Design Automation and Testing of Monolithic 3D ICs: Opportunities, Challenges, and Solutions

(Invited Paper)

Kyungwook Chang[1], Abhishek Koneru[2], Krishnendu Chakrabarty[2], and Sung Kyu Lim[1]
[1]School of ECE, Georgia Institute of Technology, Atlanta, GA
[2]Department of ECE, Duke University, Durham, NC
k.chang@gatech.edu

ABSTRACT

Monolithic 3D ICs (M3D) are fabricated using a sequential process that grows new device and interconnect tiers in a bottom-up fashion. This fabrication process is in contrast to through-silicon via (TSV) technology that aligns and bonds pre-built tiers. M3D offers key advantages over TSVs, including (1) orders-of-magnitude smaller inter-tier vias, (2) no need for high alignment accuracy, (3) finer-grained tier partitioning options, etc. Recent studies have shown power, performance, area, and reliability (PPAR) advantages of M3D over TSV. However, M3D also suffers from its own problems, including (1) device and interconnect performance mismatch between tiers, (2) lack of EDA solutions, (3) testing challenges, (4) cost, etc. Research efforts have also been made to model and mitigate the impact of these undesirable characteristics of M3D. We will provide a survey of work that address the above issues and conclude with future directions.

1. INTRODUCTION

Three-dimensional (3D) integration is a promising way to achieve high-performance ICs with more functionality and reduced die footprint. The improvement in performance and device-integration density is mainly due to the decrease in interconnect length compared to conventional 2D ICs. Today's 3D integration process is primarily based on die/wafer stacking since it does not require substantial changes to the existing fabrication flow. However, the keep-out-zone (KOZ) required for through-silicon vias (TSVs) and limitations on the die alignment precision impose limits on the device integration density that can be achieved using TSV-based 3D stacking [1, 2].

Monolithic 3D (M3D) integration is receiving considerable interest as it has the potential to achieve higher device density compared to TSV-based 3D stacking [3]. In this technology, transistors are processed tier-by-tier on the same wafer. Sequential integration of transistor tiers enables high-density vertical interconnects, known as the monolithic inter-tier vias (MIVs). Typically, the size and pitch of an MIV is one to two orders of magnitude smaller than those of a TSV [4]. Therefore, M3D integration can result in significantly reduced area and higher performance when compared to 3D die stacking.

There are three categories of M3D integration depending on the granularity of tier partitioning; transistor-level, gate-level, and block-level M3D integration. In transistor-level M3D, P-channel transistors are placed on one tier and N-channel transistors on another tier, and connection between them is established by MIVs. In gate-level M3D, standard cells are split across multiple tiers, and inter-cell connections which cross tiers utilize MIVs. Block-level M3D design, which is the coarsest-grained design style, partitions functional blocks, and inter-block MIVs are deployed to deliver signals between functional modules in different tiers.

Unfortunately, commercial EDA tools to support 3D integration are currently not available. Therefore, various studies have been reported on M3D IC design using 2D commercial tools. In [5], the authors a propose gate-level M3D implementation flow that estimates cell placement and the wire-length of a M3D design by shrinking the dimension of cells and wires. Since the shrinking is susceptible to violations on design rule check (DRC), the authors in [6] utilize the RC derating technique instead of dimensional shrinking. On the other hand, block-level M3D implementation flow, which partitions functional modules based on the architectural organization of a design, is presented in [7], In order to build a M3D design, it utilizes design partitioning along with sets of dummy wires and anchor cells to model MIVs in the 2D space.

Test challenges for M3D integration have remained largely unexplored, as highlighted at the 2015 IEEE 3D Test Workshop. They have to be identified and addressed in order to ensure that M3D is primed for use when industry is ready to adopt it. In this paper, we present two such challenges. First, in order to realize high-density MIVs, the inter-layer dielectric (ILD) thickness is being aggressively scaled [4], which leads to electrostatic coupling between device tiers. The timing of an M3D design can vary significantly due to coupling [8, 9]. Next, defects that are new to M3D (e.g., defects that arise in the ILD during the wafer-bonding step [8, 10]) have to be analyzed, and methods to test for these defects have to be developed.

The rest of the paper is organized as follows. Section 2 provides an overview of M3D integration. Section 3 describes the benefits in performance, power, and area (PPA) obtained using M3D. Section 4 presents the challenges associated with EDA of M3D ICs and the proposed solutions. An overview of test challenges is presented in Section 5. Section 6 presents the inter-tier variation challenges and solutions. Finally, Section 7 lists several open problems that need to be addressed.

2. M3D TECHNOLOGY OVERVIEW

M3D integration involves the sequential processing of transistor layers on the same wafer. First, the bottom-tier transistors and their associated interconnects are processed using a standard high temperature process, (e.g., silicon-on-insulator (SOI), FinFET, or bulk CMOS technology). Next, a thin silicon layer is created over the bottom layer. The top-layer transistors are then processed under a strict thermal budget. Finally, MIVs are processed to connect the two layers. The last three steps are repeated to fabricate additional metal layers.

