Journal of information and
communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 12(3): 186-192, Sep. 2014 Regular paper

Design Challenges and Solutions for Ultra-High-Density Monolithic 3D ICs

Shreepad Panth[1], Sandeep Samal[1], Yun Seop Yu[2], and Sung Kyu Lim[1*], *Member, KIICE*

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA
[2]Department of Electrical, Electronic, and Control Engineering, Hankyong National University, Anseong 456-749, Korea

Abstract

Monolithic three-dimensional integrated chips (3D ICs) are an emerging technology that offers an integration density that is some orders of magnitude higher than the conventional through-silicon-via (TSV)-based 3D ICs. This is due to a sequential integration process that enables extremely small monolithic inter-tier vias (MIVs). For a monolithic 3D memory, we first explore the static random-access memory (SRAM) design. Next, for digital logic, we explore several design styles. The first is transistor-level, which is a design style unique to monolithic 3D ICs that are enabled by the ultra-high-density of MIVs. We also explore gate-level and block-level design styles, which are available for TSV-based 3D ICs. For each of these design styles, we present techniques to obtain the graphic database system (GDS) layouts, and perform a signoff-quality performance and power analysis. We also discuss various challenges facing monolithic 3D ICs, such as achieving 50% footprint reduction over two-dimensional (2D) ICs, routing congestion, power delivery network design, and thermal issues. Finally, we present design techniques to overcome these challenges.

Index Terms: Floorplanning, IR drop, Monolithic 3D ICs, Placement, SRAM, Thermal analysis

I. INTRODUCTION

Three-dimensional integrated circuits (3D ICs) have emerged as a promising solution to extend the 2D scaling trajectory predicted using Moore’s Law. The currently available 3D ICs are enabled by through-silicon-vias (TSVs), where two prefabricated dies are aligned and bonded together. The TSV pitch is limited by the micro-bump pitch as well as the alignment accuracy. TSV-based 3D ICs, while ideal for integrating discrete board components onto a single package, do not provide a sufficient integration density to solve the on-chip interconnect problem.

Monolithic 3D ICs (M3D) are an emerging technology that enables an integration density that is some orders of magnitude higher than TSV-based 3D ICs, due to the extremely small size of the monolithic inter-tier vias (MIVs). In the case of monolithic 3D integration technology, two or more tiers of devices are fabricated sequentially, one on top of another. This eliminates the need for any die alignment, which enables considerably smaller via sizes. Each MIV has essentially the same size as a regular local via (diameter <100 nm) [1].

This ultra-high density enables several design styles, as shown in Fig. 1. First, with respect to static random-access memory (SRAM), the PMOS and NMOS of the bit-cell can be split onto multiple tiers. This gives us the opportunity to tune the PMOS and NMOS processes separately, leading to an optimum process for each device type. Next, a similar

Received 12 February 2014, **Revised** 27 February 2014, **Accepted** 16 May 2014
***Corresponding Author** Sung Kyu Lim (E-mail: limsk@ece.gatech.edu)
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA.

Open Access http://dx.doi.org/10.6109/jicce.2014.12.3.186 print ISSN: 2234-8255 online ISSN: 2234-8883

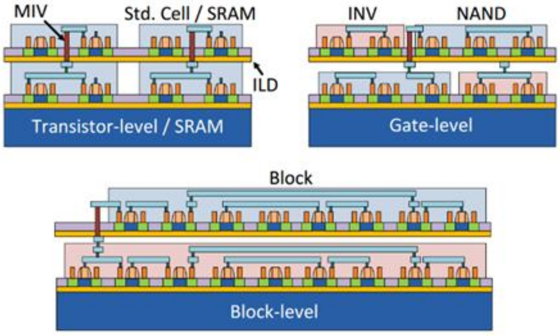

Fig. 1. Various design styles available for monolithic 3D ICs.

separation can be done for standard cells themselves, and this is known as transistor-level M3D. This design style has both intra-cell and inter-cell MIVs. Another design style is gate-level M3D, where the standard cells themselves are 2D, but they are placed in a 3D space, and interconnected using MIVs. This design style has only inter-cell MIVs. Finally, the coarsest level of integration is provided by block-level M3D, where each functional block is 2D and the 2D blocks are floorplanned onto a 3D space. In this design style, the MIVs are limited to the whitespace between blocks. We will now discuss each of these design styles in detail.

