Logic Monolithic 3D ICs: PPA Benefits and EDA Tools Necessary

Sai Surya Kiran Pentapati
sai.pentapati@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Da Eun Shim
daeun@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

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

ABSTRACT

Monolithic 3D (M3D) ICs provide a way to achieve high performance and low power designs within the same technology node, thereby bypassing the need for transistor scaling. M3D ICs have multiple 2D tiers sequentially fabricated on top of each other and connected through Monolithic Inter-tier vias (MIVs) which go from the top metal of the bottom die to the top metal of the top die. MIVs go through the dielectric separation between tiers and the active layer of the top die. The traditional APR tools only support place and route optimization in 2D IC designs and thus cannot handle the M3D designs. This paper provides a survey of all the efforts done to implement commercial quality M3D ICs using 2D APR tools and the benefit of M3D ICs over traditional 2D ICs.

CCS CONCEPTS

• **Hardware → 3D integrated circuits**; Software tools for EDA.

KEYWORDS

Monolithic 3D ICs, EDA tools, Design flows for Logic M3D ICs

ACM Reference Format:
Sai Surya Kiran Pentapati, Da Eun Shim, and Sung Kyu Lim. 2019. Logic Monolithic 3D ICs: PPA Benefits and EDA Tools Necessary. In *Great Lakes Symposium on VLSI 2019 (GLSVLSI '19), May 9–11, 2019, Tysons Corner, VA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3299874.3319486

1 INTRODUCTION

In order to support ever growing features in modern-day mobile devices and computers, there is an increased demand of high performance hardware. Traditionally transistor scaling has been employed to meet such demands, but it is becoming increasingly difficult to continue on with the traditional approach due to physical limitations in achieving gate-lengths that are few nanometers wide and constraints in the lithography process. 3D ICs show promising PPA improvement compared to 2D ICs within the same technology node. By splitting a 2D design into multiple dies and reducing the footprint of the chip, 3D ICs reduce the distance between a pair of cells.

GLSVLSI '19, May 9–11, 2019, Tysons Corner, VA, USA

ACM ISBN 978-1-4503-6252-8/19/05…$15.00
https://doi.org/10.1145/3299874.3319486

There have been efforts to use 3D integration in different fields of electronics design. For example, 3D NAND Flash memories are now mass-produced by companies such as Samsung and Micron with more than ninety stacked cell layers[1]. SRAM and other large memory structures have a relatively uniform structure that are based on repeating arrays which makes it easier to design in 3D manually. Logic design on the other hand requires a lot of automated optimization of different inter-dependent constraints and thus it is not possible to design a near-optimal logic 3D design manually. A 3D IC has multiple dies that are connected using vias from the bottom die to the top die.

3D designs can be differentiated into block-level, transistor-level, and gate-level 3D designs based on the granularity of 3D integration. Transistor level 3D designs require us to generate our own 3D standard cell layouts as done in [10]. Block level 3D designs can only improve the inter-block communication which limits the benefits of 3D and leaves less room for improvement. Gate level 3D designs enjoy the freedom of being able to use existing standard cells while being able to target all the nets in a design. Based on the fabrication technology, different levels of 3D integration are possible. Monolithic 3D ICs support the highest level of 3D connection density[2] and thus we will use M3D to achieve gate-level 3D ICs. M3D ICs are fabricated by processing each tier sequentially. The Front-End-Of-Line (FEOL) of the bottom tier is first grown followed by the Back-End-Of-Line (BEOL) of the bottom tier. This is then followed by FEOL of top tier containing its active layer and finally the BEOL of top tier. The standard Copper metals in BEOL of the bottom tier puts a limit on the maximum temperature to 450°C to be used in later steps to ensure stability. High temperature processes (>500°C) in the top tier also degrades the bottom tier transistor performance due to dopant de-activation. During normal semiconductor fabrication, processes like dopant activation is carried out using spike-annealing at 1050°C. Authors in paper [8] show that it is possible to perform dopant activation using solid phase epitaxial regrowth(SPER) at 450°C. There are also processes such as spacer formation, selective epitaxy that require temperatures of 600-650°C which are over the temperature limit.

