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]
[1]Dept. of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA
[2]Hankyong National University, Korea

***Abstract*—Monolithic 3D ICs (M3D) are an emerging technology that offers an ultra-high-density 3D integration due to the extremely small size of monolithic inter-tier vias. We explore various design styles available in M3D and present design techniques to obtain GDSII-level signoff quality results for each of these styles. We also discuss various challenges facing each style and provide solutions to them.**

I. Introduction

Monolithic 3D ICs (M3D) is an emerging technology that offers orders of magnitude higher integration density than conventional through-silicon-via (TSV) based 3D ICs. This is because it utilizes a sequential stacking process, which eliminates the need for tier alignment. This enables the monolithic inter-tier vias (MIVs) to be the same size as regular local vias ($< 100nm$) [1].

This ultra-high-density enables several design styles, as shown in Figure 1. First, with respect to 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 process separately. Next, a similar 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 in 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 whitepsace between blocks. We now discuss each of these design styles in detail.

II. Monolithic 3D SRAM

Monolithic 3D offers a unique optimization opportunity for SRAM designs [2]. We can split the PMOS and NMOS onto different tiers, which 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 22nm node, and it has an area of $0.1\mu m^2$. The default 6T SRAM cell has a 2 PMOS and 4 NMOS (2P4N) configuration. The obvious choice is to blindly split up 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. We therefore explore various alternative design options to give us a larger footprint reduction.

The first option we explore is the same 2P4N configuration, but with different sizing. We are able to obtain a footprint reduction of 44% with the same static noise margin (SNM) as 2D, but slightly worse write stability. Next, we explore a 3P3N configuration, replacing one pass transistor with an NMOS. The footprint reduction in this case is 45%. Using a single-ended read technique, we are able to achieve a high SNM margin. Lastly, we explore an 8T bit-cell, changing the conventional 2P6N configuration to 4P4N for area balance. This gives us a 40% footprint reduction under the same read margin, write margin and access time as 2D.

This research is supported by Intel Research, Qualcomm Research, and the Center for Integrated Smart Sensors (CISS-2012366054194).

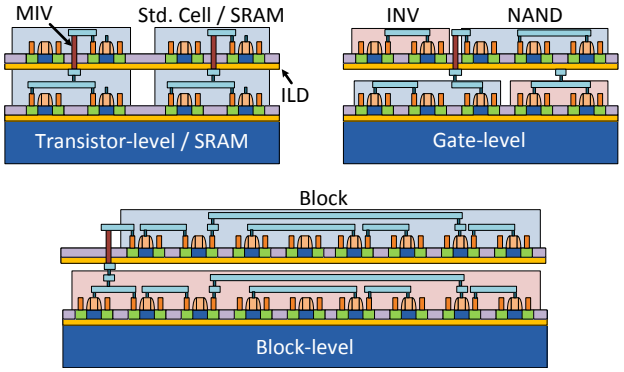

Fig. 1. Various design styles available for monolithic 3D ICs.

III. Transistor-level Monolithic 3D ICs

Transistor-level M3D is similar to the SRAM case in the sense that the PMOS and NMOS are split onto multiple tiers. In this design style, each standard cell is redesigned such that its PMOS and NMOS are on different tiers [3], [4], [5]. As in the case of SRAM, the advantage of doing this is that the PMOS and NMOS can be optimized separately.

We begin by constructing a library of 66 monolithic 3D standard cells using a cell folding technique. When compared with 2D, we observe a footprint reduction of about 40% because of an imbalance between PMOS and NMOS sizes. We re-characterize the cells taking into account the new cell internal parasitics. The advantage of this design style is that we can utilize existing 2D P&R tools to perform all the design steps for us. The standard cells have pins on different metal layers, and the router is capable of connecting all these pins together, inserting inter-cell MIVs in the process.

Since the total number of pins remain the same and the footprint is reduced, there is a 1.7-2$\times$ increase in the pin density of the chip, which causes several routability issues. We explore several interconnect options to mitigate this impact. Our first comparison set is the default case with one metal layer on the bottom tier (1BM), three additional metal layers on the top tier (3TM), and three additional metal layers on the bottom tier (4BM). We observe that the 4BM case leads 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. Overall, we observe that the 3TM case gives best results with up to a 22% reduction in the total power of the chip. We also 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.

We also study the benefit of this design style at more advanced and future technology nodes such as $22nm$ and $7nm$. We observe

978-1-4799-7439-9/14/$31.00 ©2014 IEEE

that at the $22nm$ and $7nm$ nodes, we get an additional 4% and 23% power benefit respectively.

IV. Gate-level Monolithic 3D ICs

In this design style, existing 2D standard cells are placed onto a 3D space [6], [3], [7]. The advantage of this design style is that it offers reuse of existing standard cells, a 50% or more footprint area reduction, and since each tier has equal number of metal layers as 2D designs, no increase in pin density.