II. MONOLITHIC 3D SRAM

Monolithic 3D technology offers a unique optimization opportunity for SRAM designs [2]. We can split PMOS and NMOS onto different tiers; this allows us to optimize the process of each type of transistor independently. We pick a state-of-the art 6T SRAM cell as our 2D baseline. This is designed in a 22-nm node and has an area of 0.1 μm^2, as shown in Fig. 2(a). A default 6T SRAM cell has a 2-PMOS and 4-NMOS (2P4N) configuration. The obvious choice is to blindly split this bit-cell into two tiers, but we observe that it only gives us a 33% footprint reduction due to the imbalance in the PMOS and NMOS count, as shown in Fig. 2(b). Therefore, we explore various alternate design options to give us a larger footprint reduction.

The first option that we explore is the same 2P4N configuration, but with different sizing (Fig. 2(c)). We can obtain a footprint reduction of 44% with the same static noise margin (SNM) as 2D, but slightly worse write stability. Next, we explore changing the number of PMOS and NMOS devices in the bit-cell while keeping the total transistor count the same. We first explore a 3P3N configuration, replacing one pass-transistor with an NMOS device (Fig. 2(d)). The footprint reduction in this case is 45%. Using a single-ended read technique, we can achieve a high SNM margin. Lastly, we explore an 8T bit-cell. The conventional 2P6N configuration for 2D is shown in Fig. 2(e). However, since the PMOS and NMOS devices are on separate tiers, splitting this configuration will lead to an area imbalance. Therefore, we move to a 4P4N configuration that can give us a better area balance. This gives us a 40% footprint reduction under the same read margin, write margin, and access time as 2D.

III. TRANSISTOR-LEVEL MONOLITHIC 3D

Transistor-level M3D is similar to the SRAM case in the sense that the PMOS and NMOS devices are split onto multiple tiers. In this design style, each standard cell is redesigned such that its PMOS and NMOS devices are on different tiers [3-5]. As in the case of SRAM, the advantage of doing this is that the PMOS and NMOS devices can be optimized separately.

We begin by constructing a library of 66 monolithic 3D standard cells by using a cell-folding technique. An overview of the proposed approach for a simple inverter is shown in Fig. 3. We draw a cut-line in the center of the cell and then, fold the cell along this line. All the connections that exist along the cut-line will now become MIVs.

When compared with 2D, we observe a footprint reduction of approximately 40% because of an imbalance between the PMOS and the NMOS sizes, and some additional area required for the MIVs themselves. We then perform extraction on each cell and recharacterize the cells taking into account the new cell’s internal parasitics. The advantage of this design style is that we can utilize the existing 2D P&R tools to perform all the design steps for us. From the tool’s perspective, the standard cells have pins on different metal layers, and the router is capable of connecting all these pins together, inserting the inter-cell MIVs in the process.

Since the total number of pins remains the same and the footprint area is reduced, there is a 1.7–2 times increase in the pin density of the chip, with fewer routing resources than 2D. Therefore, this causes several routability issues. We explore several interconnect options to mitigate the congestion in transistor-level M3D. An overview of the various interconnect options considered is shown in Fig. 4. We consider three different interconnect options: (1) one metal layer on the bottom tier (1BM), (2) three additional metal layers on the top tier (3TM), and (3) three additional metal layers on the bottom tier (4BM). In the 4BM case, we observe that the additional stacked vias within each cell lead to a significant increase in the cell internal parasitics, which increases the cell delay and power by up to 9.86% and 15.65%, respectively. Among the three options considered, we observe that the 3TM case gives us the best results with up to 22% reduction in the total power of the chip. To