In the following sections of the paper, we will present the different tool flows developed for optimizing gate-level M3D designs and how these flows handle M3D designs using commercial 2D EDA tools. We also talk about the characteristics of an ideal M3D flow and issues faced by M3D designs.

2 RTL-TO-GDSII TOOL FLOWS FOR M3D ICS

Current EDA tools only offer optimization for 2D designs and require special tricks to optimize M3D designs. The most important difference between 2D and M3D designs is that M3D designs has die-to-die connections inside the chip boundary whereas a 2D die

only has IO pins on the edges of the die. We review the following RTL-to-GDSII tool flows:

- Shrunk2D [11]
- Compact2D [9]
- Cascade2D [3]

2.1 Shrunk2D

Shrunk2D is the first RTL-to-GDSII tool flow that creates near-optimal M3D designs using 2D EDA tools. In the Shrunk2D flow, the width and height of the standard cells are first shrunk by a factor of $1/\sqrt{2} = 0.707$ thereby halving the area of standard cells (Hence the name Shrunk2D). With the shrunk dimensions, all the cells fit in a floorplan half the area of a 2D floorplan with the same cell density of a normal 2D design. With the halved footprint, the distance between cells in an M3D design is approximately 0.707 of 2D designs which gives a theoretical wirelength savings of 30%. The shrunk design can be treated as a proxy for M3D design because the cell distances are almost equal to that of the M3D design (ignoring the z-direction distances which are relatively short). At this stage, the design is still considered 2D and all the optimization capabilities of a commercial EDA tool such as cell sizing, buffer insertion, removal, routing, power optimization, timing closure etc., can be leveraged. The next step is to transform it to a 3D design without

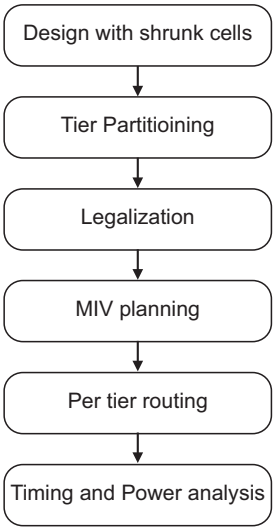

Figure 1: Shrunk2D flow as described in [11]

losing the optimization done by EDA tools. In this step, the cells are expanded back to their original sizes without changing the cell locations. This creates many overlaps among close-by cells which are resolved through a bin-based tier partitioning method. The floorplan is divided into grids and the cells in each grid/bin are partitioned into two tiers while minimizing the number of connections between the tiers using the Fiduccia-Mathheyses(FM) algorithm. This step removes most of the overlaps but there a few overlaps still remain in each tier which are then removed by legalization. The bin-based partitioning alleviates the loss of placement context from shrunk2D optimization by having small bins. Authors in [6] show that the average displacement of cells during legalization is directly proportional to the size of each grid which affects the PPA. If the bin size is too small so that it contains only a couple cells, FM can only optimize the cut-size of the nets among those cells which results in 3D overhead and a degradation in the whole chip PPA. On the other extreme, having a single bin for the entire design can minimize the number of 3D connections but can leave many overlaps unresolved. This will have a very high impact on run-time of FM and cell displacements during legalization. After the tier-partitioning, a netlist for each tier and a top level netlist that contains both tiers are created. MIV locations are then found with the help of EDA tools. A 3D tech LEF is created with the metal layer information of both tiers as well as the MIV cut layer that connects the two tiers. Then, cells from both tiers are projected on a single die and the pins are assigned to their respective tier metal layers. Routing is performed on all the 3D nets and the locations of vias between the top metal of the bottom tier and the bottom metal of the top tier are the locations of the MIVs. These locations are saved and treated as IO pin locations for each tier. Each tier is treated as a 2D die and die-by-die routing is performed. Sign-off timing analysis is done using tempus/Prime-time considering both dies as a whole with the help of a top level netlist. This gives us the final commercially viable M3D design.

