of MIVs are also scaled down with respect to reduced

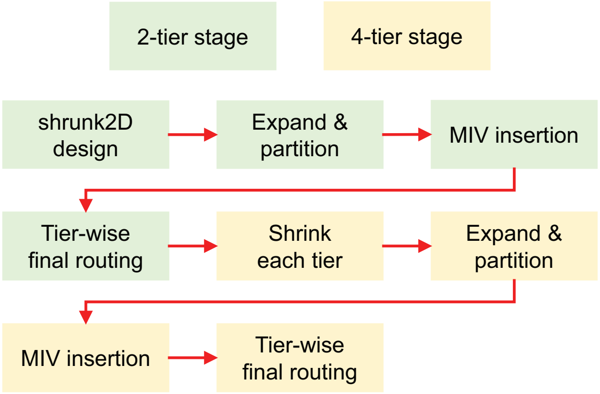


Figure 4: Overall flow of Method 1

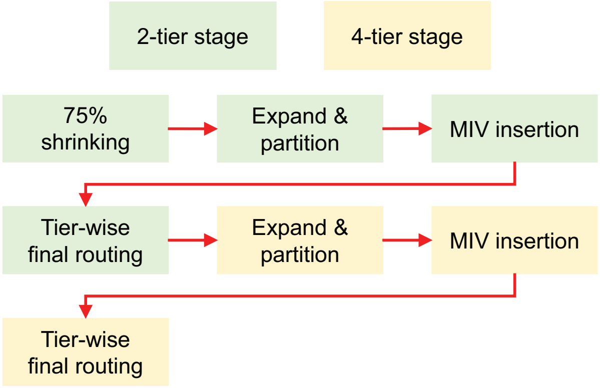


Figure 5: Overall flow of Method 2

Table 1: Benchmark settings

Circuit	Parameter	2D	2-tier M3D	4-tier M3D
LDPC	Clock period	0.9ns	0.9ns	0.9ns
	Footprint (μm^2)	800x800	566x566	400x400
ECG	Clock period	0.6ns	0.6ns	0.6ns
	Footprint (μm^2)	500x500	354x354	250x250
RCA_16	Clock period	0.5ns	0.5ns	0.5ns
	Footprint (μm^2)	650x650	456x456	325x325
TMU	Clock period	0.9ns	0.9ns	0.9ns
	Footprint (μm^2)	950x950	672x672	475x475

footprint to ensure that their relative locations with respect to the entire footprint are consistent. The motivation for scaling down Tier AB and Tier CD is to reflect further 29.3% HPWL reduction and 50% area reduction gained by splitting each tier of 2-tier M3D design into two. Overall, we assume theoretical 50.0% reduction in HPWL and 75% reduction in area compared with 2D counterpart. For each scaled Tier AB and Tier CD, we perform placement, post-placement optimization, routing, and post-routing optimization.

Cells and metal layers are then expanded in Tier AB. At this point, pre-inserted MIVs are also restored back to their original sizes (a-5). Tier AB is then partitioned into Tier A and Tier B, forming the top two tiers for 4-tier design (a-6). To properly connect intermediate two tiers, MIVs in Tier AB from the first stage must be partitioned and fixed into Tier B. Then, we legalize cells to remove placement overlaps and insert MIVs between Tier A and Tier B. Finally we perform trial routing, estimated parasitics extraction, timing-driven routing, and sign-off parasitic extraction. Same process is repeated to build Tier C and Tier D from Tier CD. As final products, one top-level and four tier-wise netlists, and sign-off parasitics are created along with DRC-clean GDSII layouts. Finally, we perform timing and power simulation in PrimeTime.

2.2 Method 2

The second method further exploits the advantage of scaling approach in shrunk2D flow. As shown in Figure 3 (b) and 5, the initial design is created using a scaling factor of 1/2 instead of $1/\sqrt{2}$. We start with first stage 2D design with a scaling factor of 1/2 for both standard cells and metal layers to reflect theoretical 50% HPWL and 75% area reduction. This 2D design grasps entire 3D timing path and inserts minimum number of buffers required to meet timing. This new scaled shrunk2D is called “75% shrinking.”