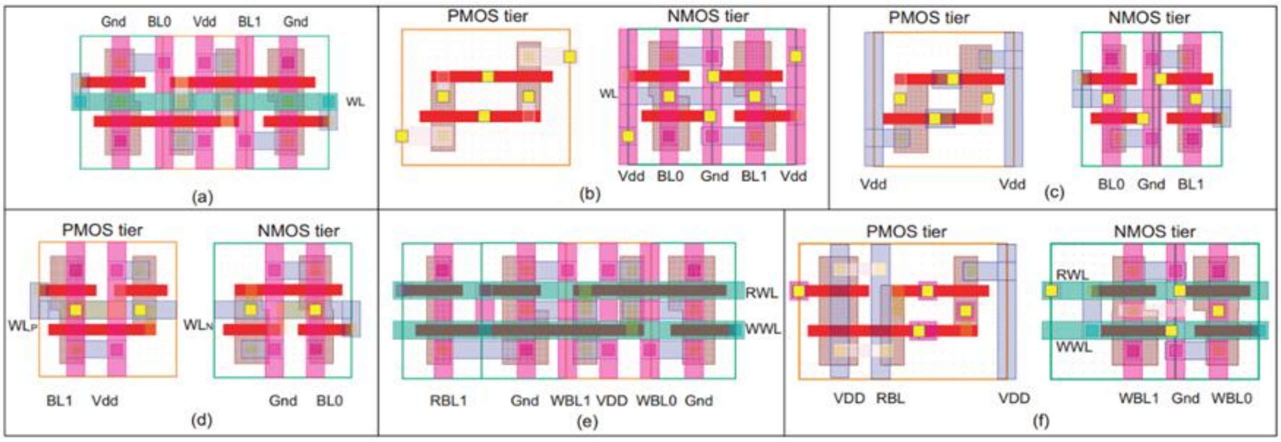

Fig. 2. Layout of different SRAM cell designs. Yellow squares denote inter-tier vias. (a) 2D 6T SRAM (=2P4N) cell, (b) 3D 2P4N SRAM without transistor re-sizing, (c) 3D 2P4N SRAM with 3D-oriented sizing, (d) 3D 3P3N SRAM, (e) 2D 2P6N 8T SRAM, and (f) 3D 8T SRAM with a modified structure.

further improve this benefit, we fine-tune the 3TM case. For example, the routing congestion is not always on the intermediate metal level. Therefore, we explore other options like utilizing two intermediate and two global metal layers instead of three intermediate metal layers, and this gives us a further 2.8% power benefit.

Since interconnects play a more dominant role at the lower nodes, we also study the benefit of this design style at the more advanced and future technology nodes, such as 22 nm and 7 nm. We observe that at the 22-nm and 7-nm nodes, we get an additional 4% and 23% power benefit, respectively.

IV. GATE-LEVEL MONOLITHIC 3D

In this design style, each standard cell is 2D, and these 2D standard cells are placed onto a 3D space. The advantage of this method is the reuse of the existing standard cells, which can avoid the need for library re-characterization. Once the gates are placed in 3D, MIVs are inserted into the existing whitespace available between the cells.

We propose a design flow based on the ‘shrunk 2D’ gate placement (shown in Fig. 5) that leverages the existing commercial 2D placers. This approach first halves the footprint area to represent a monolithic 3D footprint so that there is exactly 0% total silicon area overhead over 2D. Next, the placement capacity is doubled (or the area of the standard cells is halved), and the commercial 2D engine is run to obtain the initial placement. The shrunk 2D placement then needs to be partitioned to obtain a legal monolithic 3D IC placement. We define the partition bins and partition the design with a local area balance in each placement bin to create a gate-level M3D design. We demonstrate that this approach can give us up to 30% HPWL savings when compared with 2D ICs.

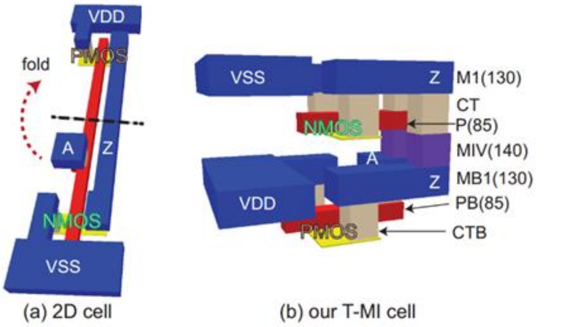

Fig. 3. Layout of an inverter from (a) Nangate 45-nm library, and (b) our transistor-level monolithic 3D library. P, M, and CT represent poly, metal, and contact, respectively. The suffix ‘B’ denotes the bottom tier. Top/bottom-tier silicon substrate and p/n-wells are not shown for the sake of simplicity. Numbers in parentheses denote thickness in nanometers.