Sequential integration of device layers would not have been possible without the development of: (1) a low-temperature process to create a thin silicon film over the bottom tier, and (2) a process to realize top-tier transistors without damaging the underlying

978-1-5386-3093-8/17/$31.00

interconnects and degrading the properties of the bottom-tier transistors. Several approaches have been proposed in the literature to create a thin silicon film at low temperatures: (1) thin-film polysilicon deposition (which is based on the method used to create thin-film transistors and is currently for 3D-NAND flash memories), (2) pulsed-laser crystallization of amorphous silicon, (3) excimer laser crystallization, (4) epitaxial layer transfer (ELTRAN) [11], (5) wafer bonding followed by grinding and etchback [12], and (6) modified Smart-Cut process [13] (originally used to manufacture SOI wafers). The M3D technology developed by CEA Leti, which is known as CoolCube^{TM} and has attracted the attention of companies such as IBM, ST Microelectronics, and Qualcomm, uses the last two techniques for creating a thin silicon film over the bottom tier [14].

Dopant activation, which is usually performed at temperatures higher than 800 °C, is one of the key steps in forming transistors. However, dopant activation at such high temperatures to realize top-tier transistors can damage the underlying interconnects and degrade the properties of the bottom-tier transistors. Several process flows have been proposed to perform dopant activation at low temperature: (1) solid-phase epitaxy at 600 °C along with the use of Tungsten (W) back-end-of-the-line (BEOL) in the bottom tier [12], (2) recessed channel array transistor (RCAT) process flow [15], and (3) pulsed laser annealing along with a shielding layer between the top and bottom tier [16]. The last two processes achieve dopant activation at temperatures below 400 °C; therefore, these processes are compatible with Copper (Cu)/low-k BEOL in the bottom tier.

3. FULL-CHIP PPA BENEFITS OF M3D

By placing cells on multiple tiers, M3D designs offer PPA benefits over 2D designs [5, 17, 18]. We investigate the impact of technology nodes, operating clock frequency, and the characteristic of benchmarks on PPA benefits of M3D designs.

We first investigate the PPA benefits of M3D depending on technology nodes with foundry 28 nm and 14/16 nm technology nodes as well as a predictive 7 nm node using a commercial, in-order, 32-bit application processor as a benchmark [17]. The technology model as well as the standard cell layouts are created with accurate dimensional and material parameters from previous publications, which reflect the features of FinFET devices and sub-20 nm technology nodes.

Table 1 shows the normalized key design and power metrics of 2D and M3D application processor designs in the three technology nodes. With M3D technology, standard cell area and wire-length of the designs are reduced by 9.9 % and 24.5 % on average, respectively, offering more than 10 % power saving for all three technology nodes. We also observe that the total power saving is maximum in 28 nm technology node.

To analyze the trend, we employ Equation (1) which describes the components of dynamic power of a design.

$$P_{dyn} = P_{INT} + \alpha \cdot (C_{pin} + C_{wire}) \cdot {V_{DD}}^2 \cdot f_{clk}, \quad (1)$$

where P_{INT} is internal power of the cells, and the second term describes switching power where C_{pin} and C_{wire} are the pin and wire capacitance of the design, α is the switching activity factor, f_{clk} is the operating clock frequency of the design. Owing to the reduced footprint and utilizing short vertical connections (MIVs) instead of using long wires on x-y plane, M3D technology effectively reduces wire-length of the design, which reduces C_{wire}. In addition, since the cells drive reduced wire load, the number of buffers as well as the drive-strength of cells decreases, which in turn, reduces the standard cell area. The reduced standard cell area contributes P_{INT} as well as C_{pin} reduction. Therefore, the total power reduction

Table 1: Normalized iso-performance comparison of 2D vs. M3D of a commercial application processor across technology nodes. All values are normalized to corresponding 28 nm 2D data.

	normalized 2D			M3D percentage gain over 2D (%)		
metric	28 nm	14/16 nm	7 nm	28 nm	14/16 nm	7 nm
footprint	1x1	0.6x0.6	0.4x0.4	-51.1	-50	-54.7
density	1	0.90	0.80	-10.9	-8.9	-12.3
cell count	1	1.03	1.25	-7.8	-7.3	-9.5
cell area	1	0.32	0.09	-12.6	-7.3	-9.8
wirelength	1	0.65	0.44	-23.6	-22.3	-27.5
wire cap	0.54	0.33	0.21	-23.3	-13.1	-13.2
pin cap	0.46	0.38	0.21	-16.5	-9.1	-12
total cap	**1**	**0.71**	**0.41**	**-20.2**	**-11**	**-12.6**
int pw	0.44	0.28	0.14	-11.4	-7.9	-8.6
wire sw pw	0.27	0.13	0.06	-21.8	-14	-12.7
pin sw pw	0.23	0.15	0.06	-14.9	-10.1	-11.5
lkg pw	0.06	0.00	0.00	-13.4	-5	-3.2
total pw	**1**	**0.56**	**0.26**	**-15.1**	**-9.9**	**-10.3**