Similar to Method 1, Method 2 also consists of two design stages: (1) scaled 2-tier M3D design and (2) recursive 4-tier M3D design. In (1), we create 2D design on a footprint whose area is reduced by 75% from the 2D footprint (b-1). This footprint is used throughout all design stages in this method. As in Method 1, we perform placement, routing, CTS, and post-routing optimization at this design stage. We now expand cells and metal layers in each tier from its scaled dimensions (b-2). However, at this point, we do not restore them back to their original dimensions yet. Instead, their X-Y dimensions are expanded by a factor of $\sqrt{2}$. The reason for this is similar to that of Method 1. Since each of Tier AB and Tier BC will be recursively split into 2 tiers in the later stage, we must assume 29.3% reduction in wirelength at this stage.

This design is partitioned into Tier AB and Tier CD (b-3). After expanding and partitioning steps, cells are legalized to remove overlaps. MIVs are inserted between Tier AB and Tier CD. Like Method 1, these MIVs propagate to Tier B and Tier C with their locations fixed on corresponding tiers. Then we perform same design steps taken in Method 1: trial routing, estimated parasitic extraction, timing constraints extraction, timing-driven routing, sign-off parasitic extraction, and updated timing constraints extraction.

In (2), we perform 2D design stages again for Tier AB and Tier CD using the updated timing constraints. In contrast to Method 1, scaling down cells, metal layers, pins, MIVs, and footprint are not required due to our expansion strategy in previous step. From the results of the first stage, cells and metal layer dimensions in Tier AB and Tier CD already reflect 29.3% wirelength reduction. We perform placement, post-placement optimization, routing, and post-routing optimization. Then, Tiers A, B, C, and D are built from Tier AB and Tier CD in the same manner as in Method 1 (b-4 and b-5). Final products of this method are also one top-level and four tier-wise netlists, sign-off parasitics, and GDSII layouts.

3. POWER BENEFIT STUDY

In this section, we compare our proposed methodologies and explore the power benefits of sign-off 4-tier M3D designs over 2D and 2-tier M3D design using a 28 nm technology library and a set of four benchmark circuits from opencores.org. We perform iso-performance comparison at the fastest operating frequencies for 2D designs. Our design settings for benchmarks are shown in Table 1. 2-tier M3D designs are obtained from the original shrunk2D flow. Sample GDSII layouts of 2D, 2-tier M3D, and 4-tier M3D designs are shown in Figure 6. In this study, we assume same MIV parameters as authors of [4] have used. The diameter of MIV is assumed to be 100 nm with resistance of 2 Ω and capacitance of 0.1 fF.

3.1 Method 1 vs. Method 2 Comparison

Table 2 shows a comparison between initial 2D design stages of

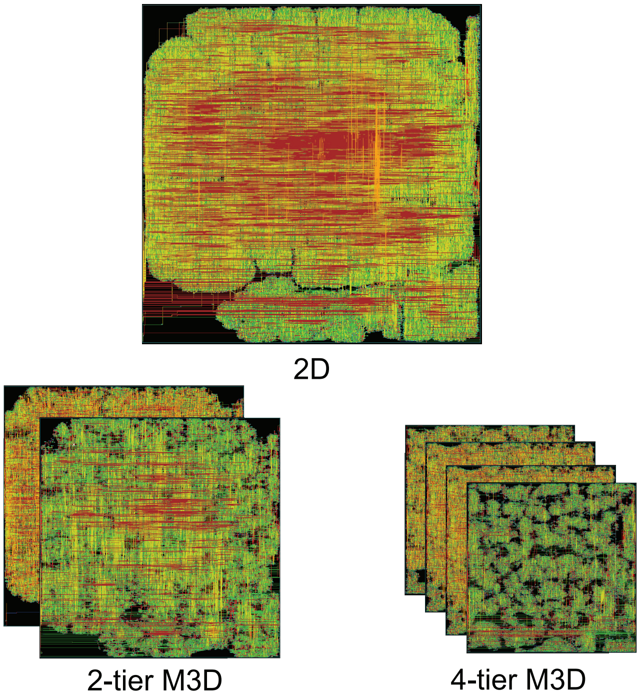


Figure 6: GDSII die shots of (a) 2D (950x950um), (b) 2-tier M3D (672x672um), (c) 4-tier M3D (475x475um) implementation of TMU (Method 2)