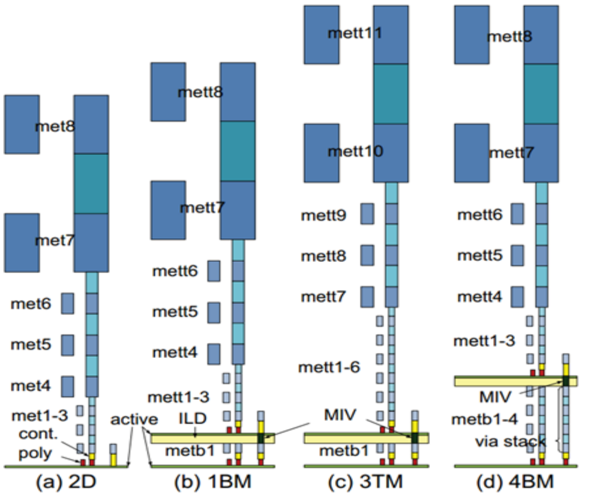

Fig. 4. Metal layer stack options: (a) 2D, (b) baseline MI-T, (c) 3 local metal layers added to the top tier, and (d) 3 local metal layers added to the bottom tier. ILD stands for inter-layer dielectric between the top and the bottom tier. The bottom-tier substrate and ILD for metal layers are not shown for the sake of simplicity. Objects are drawn to scale.

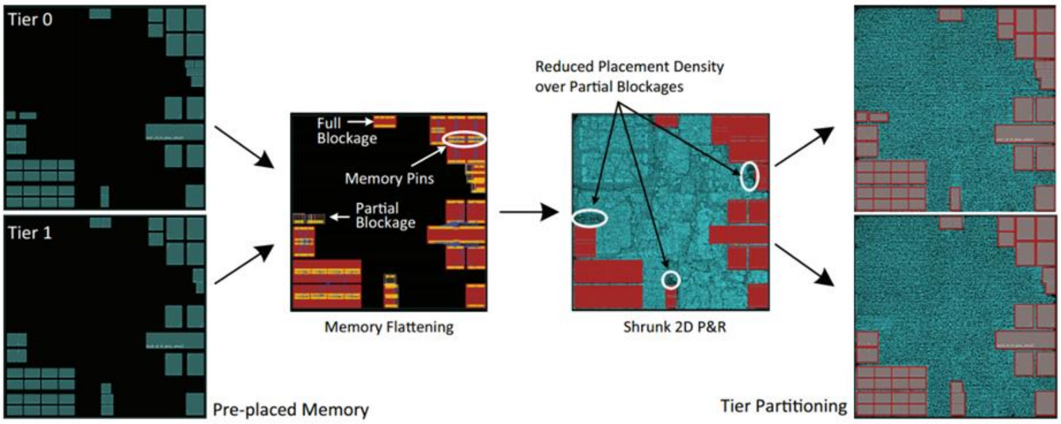

Fig. 5. Shrunk 2D technique for gate-level monolithic 3D IC placement. Pre-placed memory is projected to obtain a shrunk 2D footprint on which 2D P&R is performed. This is then partitioned to obtain a monolithic 3D solution.

To insert MIVs into the layout, the conventional approach is to perform a cell and 3D-via co-placement step. We propose a commercial-router-driven MIV insertion algorithm that improves the routed wirelength (WL) by up to 16.6% and the power delay product (PDP) by up to 6.1%.

To further reduce the routed WL, we propose a routability-driven partitioner that utilizes the fine-grained nature of MIVs to reduce routing congestion. By intelligent partitioning, we move nets that are in the congested regions in one tier to a non-congested tier. The proposed approach provides us with an additional benefit of 4% WL and 4.33% PDP. We also demonstrate that the use of multiple MIVs per 3D net can provide a benefit of 8.43% WL and 2.25% PDP.