The intermediate shrunk2D design has some shortcomings as a proxy for 3D designs. In the original shrunk2D flow presented in the paper [11], the wire widths and pitches are scaled by a factor of 0.707 to allow for routing on a smaller footprint. In this case, the authors did not consider the change in wire parasitics due to the geometry scaling. In paper [7] this issue is addressed to create better RC-lookup tables so that the parasitics in the shrunk2D stage mirror the parasitics in the M3D design. Correct parasitic estimation is necessary for EDA engines to optimize M3D designs. Another drawback is that the shrunk2D flow does not provide optimization after tier partitioning meaning that the cell movement during legalization and the routing overhead of 3D nets are not considered. When we refer to the shrunk2D flow in the paper we simply refer to the original shrunk2D flow without the parasitic adjustments.

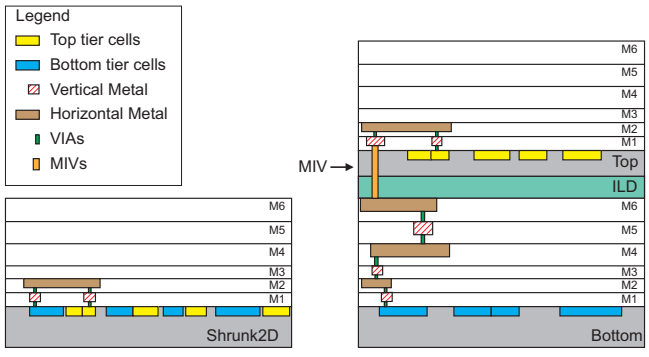

Figure 2: Routing overhead of a 3D net

2.2 Compact2D

Compact2D is another RTL-to-GDSII tool flow. It has an advantage over shrunk2D as it does not need to shrink cells which removes the need for one-technology node lower licenses for EDA tools. While the shrunk2D flow is based on the assumption that cells in M3D are on average 0.707x closer, compact2D goes one step further uses the fact that closer cells and smaller wirelengths implies smaller parasitics. This means the wire load in M3D is on average 0.707x times the wire load in 2D. Thus, compact2D simply scales the RC parasitics without shrinking any geometry. As EDA tools perform optimization based on the output load and the connectivity of cells, compact2D employs scaling of RC parasitics to mimic the M3D design. Once the design is optimized, the cells are then mapped to a halved footprint by linearly scaling the x,y coordinates by 0.707. Then we have a complete cell placement in a half-footprint of the 2D design with overlaps that are resolved in the same fashion as shrunk2D using bin-based partitioning.