Method 1 and 2. As discussed, Method 1 uses a factor of $1/\sqrt{2}$ to scale down X-Y dimensions of cells, metal layers, and footprints while Method 2 uses a factor of 1/2. Since Method 2 assumes 25% smaller area and 20.7% shorter wirelength compared with Method 1, its initial 2D designs are better optimized for final 4-tier M3D designs. Initial designs of Method 2 use less number of cells and buffers due to shorter wirelength and smaller area.

In both methods, Tiers AB and CD are optimized tier-wise using updated timing constraints. Since Method 1 only assumes conditions of 2-tier M3D designs in its initial stage designs, each tier must be optimized to reflect further reduced wirelength and area that matches those of final 4-tier M3D designs. Depending on partitioning solutions, number of MIVs, routing congestion, and parasitics, results of these optimization steps improve or worsen the initial design as shown in Table 2. However, despite possible degradation from initial design quality for both Method 1 and Method 2, tuning each design using updated timing constraints is a required step to enforce timing closure in 4-tier M3D designs. As shown in Table 3, the effects of design qualities translate to 4-tier wirelength and power benefits. Overall, we observe that Method 2 outperforms Method 1 by using a scaling factor that exploits wirelength benefits of 4-tier M3D more accurately. Therefore, we adopt Method 2 for subsequent experimental studies in this work.

3.2 1-tier vs. 2-tier vs. 4-tier Comparison

The summary of our designs is presented in Table 4. By having 4-tier M3D design, we reduce the total wirelength by 34.1% and cell count by 14.5% on average. Compared to 2-tier M3D, we achieve additional 12.7% and 5.3% reduction in wirelength and cell count respectively. The wirelength reduction comes from the 75% less and 25% less area of 4-tier M3Ds compared with 2Ds and 2-tier M3Ds. Due to shorter wirelength and smaller footprint, 4-tier M3D achieves an average of additional 18.0% savings in buffer usage compared with 2-tier M3Ds. Reduced buffer usage translates to an average of 14.5% reduction in total gate count. Overall, we observe that by advancing to 4-tier M3D, we can consistently enhance benefits already introduced by 2-tier M3Ds by at least 50%.

Table 2: Method 1 vs. Method 2 area and cell count comparison. The number of cells and buffers is normalized to the 2-tier design of Method 1.

Circuit	Parameter	Method 1		Method 2	
		2-tier	4-tier	2-tier	4-tier
LDPC	Footprint	0.5	0.25	0.25	0.25
	# Cells	1.0	1.078	0.824	0.896
	# Buffers	1.0	1.036	0.666	0.731
ECG	Footprint	0.5	0.25	0.25	0.25
	# Cells	1.0	0.985	0.957	0.955
	# Buffers	1.0	0.917	0.835	0.825
RCA_16	Footprint	0.5	0.25	0.25	0.25
	# Cells	1.0	1.074	0.948	0.989
	# Buffers	1.0	1.133	0.738	0.860
TMU	Footprint	0.5	0.25	0.25	0.25
	# Cells	1.0	1.008	0.919	0.926
	# Buffers	1.0	1.030	0.645	0.675

Table 3: Method 1 vs. Method 2 wirelength and power comparison. We show normalized values of Method 2 with respect to Method 1.

Circuit	LDPC	ECG	RCA_16	TMU
Wirelength	0.995	1.004	0.963	0.947
Switching power	0.946	0.975	0.951	0.923
Internal power	0.934	0.997	0.976	0.961
Leakage power	0.667	0.963	0.956	0.781
Total power	**0.935**	**0.981**	**0.960**	**0.923**

Table 5 shows summary of power benefits of 2-tier and 4-tier M3D designs. The wirelength reductions achieved in 4-tier M3Ds contribute to an average saving of 30.7% in switching power compared with their 2D counterparts. Due to reduction in cell count and buffer usage, internal power and leakage power are reduced by 20.4% and 27.0% on average. Overall, we achieve an average of 26.8% total power reduction. Again, we observe similar advantage of advancing to 4-tier from 2-tier M3D by enhancing power reductions by at least 50%.