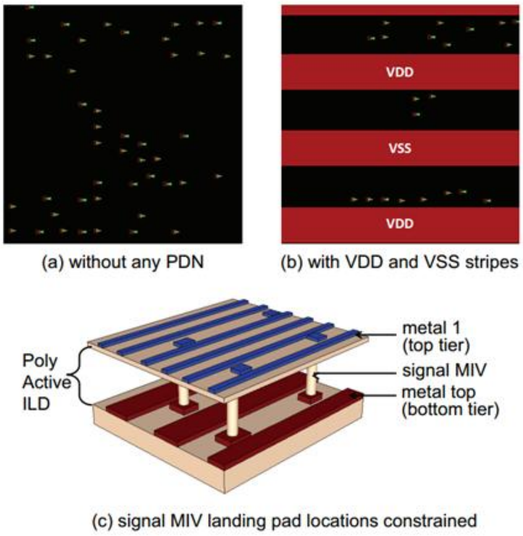

Fig. 6. Impact of PDN on MIV landing pads: (a) MIVs freely distributed without any PDN blockages in top metal, (b) PDN blockages affect MIVs in top metal, and (c) isometric view showing the constraints on signal MIV landing pad locations in top metal and metal1 of the next tier. PDN: power delivery network, MIV: monolithic inter-tier vias.

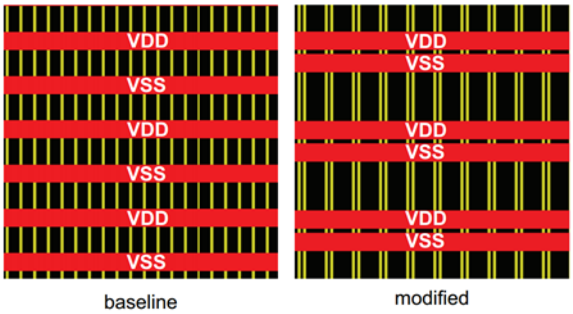

Fig. 7. Baseline PDN vs. modified PDN. Note the extra continuous space between the red top metal wires, which enhances MIV insertion and routing. The yellow wires are placed on the intermediate metal. PDN: power delivery network, MIV: monolithic inter-tier vias.

We also propose techniques for utilizing a commercial tool for timing optimization and clock-tree synthesis. We demonstrate that keeping the clock backbone on a single tier gives us 29.82% clock power reduction as compared to the case where we have one separate clock tree per tier. Overall, we demonstrate on the OpenSparc T2 design that M3D can give a 15.57% power benefit as compared to commercial-quality 2D designs. We also demonstrate that this benefit increases to 16.08% when utilizing dual-Vt libraries.

One of the main challenges facing M3D is the design of the power delivery network (PDN). In conventional 2D ICs, the entire top metal layer is available for PDN design. However, in M3D, the top metal layer needs to be used for both PDN and MIV landing pads in the top tier [6]. An example of this is shown in Fig. 6. In Fig. 6(a), we show the MIV landing pad locations without any PDN present in the top tier. We observe that they are spread out all over the footprint of the chip. In Fig. 6(b), we first create a PDN and then, perform MIV planning. We observe that the MIV locations are now confined to the space between the PDN wires. An isometric view of the MIVs in both tiers is