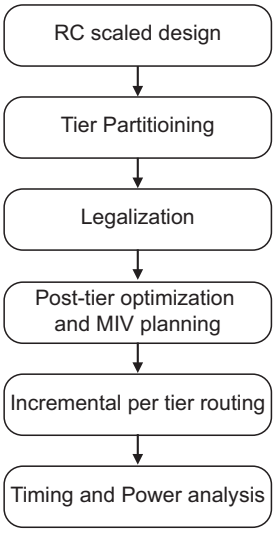

Figure 3: Compact2D flow as described in [9] for F2F 3D ICs

The second big improvement of compact2D is in the MIV planning stage, as it re-optimizes the whole design after tier-partitioning to consider the cell movement during legalization and the 3D net overhead. This particular step is called post-tier optimization and is the most important addition in compact2D. In the case of shrunk2D, placing cells on a single half-footprint die creates overlaps in the MIV planning stage. Compact2D employs placement row splitting during it's MIV planning stage to prevent overlaps. The standard cell row and cell heights are halved and by assigning every odd row to cells from one tier and even row to the other tier, the overlaps are completely removed. The pin locations stay the same as they are before halving the cells and therefore lie outside cell boundaries. With this placement row splitting, optimization engines can run without any issues. The design is therefore optimized with 3D overhead and 3D metal layer structure in consideration and is a better proxy for the final M3D design. Routing is saved for each tier in a DEF format. Each tier is then loaded with routing and MIV locations as IO ports and incremental routing is done for each tier to get a final commercial quality M3D design. Note that, in [9] the authors designed the compact2D flow for face-to-face bonded 3D ICs but it can also be used for monolithic 3D ICs.

The main drawback in compact2D is that the scaling factor of 0.707 is not accurate as not all nets scale exactly by a factor of 0.707 in M3D. The scaling factor should be modified depending on the design as wirelength savings can vary. Also, to get a true proxy of 3D parasitics a net-by-net scaling factor control is necessary which cannot be determined until the 3D design is completed.

2.3 Cascade2D

Cascade2D is a different type of 3D RTL-to-GDSII flow where the z-location(tier assignment) of the cells are determined before performing x-y optimization unlike the previous two flows. The authors present the following methods to determine the z-location:

- Decided by the designer based on the micro-architecture of the RTL
- Implement 2D design and perform tier partitioning based on the 2D results

The first option completely leaves the partitioning choice up to the designer. In case it is not feasible, the second alternative is chosen. Based on 2D design results, the authors first fix the critical timing paths on different tiers to minimize the distance between the cells within these paths. Next, the remaining cells are partitioned to maximize the number of timing paths crossing between the tiers with the cell area balance in consideration. After deciding the tier assignment, the next step is MIV planning where MIV locations are decided. This is again done in a different manner from both shrunk2D and compact2D. In cascade2D, the top tier is first designed independently of the bottom tier with MIV port locations placed on top of its respective receiving or driving cell to decrease the distance between MIV and cell-pin. This still leaves the locations of MIVs that are directly connected to IO ports to be decided. Using the MIV port locations calculated from the top tier design, the bottom tier is then optimized independently to fix the locations of the remaining MIVs by placing them on top of their connected cells on the bottom tier. Fixing the predetermined MIV locations from the top tier helps guide the placement of bottom tier cells. After this stage, only the MIV locations are preserved while discarding the routing and cell information.

A 2D floorplan is then made with two partitions side by side where each partition represents a tier. Partitions and blockages are set so that cell locations and routing of each tier is contained in their respective partitions. The connections between tier partitions are made by adding layers on top of the metals used for die-by-die routing. These additional metal layers have zero resistance and capacitance as they are used just for connectivity in the cascade2D footprint and are replaced by MIVs in direct contact in the final M3D design. All the wiring done in these additional metals is only for inter-die routing and has zero delay wire (dummy wires). Anchor cells, which are buffers with zero delay-power, are placed at the MIV locations in each partition. For the anchor cells on the bottom tier, both input and output pins are located on the top metal layer as the landing metal of MIVs is the top metal layer for the bottom tier. For the top tier, two configurations of anchor cells are designed:

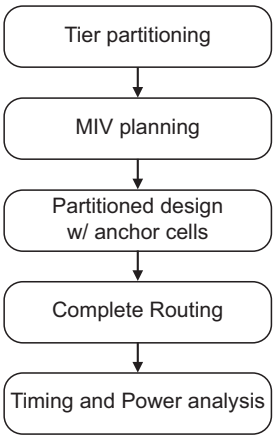

Figure 4: Cascade2D flow as described in [3]

receiving anchor cells (for MIVs connected to the input pin of a top tier cell), driving anchor cells (for MIVs connected to the output pin of the top tier cell). The receiving anchor cell has the input pin on the top metal as it receives a connecting net from the bottom tier partition. The output pin is located on the bottom-most metal of the top tier, which is the landing metal for MIVs in M3D design. The driving anchor cell receives a connection to the input pin located on the bottom metal and is connected to the bottom tier through the output pin that is located on top metal layer of the top tier. These anchor cells act as wormhole connections between the bottom metal of the top tier, where MIVs are connected in M3D designs, to the top metal which is used to connect to bottom tier partition in a Cascade2D design. This setup allows use of EDA optimization tools for routing and placement of the cells in both partitions as a whole.