4. DESIGN STRATEGY

In the original shrunk2D flow and methodologies developed in this study, MIV usage can be controlled in two ways. First, multiple MIVs can be inserted per each net. In [4], authors show additional reduction in wirelength and power when multiple MIVs were used. Second, overall MIV usage is controlled by varying partitioning bin sizes as discussed in [5]. In this previous study, authors also showed M3D benefits are not necessarily maximized by using the minimum of MIVs. In this section, we examine impacts of MIV usage on 2-tier and 4-tier M3D ICs and present an effective design strategy to obtain optimized 4-tier M3D designs.

4.1 Impact of MIV Usage

To investigate impacts of MIV configurations on M3D benefits, we first conduct this study in 2-tier M3D. Number of MIVs is controlled by varying the partitioning bin sizes similarly to [5]. Figure 7 shows the impact of MIV usage on wirelength, switching power, and total power benefits. Due to limited space, corresponding bin size for each MIV configuration is not shown in this work. Since wirelength saving translates to both switching and total power sav-

Table 4: Comparison of design metrics. The unit for silicon area is μm^2, and wirelength is m. We use Method 2.

Circuit	Parameter	2D	2-tier M3D	4-tier M3D
LDPC	Si. Area	215,227	179,537 (-16.6%)	158,643 (-26.3%)
	WL	6.245	3.939 (-36.9%)	2.993 (-52.1%)
	# Cells	76,201	59,801 (-21.5%)	53,553(-29.7%)
	# Buffers	47,439	30,851 (-35.0%)	22,553 (-52.5%)
	# MIVs	-	14,142	32,889
ECG	Si. Area	162,983	150,595 (-7.6%)	144,124 (-11.6%)
	WL	1.566	1.217 (-22.2%)	1.014 (-35.2%)
	# Cells	57,182	52,814 (-7.6%)	50,425(-11.8%)
	# Buffers	17,985	13,646 (-24.1%)	11,264 (-37.4%)
	# MIVs	-	10,132	28,713
RCA_16	Si. Area	284,104	142,342 (-2.2%)	269,799 (-5.0%)
	WL	1.448	1.256 (-13.3%)	1.155 (-20.2%)
	# Cells	68,592	66,500 (-3.0%)	65,800(-4.1%)
	# Buffers	15,145	13,162 (-13.1%)	11,318 (-25.3%)
	# MIVs	-	15,423	42,516
TMU	Si. Area	559,411	543,615 (-2.8%)	504,354 (-9.8%)
	WL	3.732	3.2405 (-13.2%)	2.654 (-28.9%)
	# Cells	152,690	145,637 (-4.6%)	133,778 (-12.4%)
	# Buffers	40,761	34,426 (-15.5%)	22,571 (-44.6%)
	# MIVs	-	37,747	78,984

Table 5: Comparison of power metrics. All power values are in mW. We use Method 2.

Circuit	Parameter	2D	2-tier M3D	4-tier M3D
LDPC	Switching power	358.6	225.3 (-37.2%)	169.5 (-52.1%)
	Internal power	46.7	38.7 (-17.1%)	32.4 (-52.7%)
	Leakage power	8.83	6.59 (-25.4%)	5.18 (-41.3%)
	Total power	414.4	**270.7 (-34.5%)**	**207.1 (-50.0%)**
ECG	Switching power	88.5	71.2 (-19.5%)	62.0 (-29.9%)
	Internal power	34.6	32.7 (-5.5%)	31.7 (-8.4%)
	Leakage power	5.75	4.46 (-22.4%)	4.04 (-29.7%)
	Total power	128.9	**108.4 (-15.9%)**	**97.8 (-24.1%)**
RCA_16	Switching power	42.8	38.4 (-10.3%)	34.7 (-18.9%)
	Internal power	28.2	26.9 (-4.6%)	24.4 (-13.5%)
	Leakage power	8.84	8.47 (-4.1%)	7.81 (-11.6%)
	Total power	79.8	**73.7 (-7.6%)**	**66.9 (-16.2%)**
TMU	Switching power	127.7	115.8 (-9.3%)	100.0 (-21.7%)
	Internal power	87.2	86.0 (-1.4%)	81.0 (-7.1%)
	Leakage power	23.4	22.4 (-4.3%)	17.5 (-25.2%)
	Total power	239.3	**224.1 (-6.4%)**	**198.5 (-17.0%)**