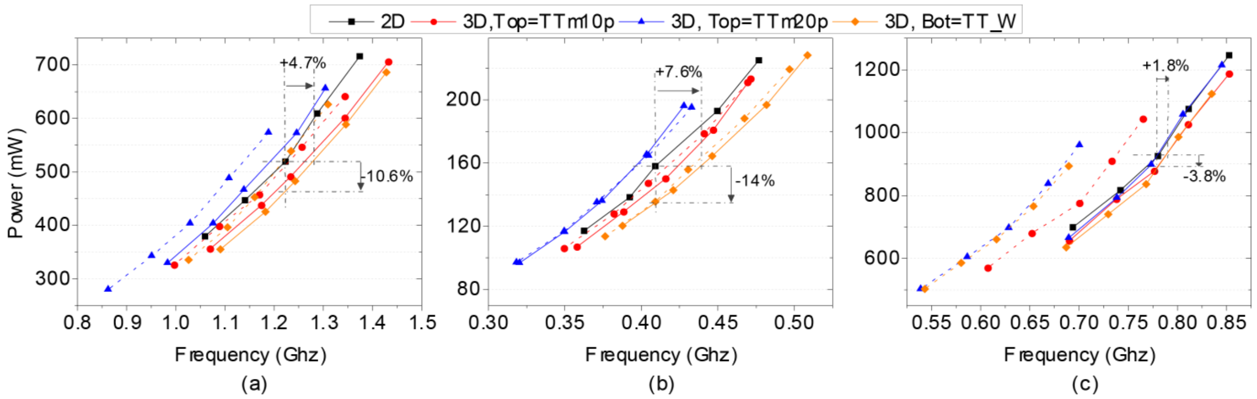

Fig. 8. Power-performance trade-off curves assuming degraded transistors and interconnects. Dashed lines represent non-variation-aware floorplanning, and solid lines represent variation-aware floorplanning (TTm10p denotes 10% worse transistors; TTm20p, 20% worse transistors; and TT_W, tungsten interconnects). (a) des3, (b) b19, and (c) mul128.

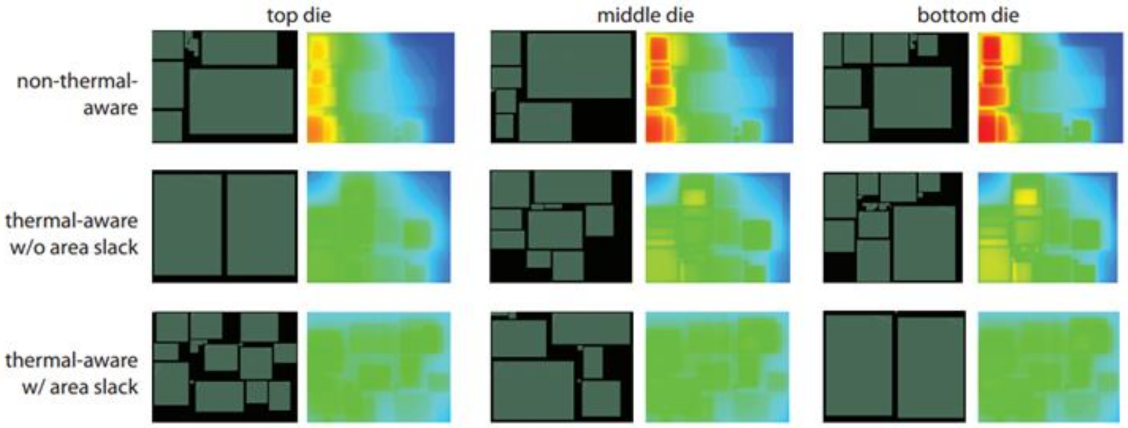

Fig. 9. Three-tier floorplanning layouts with the corresponding absolute temperature maps. The thermal-aware floorplans avoid the stacking of high-power-density blocks and result in 22% temperature reduction in a lesser total area. The temperature range is [47°C, 68 °C].

shown in Fig. 6(c). Because of the conflict between the PDN and the MIV landing pads, the addition of the PDN increases the M3D WL by 20.5%, as compared to an only 7.1% WL increase when the PDN was added to the 2D ICs. This in turn increases the net power and temperature, reducing the benefit of M3D.

We propose a PDN optimization technique that helps reduce the routed WL. An overview of this technique is shown in Fig. 7. In the original case, the VDD and VSS lines are spaced evenly. This leads to fragmented whitespace, which makes MIV planning difficult, there-by increasing the routed WL. In the proposed optimization approach, we move the VDD and VSS lines close to each other. This leads to larger continuous whitespace available for MIV insertion. This lowers the routed WL by up to 8%. We also perform a GDSII-level thermal analysis on the M3D chip with and without the presence of PDN. With the proposed PDN technique, the power consumed by the chip decreases because of the lower routed WL. This in turn leads to a 5% reduction in the maximum temperature of the chip, while still meeting the given IR-drop budget.

V. BLOCK-LEVEL MONOLITHIC 3D ICs

Block-level M3D utilize the existing functional 2D IP blocks and floorplan them onto a 3D space [7, 8]. This design style can be used for SoC-level integration and it has the benefit of IP reuse.