2.4 Ideal Gate-level M3D Flow

There are three main aspects in a 2D IC design: Cell Placement, Clock tree layout, Routing. Optimizations are performed after each of the stages. Similarly in a 3D IC design, all three stages along with the ability to perform optimizations should remain possible for an optimized design.

Cell Placement: The flow should be able to optimize x,y,z locations simultaneously. This placement should be done based on timing and power of the design. The z-location determination is akin to the tier partitioning strategies from the tool flows discussed previously. All flows used some heuristics to find the z-location without considering the complete M3D tech LEF and M3D net overheads.

Clock tree layout: Clock tree is an important part of 2D IC design. The design of clock tree aims to reduce clock skew and improve overall timing. In the M3D flows described previously, the clock tree is designed before tier-partitioning which may not translate to the best 3D clock tree. Thus, the clock tree should be optimized with both tiers and the total metal stack in consideration.

Routing: Routing should be performed on the whole design (both tiers) at the same time. Compact2D and cascade2D flow allow for complete routing and have the perfect characteristics of an ideal 3D routing stage. In cascade2D the MIV planning is done die-by-die separately and thus the MIV planning stage is not ideal.

Optimizations: The ability to change cell locations, cell sizes, add and remove buffers at each stage of the design should be supported. All the tool flows have this ability to some extent. However, in the case of the shrunk2D flow, this type of optimization is not supported after tier partitioning. Compact2D has this ability all through-out the design stages but after tier partitioning it does not allow for cells to change tiers. Cascade2D also has this ability to the same extent as compact2D.

The ability to handle incremental updates such as ECO routing and re-optimization after manual buffer insertion, deletion, and tier changing while considering the two levels of M3D design as a whole should be supported. These features are useful for manual changes to the design after initial M3D design.

3 PPA BENEFITS OF THE M3D TOOL FLOWS

3.1 Shrunk2D

A comprehensive study of performance and power benefits of M3D ICs was done in [7]. The shrunk2D tool flow was used and improved through parasitic adjustments. The authors also incorporate other timing aware tier partitioning techniques such as placing long wires on a single tier to reduce 3D net overhead, and fixing clock tree buffers to a single tier to reduce clock skew. Using nangate45nm, up to 15.6% performance improvement is observed from Table 1. At iso-performance the authors also demonstrated power benefits ranging from as low as 5.6% in rocket-core to as high as 40% savings in LDPC as depicted in Table 2.

3.2 Compact2D

The main improvement in compact2D was post-tier partitioning optimization. This is demonstrated by the improvement in timing of M3D designs compared to both shrunk2D-based M3D and 2D designs which is observed in Table 3. Note that the shrunk2D is the original shrunk2D flow without parasitic adjustments. All the numbers in the compact2D section are from designs using a dual-Vt commercial-grade 28nm PDK.

Compared to 2D ICs, compact2D based F2F 3D designs show 5.7% to 26.8% iso-performance power improvements at the 28nm technology node as can be seen in Table 3. Please note that the results are not from M3D ICs where the die integration is front-to-back but are from the wafer bonded front-to-front fabrication considered in the paper. This table still shows the improvement of compact2D in improving timing due to its post-tier optimization with little overhead on power.