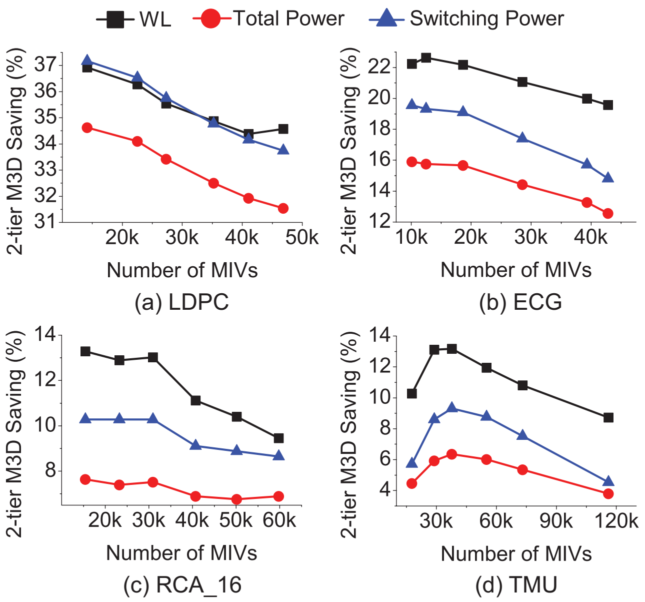
ings, targeting for minimizing wirelength improves power savings. Our results show that reducing MIV usage tends to improve M3D benefits in general. However, minimal usage of MIV does not always guarantee the best design. The best MIV configurations and corresponding bin sizes are heuristics that depend on benchmarks.

Using similar strategy, we explore the impact of MIV usage in 4-tier M3D ICs. We first control the number of MIVs inserted between Tier B and Tier C by varying partitioning bin sizes in the first 2-tier split from "75% shrinking" design. Then, we control the number of MIVs inserted between Tier A and Tier B and between Tier C and Tier D by varying bin sizes in the following 4-tier split.

In 4-tier case, we observe similar impact of MIVs. Reducing MIV usage improves 4-tier benefits in general. However, as shown in Figure 8, impact of MIV configurations on total power benefits is more severe in 4-tier designs. If MIVs are poorly controlled, total power benefits reduce as much as 6.05% in the worst case.

4.2 Design Complexity

Finding M3D IC design that maximizes its benefits requires a wide exploration of different configuration of MIVs and partitioning bin sizes since minimal MIV usage does not guarantee best design solution. Table 6 shows a summary of partitioning bin sizes used for best 2-tier and 4-tier M3D designs found in Section 3. These designs are obtained by performing extensive trials on combinations of bin sizes. There is weak correlation between best bin sizes used for 2-tier and 4-tier designs due to difference between scaling factors and footprints used in shrunk2D and "75% shrinking." Hence, optimizing 2-tier and 4-tier M3D designs must be performed independent of each other. In 2-tier M3D designs, complexity of varying bin sizes is $\mathcal{O}(n)$ in its nature. In 4-tier M3D design that requires two partitioning stages, its complexity is $\mathcal{O}(n^2)$.