We present a simulated annealing framework for M3D floorplanning, which uses a weighted sum of the WL and the area as the cost function. Unlike TSV-based floor-planners, the MIV count is not included in the cost function because MIVs are so small that we need not minimize their number. Once the blocks are floor-planned onto a 3D space, we need to insert MIVs in the whitespace between them. For this purpose, we present a commercial router-driven MIV insertion algorithm. For performance evaluations, we first assume that the monolithic 3D fabrication process is mature and that both tiers have an identical performance. In this case, we demonstrate that we can close the gap to the ideal block-level implementation by up to 50% in terms of both power and performance. The ideal block-level implementation is obtained by designing the chip assuming perfect inter-block interconnects, i.e., the inter-block nets have zero resistance and capacitance. This is the best possible block-level design for a given benchmark, given the same set of blocks.

However, currently, we cannot achieve identical tier performance. During the manufacturing process of the top tier, we need to take care not to damage the underlying interconnects and transistors. This can be achieved by using either a low-temperature process on the top tier or tungsten on the bottom tier. We model the impact of both these options and present a variation-aware floorplanning scheme that makes the design tolerant to such manufacturing-induced performance variations. A summary of the power-performance results for these variations is shown in Fig. 8. These results demonstrate that our variation-tolerant floorplanning scheme improves the chip performance and power by up to 12.6% and 10.6%, respectively. We also demonstrate that tungsten interconnects on the bottom tier are preferable to degraded transistors on the top tier. Finally, we demonstrate that even under such performance variations, we can still close the gap to the ideal block-level implementation by up to 50% in terms of performance and 36% in terms of power.

The increase in power density associated with 3D ICs means that thermal-aware design methodologies have become necessary [9]. We first study the thermal properties of monolithic 3D ICs and observe that it has several unique properties. First, the extremely thin tiers mean that there is negligible lateral thermal coupling. In addition, the absence of any bonding or adhesive layer for underfill implies that heat is not trapped in a given tier and that the vertical thermal coupling is very high. In addition, the small size of MIVs means that they do not serve as a conduction path and their location need not be optimized for thermal reasons.

These properties enable us to develop a multi-adaptive regression spline (MARS) model to quickly estimate the temperature of a monolithic 3D IC. We generate a large sample set considering different block powers and use it to train our model. We demonstrate that the proposed model has an error of less than 5% when compared to GDSII-level FEA simulations. Further, the proposed model is extremely fast and is 10^5 times faster than prior quick thermal approaches. This extremely quick computation means that the proposed model can be used within a simulated annealing floorplanning framework, which makes millions of cost function evaluations to come up with a floorplanning solution. We modify the cost function of the proposed simulated annealing framework to be a weighted sum of the area, WL, and the maximum temperature of the chip. We first run the non-modified floorplanner until certain area and WL targets are met. Next, the temperature term in the cost function is introduced, and the area and the WL serve as constraints instead of objectives. Therefore, the floorplanner does not increase the chip area to reduce the maximum temperature. Using this approach, we demonstrate up to 22% reduction in the maximum temperature of the chip, without affecting other design metrics such as WL and area. Floorplan screenshots with and without thermal-aware floorplanning are shown in Fig. 9. We perform two thermal-aware runs—with and without area slack. In the case with an area slack, the footprint area constraint is relaxed slightly so that the thermal-aware floorplanner can achieve a better solution. We observe that relaxing the constraint leads to 22% area reduction in addition to temperature reduction.

VI. CONCLUSION

We have explored several design styles that are available for monolithic 3D ICs—SRAM, transistor-level, gate-level, and block-level. For each design style, we have presented design flows to obtain GDSII-level signoff-quality power and performance results. We have enumerated various challenges facing M3D, and the techniques to overcome them. Overall, ultra-high-density monolithic 3D ICs offer significant benefits over 2D ICs.

REFERENCES

[1] P. Batude, M. Vinet, A. Pouydebasque, C. Le Royer, B. Previtali, C. Tabone, et al., “Advances in 3D CMOS sequential integration,” in *Proceedings of the IEEE International Electron Devices Meeting*, Baltimore, MD, pp. 1-4, 2009.