3.3 Cascade2D

All the results in Table 4 regarding cascade2D are from [3]. The authors use a foundry based 28nm planar PDK as their technology node. Shrunk2D has higher wirelength savings as it starts with placing cells closely. Cascade2D does tier partition first based on architecture aware heuristics and has lower wirelength savings than shrunk2D. However, the main improvement is in the cell area savings where cascade2D beats shrunk2D because of its choice of tier partitioning and all around optimization of both tiers in M3D. Standard cell area contributes to a higher portion of total

Table 1: Max Performance Results of shrunk2D with Parasitic Adjustment as shown in [7]

parameters	rocket-core			ride-core			AES-128			LDPC		
	2D	M3D		2D	M3D		2D	M3D		2D	M3D	
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(MHz)	**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%)
target freq(MHz)	813	938	(15.4%)	530	562	(5.9%)	1,375	1,750	(27.3%)	750	875	(16.7%)

Table 2: Iso-performance power comparison of the M3D designs implemented with baseline shrunk-2D flow (S2D-base) and shrunk-2D flow with parasitic adjustment(S2D-PA). The percentage values in the M3D designs are w.r.t. their 2D counterparts. From [7]

designs	paramaters	M3D /w S2D-base	M3D /w S2D-PA
rocket -core	sw. pwr	-9.1%	-11.8%
	int. pwr	-1.4%	-3.1%
	lkg. pwr	-2.7%	-6.8%
	tot. pwr	**-3.5%**	**-5.6%**
ride -core	sw. pwr	-16.7%	-23.0%
	int. pwr	-3.4%	-10.0%
	lkg. pwr	-4.8%	-14.2%
	tot. pwr	**-9.2%**	**-15.7%**
AES -128	sw. pwr	-7.6%	-12.2%
	int. pwr	-1.7%	-6.6%
	lkg. pwr	-3.0%	-13.5%
	tot. pwr	**-4.1%**	**-9.0%**
LDPC	sw. pwr	-41.9%	-47.5%
	int. pwr	-21.4%	-30.2%
	lkg. pwr	-22.2%	-32.7%
	tot. pwr	**-33.5%**	**-40.5%**

Table 3: Iso-performance power comparison of the F2F designs of OpenSparc T2 core implemented with original shrunk-2D design flow (S2D) and Compact-2D(C2D) Flow. The percentage values in the F2F designs are w.r.t. their 2D counterparts. From [9]

designs	paramaters	F2F /w S2D	F2F /w C2D
OpenSPARC -T2 core	**tot. pwr**	**-11.0%**	**-11.1%**
	tot. neg. slack	-1.7%	83.6%
	tot. pos. slack	8.0%	9.5%
	tot WL	-23.4%	-24.8%

power in cell dominant circuits and in improved technology nodes, cascade2D becomes more powerful in such cases.

4 FACTORS IMPACTING M3D ICS

4.1 Issues caused by sequential fabrication

As discussed before Monolithic 3D ICs use sequential fabrication that allow the use of MIVs with a high 3D interconnect density. As sequential fabrication either needs low temperature processes for the top tier FEOL or tungsten based bottom tier BEOL, one

Table 4: Iso-performance power comparison of the M3D designs of a commercial in-order processor implemented with original shrunk-2D design flow (S2D) and Cascade-2D Flow. The percentage values in the M3D designs are w.r.t. their 2D counterparts. From [3]

designs	paramaters	M3D /w S2D	M3D /w Cascade2D
32-bit application processor	tot. WL	-19.3%	-11.9%
	std. cell area	-7.6%	-9.5%
	sw. pwr	-13.4%	-13.0%
	int. pwr	-4.8%	-14.5%
	lkg. pwr	-7.7%	-9.5%
	tot. pwr	**-9.3%**	**-13.4%**

of the tiers will experience degradation either due to weak cells on the top tier or high resistance wires on the bottom tier. The authors of [12] have developed BEOL/FEOL degradation aware tier-partitioning strategies to minimize the impact of this degradation. They use shrunk2D as their baseline flow and in the case of FEOL degradation, the tier partitioning is done in such a way that cells with most available slack are assigned to the top tier. This helps in combating the timing degradation that arises due to slow top tier cells. In the case of BEOL degradation due to tungsten wires, the authors assign long wires to be routed exclusively on the top tier where the interconnects are still based on Copper.