(a) LDPC (b) ECG (c) RCA_16 (d) TMU

Figure 7: Impact of MIV usage on 2-tier wirelength, switching power, and total power saving

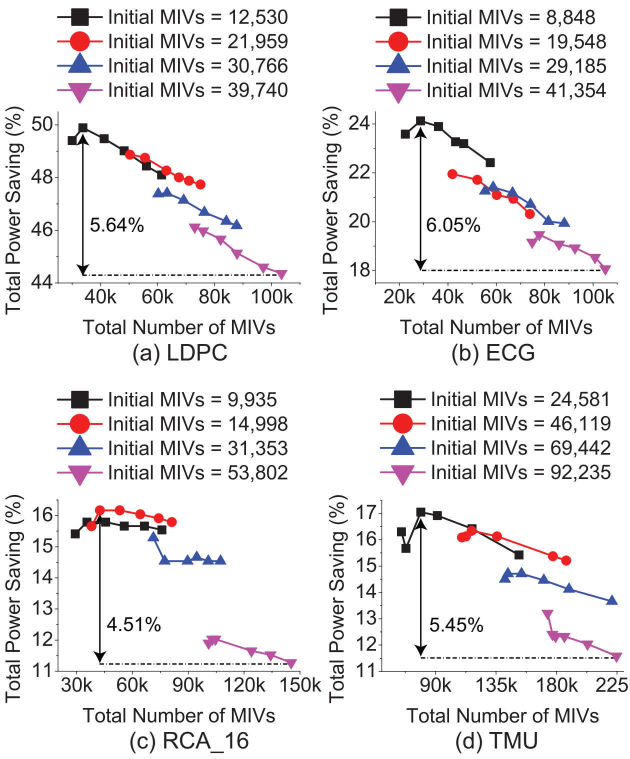
Table 6: Partitioning bin sizes used for the best 2-tier and 4-tier M3D designs. Bin sizes are measured in μm.

Circuit	2-tier M3D	4-tier M3D	
	Bin size	Initial bin size	4-tier bin size
LDPC	113	80	80
ECG	50	35	12.5
RCA_16	91	16	16
TMU	67	47.5	15

For each partitioning solution, we use Cadence Encounter to obtain sign-off quality designs and Synopsys PrimeTime to measure power and timing. In 2-tier case, this $\mathcal{O}(n)$ process is unavoidable yet manageable. However, in 4-tier case, this $\mathcal{O}(n^2)$ process becomes too repetitive and time-consuming. For instance, after partitioning initial shrunk2D design, each 2-tier M3D design requires 16 minutes to one hour in CPU time. In 4-tier case, each design requires 40 minutes to 2 hours. Thus, we are motivated to reduce the complexity and effort to search for optimized 4-tier M3D solutions.

4.3 Strategy to Optimize 4-tier M3D Designs

In this section, we present an efficient design strategy to obtain a quality power-optimized 4-tier M3D design while reducing iterative efforts. Our approach is based on the relationship between cutsize among partitions and number of MIVs. As shown in Figure 9, cutsize and number of MIVs exhibit strong dependency between them since both metrics describe number of inter-tier connections. Hence, although MIV usage is controlled by varying partitioning bin sizes, it is also equivalent to controlling cutsizes. Based on our



(a) LDPC (b) ECG

(c) RCA_16 (d) TMU

Figure 8: Impact of MIV usage on 4-tier total power saving

findings, we present the following strategy:

- Step 1: Let
$$\alpha = \frac{\text{minimum cutsize of 4-way partitions}}{\text{minimum cutsize of 2-way partitions}}$$
from "75% shrinking" design.
- Step 2: Vary initial 2-tier bin size for partitioning "75% shrinking" design and find cutsize for each.
- Step 3: For each bin size used in Step 2, multiply the obtained cutsize by α. This will be our target 4-tier cutsize.
- Step 4: Vary partitioning bin size in recursive stage to obtain 4-tier cutsize closest to target cutsize found in Step 3. Note that it may not be possible to get the exact target cutsize.