[2] C. Liu and S. K. Lim, “Ultra-high density 3D SRAM cell designs for monolithic 3D integration,” in *Proceedings of the IEEE International Interconnect Technology Conference*, San Jose, CA, pp. 1-3, 2012.

[3] C. Liu and S. K. Lim, “A design tradeoff study with monolithic 3D integration,” in *Proceedings of the 13th International Symposium on Quality Electronic Design*, Santa Clara, CA, pp. 529-536, 2012.

[4] Y. J. Lee, P. Morrow, and S. K. Lim, “Ultra high density logic designs using transistor-level monolithic 3D integration,” in *Proceedings of the IEEE International Conference on Computer-Aided Design*, San Jose, CA, pp. 539-546, 2012.

[5] Y. J. Lee, D. Limbrick, and S. K. Lim, “Power benefit study for ultra-high density transistor-level monolithic 3D ICs,” in *Proceedings of the IEEE/ACM Design Automation Conference*, Austin, TX, pp. 1-10, 2013.

[6] S. 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 *Proceedings of the IEEE International Conference on*

Computer-Aided Design, 2014.

[7] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "High-density integration of functional modules using monolithic 3D-IC technology," in *Proceedings of the 18th IEEE/ACM Asia and South Pacific Design Automation Conference*, Yokohama, Japan, pp. 681-686, 2013.

[8] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Power-performance study of block-level monolithic 3D-ICs considering inter-tier performance variations," in *Proceedings of the 51st IEEE/ACM Design Automation Conference*, San Francisco, CA, pp. 1-6, 2014.

[9] S. K. Samal, S. Panth, K. Samadi, M. Saeidi, Y. Du, and S. K. Lim, "Fast and accurate thermal modeling and optimization for monolithic 3D ICs," in *Proceedings of the 51st IEEE/ACM Design Automation Conference*, San Francisco, CA, pp. 1-6, 2014.

Shreepad Panth
received his B.S. degree from Anna University, India, in 2009. He received his M.S. degree from Georgia Institute of Technology in 2011, where he is currently pursuing his Ph.D. His research interests include physical design methodologies for monolithic 3D ICs. He is the author of more than 20 publications in top conferences and journals and received the Best Paper award at ATS'12 and nominations for the Best Paper awards at ISPD'14 and DAC'14.

Sandeep Samal
is a Ph.D. student at the School of Electrical and Computer Engineering, Georgia Institute of Technology. He received his B.Tech. in Electronics and Electrical Communication Engineering from Indian Institute of Technology, Kharagpur, in 2012, and his M.S. in Electrical and Computer Engineering from Georgia Institute of Technology in 2013. His research interests include low-power and reliable digital design and analysis using 3D IC technology.

Yun Seop Yu
received his B.S., M.S., and Ph.D. degrees from Department of Electronics Engineering, Korea University, Seoul, Korea, in 1995, 1997, and 2001, respectively. From 2001 to 2002, he worked as Guest Researcher at Electronics and Electrical Engineering Laboratory, NIST, Gaithersburg, MD. He is Full Professor at Department of Electrical, Electronic and Control Engineering, Hankyong National University, Anseong, Korea. His main research interests are in the fields of modeling various nano devices for efficient circuit simulation, and future memory, logic, and sensor designs using these devices. He is also interested in the fabrication and characterization of various nano devices. He has authored and coauthored 60 refereed international journal papers.

Sung Kyu Lim
received his B.S., M.S., and Ph.D. degrees from Computer Science Department, University of California, Los Angeles (UCLA), in 1994, 1997, and 2000, respectively. He is Full Professor at School of Electrical and Computer Engineering, Georgia Institute of Technology. His research focus is on the architecture, design, and test of and EDA solutions for 3D ICs. His research on 3D IC reliability is featured as Research Highlight in the Communication of the ACM in 2014. Dr. Lim received the National Science Foundation Faculty Early Career Development (CAREER) Award in 2006. His work was nominated for the Best Paper award at ISPD'06, ICCAD'09, CICC'10, DAC'11, DAC'12, ISLPED'12, and DAC'14. He is Associate Editor of the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.