The degradation aware tier-partitioning methods are just heuristics and are not the best possible optimization. Compact2D can exploit the commercial router capabilities to combat BEOL degradation. During the post tier-partitioning optimization of compact2D, the timing-aware routing is done by EDA tools where it can utilize the top tier BEOL when the bottom tier BEOL is degraded to improve timing on critical paths. This produces both degradation aware MIV locations, and die-by-die routing. Tier partitioning can also be improved in case of FEOL degradation by the cell slack based partitioning method mentioned in the previous paragraph. This adds to the benefit of timing aware 3D routing.

4.2 Power Delivery in M3D ICs

Power delivery in M3D is a challenge because we do not have C4 bumps on the bottom tier due to the sequential nature of M3D fabrication. As M3D designs have half the footprint of 2D designs, the maximum number of C4 bumps in M3D also halves. This almost doubles the current drawn from each bump increasing the IR-drop. For the top tier Power Delivery Network (PDN), the current is drawn from the C4 bumps as usual, but for the bottom tier the current is delivered through power MIVs (array of MIVs). This

structure means that the PDN of top and bottom tiers are in series and the IR-drop of the bottom tier is added to the already higher IR-drop in the top tier. The authors in [5] shows the static and dynamic IR-drop in M3D ICs compared to 2D ICs. According to the authors, the static instance based IR-drop in 3D ICs is almost twice that of 2D ICs. However, the dynamic IR-drop in 3D ICs is only 11.3% higher than that of 2D ICs on average. This is attributed to the higher density of decap cell placement in 3D design as they can be placed in both top and bottom tiers. It is also observed that the voltage spiking reduces in 3D due to the increased resistance of the PDN.

4.3 Thermal issues in M3D designs

Thermal hotspot is another issue in M3D designs as they have a higher power density compared to 2D ICs. It is also difficult for the heat in the bottom tier to dissipate as it is surrounded by a thick dielectric on the bottom and top tier. Heat from the bottom tier is hence only efficiently dissipated through the routing metals and the PDN to the top tier where it can be dumped into the cooling system. This makes the operating temperature of 3D ICs higher than 2D ICs. The authors in [13] used commercial thermal analysis tools with the help of some in-house scripts to perform thermal analysis. They also reduced the temperature by 5% through optimization of the PDN to improve heat flow without violating the IR-drop limit.

4.4 Frequency impact on power Benefits of M3D ICs

Higher frequency gives better power savings in M3D design. This is simply due to the fact that a higher frequency design in 2D needs stronger cells and has more timing critical paths. This allows more optimizations in 3D ICs where it can reduce wirelengths and cell sizes as lower wire-load needs lower driving strength cells. The authors in [4] explored the impact of frequency and technology node on M3D power benefits. The results can be seen in Figure 5.

4.5 PPA benefits at various technology nodes

Technology node plays an important part in determining the impact of M3D benefit on any design. This is mostly because the portion of pin capacitance increases in advanced technology nodes based on FinFETs where transistors have vertical gates placed close to each other. This increases the cross-coupling capacitance between the fins. Figure 5 shows the impact of technology node and frequency on power savings on OpenSparc T2 SoC design. The M3D designs are obtained using the shrunk2D tool flow. 28nm planar foundry-based PDK, 14nm FinFET based PDK, predictive 7nm FinFET based technologies are considered.

5 CONCLUSIONS

From the results we can conclude that M3D designs show better performance/power numbers compared to traditional 2D designs. The three M3D design flows presented show power benefits ranging from 5% to 40% based on the technologies and benchmarks. A performance improvement of up to 15% is also demonstrated. Developing better power delivery networks and cooling solutions are crucial and can effect the overall M3D quality. With the observations made, M3D can be a viable option for improving PPA of designs in absence of technology scaling. Therefore, it would be important to develop a true 3D tool to realize the complete benefits of 3D designs.