In Step 1, we use a min-cut partitioner to obtain minimum cutsizes of 2-tier and 4-tier partitions and their ratio, α. In Step 2, we obtain bipartitions of "75% shrinking" by varying bin sizes. As discussed, we must explore various configurations other than min-cut solutions to achieve better enhanced M3D designs. In Step 3, we find target cutsizes to aim in Step 4. In Step 4, we vary bin sizes in recursive stage to find a 4-tier partitioning solution whose cutsize is the closest to the target cutsize. Then, we use EDA tools on partitions found in Step 4 to obtain and analyze 4-tier M3D design. Without our strategy, if we use n different bin sizes for each initial and recursive stage, we must create and compare n^2 designs. When our strategy is applied, we create only n 4-tier M3D designs. Although we cannot reduce runtime of EDA tools, effort of searching for 4-tier M3D solution is reduced to 2-tier equivalent.

Table 7 shows step-by-step application of our strategy and resulting wirelength and power savings compared with their 2D counterparts. In Section 3.2, we presented best 4-tier M3D designs found by performing a greedy search over an extensive variations of bin sizes. Compared with these results, application of our strategy provides slightly less M3D benefits. Resulting degradation are only 1.00% in wirelength and 0.6% in total power in the worst case.

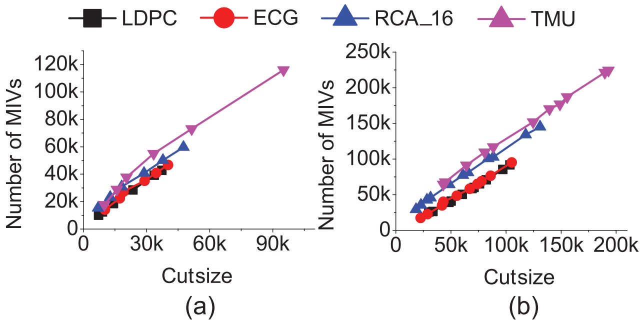


Figure 9: Cutsize vs. number of MIVs. (a) 2-tier M3D, (b) 4-tier M3D

Table 7: Effectiveness of our strategy. Wirelength and total power savings are with respect to 2D counterparts.

Circuit	Step 1 α	Step 2 Initial CS	Step 3 Target CS	Step 4 4-tier CS	WL Saving	TP Saving
LDPC	2.33	9,300	21,669	23,420	**52.0%**	**49.4%**
ECG	2.59	6,755	17,496	17,637	**34.7%**	**23.6%**
RCA_16	2.84	9,474	26,906	26,354	**19.2%**	**15.7%**
TMU	2.6	16,148	41,985	44,877	**28.2%**	**16.5%**

5. CONCLUSION

In this work, we have demonstrated tier partitioning and comprehensive CAD methodologies for 4-tier M3D IC designs. Using this methodology we have built full-chip GDSII layouts of 4-tier M3D designs. We have also investigated impacts of MIV configuration on 4-tier M3D and proposed a strategy to optimize M3D power benefits with negligible degradation. Our study shows that we achieve an average of 34.1% reduction in wirelength and 26.8% reduction in total power compared with their 2D counterparts. In addition, by advancing to 4-tier M3D from 2-tier M3D, we enhance the design benefits (e.g., area, wirelength, cell count, and buffer usage) and power reduction offered by 2-tier M3D by at least 50%. Our on-going work includes the consideration of tier-to-tier transistor performance variations caused by the low thermal budget manufacturing of M3D.

6. REFERENCES

[1] P. Batude et al. Advances in 3D CMOS sequential integration. In *Proc. IEEE Int. Electron Devices Meeting*, pages 1–4, 2009.

[2] O. Billoint et al. A comprehensive study of Monolithic 3D cell on cell design using commercial 2D tool. In *Proc. Design, Automation and Test in Europe*, pages 1192–1196, 2015.

[3] K. Chang et al. Power benefit study of monolithic 3D IC at the 7nm technology node. In *Proc. Int. Symp. on Low Power Electronics and Design*, pages 201–206, 2015.

[4] S. Panth et al. Design and CAD methodologies for low power gate-level monolithic 3D ICs. In *Proc. Int. Symp. on Low Power Electronics and Design*, pages 171–176, 2014.

[5] S. Panth et al. Placement-driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs. In *Proc. Int. Symp. on Physical Design*, pages 47–54, 2014.

[6] J. Warnock. Circuit design challenges at the 14nm technology node. In *Proc. ACM Design Automation Conf.*, pages 464–467, 2011.