We propose a design flow based on "shrunk 2D" gate placement that leverages existing commercial 2D placers. This approach halves the footprint area and doubles the placement capacity in each placement bin to give an initial placement. The shrunk 2D placement is then partitioned with local area balance in each placement bin to give us a gate-level M3D design. We demonstrate that it can give us up to 30% HPWL savings when compared with 2D ICs. We also 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%. Next, we propose a routability-driven partitioner that utilizes the fine-grained nature of MIVs to reduce routing congestion. Our approach helps give us an additional 4% WL and 4.33% PDP benefit. We also demonstrate that using multiple MIVs per 3D net can help give us a 8.43% WL benefit and 2.25% power benefit. Next, we propose techniques for utilizing a commercial tool for timing optimization and clock-tree-synthesis (CTS). We demonstrate that keeping the clock backbone on a single tier gives us 29.82% clock power reduction 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 compared to commercial-quality 2D designs. We also demonstrate that this benefit rises to 16.08% when utilizing dual-V_t libraries.

We also explore power delivery network (PDN) issues in M3D [8]. Since the top metal layers need to be used for both PDN as well as MIV landing pads, we demonstrate that PDN increases the M3D WL by 20.5% compared to only a 7.1% increase in 2D. This in turn increases the net power and temperature, reducing the benefit of M3D. We propose PDN optimization techniques that reduce the routed WL by up to 8% and the maximum temperature by up to 5% while still meeting the original IR-drop budget.

V. Block-level Monolithic 3D ICs

Block-level monolithic 3D ICs utilize existing functional 2D IP blocks and floorplan them onto a 3D space [9], [10]. This design style can be used for SoC-level integration, and it also has the benefit of IP reuse.

We present a simulated annealing framework for M3D floorplanning, which uses a weighted sum of wirelength and area as the cost function. We also present a router-driven MIV insertion algorithm that inserts MIVs into the whitespace between blocks. We first demonstrate that in the case of a perfect manufacturing process, we can close the gap to the ideal block-level implementation by up to 50% w.r.t. both power and performance. The ideal block-level implementation is obtained by designing the chip assuming perfect inter-block interconnects. This is the best possible block-level design for a given benchmark.

However, during the manufacturing process of the top tier, we need take care not to damage the underlying interconnects and transistors, which can be achieved by using tungsten on the bottom tier. We model the impact of the tungsten interconnects and present a variation-aware floorplanning scheme that improves the performance and power by up to 12.6% and 10.6% respectively. 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% w.r.t. performance and 36% w.r.t power.

The increase in power density associated with 3D ICs mean that thermal-aware design methodologies have become necessary [11]. We first study the thermal properties of monolithic 3D ICs and observe that the extremely thin tiers leads to negligible lateral thermal coupling. In addition, the absence of a bonding layer means 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 mean 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 non-linear multi adaptive regression spline (MARS) model to quickly estimate the temperature of a monolithic 3D IC. We demonstrate a modeling error of $< 5\%$ when compared to GDSII-level FEA simulations. In addition, our model is extremely fast, and is 10^5 times faster than prior quick-thermal approaches. This extremely quick computation means that our model can be used within a simulated annealing floorplanning framework. We modify our simulated annealing framework to include the temperature in the cost function. The non-modified floorplanner is first run until a certain area and wirelength target are met. Next, the temperature term in the cost function is introduced, and the area and wirelength serve as constraints instead of objectives. Using this approach, we demonstrate up to a 22% reduction in the maximum temperature of the chip, without affecting other design metrics such as wirelength and area.

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 techniques to overcome them. Overall, ultra-high-density monolithic 3D ICs offers significant benefit over 2D ICs.

References

[1] P. Batude *et al.*, "Advances in 3D CMOS Sequential Integration," in *Proc. IEEE Int. Electron Devices Meeting*, 2009, pp. 1–4.

[2] C. Liu and S. K. Lim, "Ultra-High Density 3D SRAM Cell Designs for Monolithic 3D Integration," in *Proc. IEEE Int. Interconnect Technology Conference*, 2012.

[3] C. Liu and S. K. Lim, "A Design Tradeoff Study with Monolithic 3D Integration," in *Proc. Int. Symp. on Quality Electronic Design*, 2012.

[4] Y.-J. Lee, P. Morrow, and S. K. Lim, "Ultra High Density Logic Designs Using Transistor-Level Monolithic 3D Integrationn," in *Proc. IEEE Int. Conf. on Computer-Aided Design*, 2012.

[5] Y.-J. Lee, D. Limbrick, and S. K. Lim, "Power Benefit Study for Ultra-High Density Transistor-Level Monolithic 3D ICs," in *Proc. ACM Design Automation Conf.*, 2013.

[6] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs," in *Proc. Int. Symp. on Physical Design*, 2014.

[7] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Design and CAD Methodologies for Low Power Gate-level Monolithic 3D ICs," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2014.

[8] S. K. Samal *et al.*, "Full Chip Impact Study of Power Delivery Network Designs in Monolithic 3D ICs," in *Proc. IEEE Int. Conf. on Computer-Aided Design*, 2014.

[9] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "High-Density Integration of Functional Modules Using Monolithic 3D-IC Technology," in *Proc. Asia and South Pacific Design Automation Conf.*, 2013.

[10] 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 *Proc. ACM Design Automation Conf.*, 2014.

[11] S. Samal *et al.*, "Fast and Accurate Thermal Modeling and Optimization for Monolithic 3D ICs," in *Proc. ACM Design Automation Conf.*, 2014.