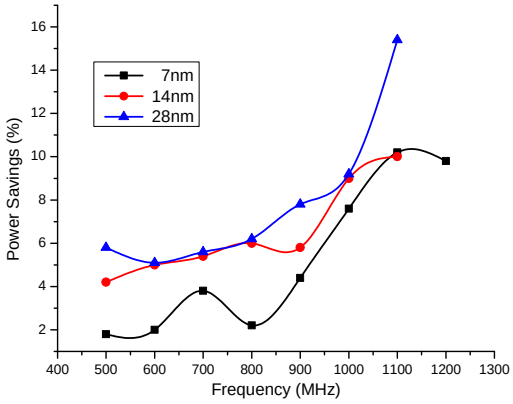

Figure 5: Impact of frequency and technology node on M3D Power savings [4]

6 ACKNOWLEDGMENTS

This research is funded by the DARPA ERI 3DSOC Program under Award HR001118C0096.

REFERENCES

[1] 3Dnand:Samsung. "https://news.samsung.com/us/samsung-fifth-generation-v-nand-mass-production/".
[2] E. Beyne. The 3-d interconnect technology landscape. *IEEE Design and Test*, 33:8–20, 2016.
[3] K. Chang et al. Cascade2D: A design-aware partitioning approach to monolithic 3D IC with 2D commercial tools. In *Proc. IEEE Int. Conf. on Computer-Aided Design*, 2016.
[4] K. Chang et al. Match-making for Monolithic 3D IC: Finding the Right Technology Node. In *Proc. ACM Design Automation Conf.*, 2016.
[5] K. Chang et al. Frequency and Time Domain Analysis of Power Delivery Network for Monolithic 3D ICs. In *Proc. Int. Symp. on Low Power Electronics and Design*, 2017.
[6] K. Chang et al. Full-chip Monolithic 3D IC Design and Power Performance Analysis with ASAP7 Library. In *Proc. IEEE Int. Conf. on Computer-Aided Design*, 2017.
[7] K. Chang, S. Pentapati, D. E. Shim, and S. K. Lim. Road to High-Performance 3D ICs: Performance Optimization Methodologies for Monolithic 3D ICs. In *Proc. Int. Symp. on Low Power Electronics and Design*, 2018.
[8] F. L. et al. Methodology for thermal budget reduction of sper down to 450c for 3d sequential integration. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 370(14-18), 2016.
[9] B. W. Ku, K. Chang, and S. K. Lim. Compact-2D: A Physical Design Methodology to Build Commercial-Quality Face-to-Face-Bonded 3D ICs. In *Proc. Int. Symp. on Physical Design*, 2018.
[10] Y.-J. Lee and S. K. Lim. Ultra High Density Logic Designs Using Transistor-Level Monolithic 3D Integration. In *Proc. IEEE Int. Conf. on Computer-Aided Design*, 2012.
[11] S. A. Panth, K. Samadi, Y. Du, and S. K. Lim. Shrunk-2D: A Physical Design Methodology to Build Commercial-Quality Monolithic 3D ICs. *IEEE Trans. on Computer-Aided Design of Int. Circuits and Systems*, 2017.
[12] S. K. Samal et al. Tier Partitioning Strategy to Mitigate BEOL Degradation and Cost Issues in Monolithic 3D ICs. In *Proc. IEEE Int. Conf. on Computer-Aided Design*, 2016.
[13] S. K. Samal, K. Samadi, P. Kamal, Y. Du, and S. K. Lim. Full chip impact study of power delivery network designs in monolithic 3D ICs. In *Proc. IEEE Int. Conf. on Computer-Aided Design*, 2014.