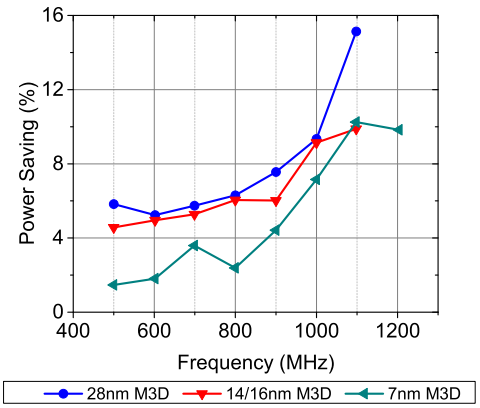

Figure 1: Power saving of the M3D application processor over 2D in foundry 28 nm, 14/16 nm, and a predictive 7 nm technology nodes.

in an M3D design depends on wire-length and standard cell area reduction, the ratio of pin versus wire capacitance, and switching power versus internal power in the 2D design.

The power saving difference from 28 nm to 14/16 nm and 7 nm technology node can be attributed to the difference in C_{pin} associated with planar and FinFET transistors. FinFET based technologies have higher C_{pin} due to the 3D fin structure and the introduction of local interconnect middle-of-line (MOL) layers. Since wire-length saving of M3D designs in all three technology nodes are much higher than standard cell area reduction, 28 nm technology benefits most from M3D as C_{wire} dominates the total capacitance of the design.

We also explore the impact of operating frequency on the PPA benefits of M3D design [17]. Figure 1 compares the power saving achieved by M3D designs implemented with different target clock frequency. As target clock frequency increases, more buffers and high drive-strength cells are required in 2D designs in order to to meet timing constraints. However, thanks to reduced wire load from wire-length saving, M3D designs utilize less number of buffers and lower drive-strength cells, offering more standard cell area saving. Therefore, power benefit of M3D technology increases with design which operates in high clock frequency by reducing P_{INT} and C_{pin}.

Table 2: Qualitative comparison of Shrunk-2D [5], Derated-2D [6] and Cascade-2D [7] flow.

property	Shrunk-2D	Derated-2D	Cascade-2D
granularity	gate-level	gate-level	block-level
key enabler	dimensional shrinking	RC derating and placement projection	design partition and wormhole model
tier partitioning	min-cut	timing slack	architects decide
MIV planning	together	together	tier-by-tier
placement optimization	together	together	together
routing optimization	tier-by-tier	together	together

The characteristic of a benchmark is another factor which affects the PPA benefits of M3D designs [18]. We use AES-128, FFT, and LDPC as the benchmarks, and observe huge difference in power saving between LDPC and other benchmarks. The M3D AES-128 and FFT designs implemented with high-performance 7 nm cell library show 12.7 % and 8.8 % power savings at their maximum frequency, respectively, whereas LDPC design offers 26.5 % power reduction. This difference is attributed to the ratio of the switching power (i.e., the second term in Equation 1) to P_{INT} of the designs. The ratio is much higher in LDPC 2D design (4.27) than AES-128 and FFT 2D designs (0.84 and 0.54, respectively) since LDPC is a wire-dominated circuit whereas other benchmarks are cell-dominated circuits. Again, based on the fact that wire-length saving is normally higher than standard cell area reduction, a wire-dominated circuit gets more power benefit than cell-dominated circuits from M3D technology.

4. FULL-CHIP RTL-TO-GDS FLOW

Since no commercial EDA tool that supports cell placement in 3D space is currently available, several M3D design methodologies which utilize 2D commercial EDA tools to implement M3D designs are proposed. In [5, 6], the authors propose the design flows for gate-level M3D designs, whereas [7] presents block-level M3D design flow. Table 2 compares the flow of the three M3D implementation methodologies.

4.1 Shrunk-2D Flow

To enable 3D cell placement, this flow uses dimensional shrinking methodology assuming that the z dimension is negligibly small. This methodology starts with scaling the x-y dimensions of all cells and metal layers by a factor of $(1/\sqrt{2})$. With shrunk cells, a *Shrunk-2D design* is implemented in the half footprint compared to the 2D counterpart. Placement optimization is performed during the *Shrunk-2D design* implementation, resulting in x-y locations cells in the final M3D design.

The cell instances in the resulting *Shrunk-2D design* are then scaled up to the original size, creating overlaps among cells. Then, with an area-balanced min-cut partitioning algorithm, cell instances are partitioned on the two tiers (i.e., determine z location of cells), so that the number of connection (i.e., MIVs) between two tiers is minimized while balancing the cell area of top and bottom tiers.

In order to determine the location of MIVs, the original metal stack is duplicated, creating full metal stack (i.e., both metal stack on top and bottom tier) of the M3D design. Additionally, the cells are annotated with their respective tiers, so that their pins are located on metal layers of their corresponding tier. The design is routed with the full metal stacks using EDA tools. The location of MIVs are determined by the location of vias between the top metal layer of the bottom tier and the bottom metal layer of the top tier.

Then, using the netlists of each tier and MIV location, we perform trial-route, and the initial routed design is used to obtain timing constraints for each tier. Once the timing constraints are determined, timing-driven routing is performed for each tier, resulting in final M3D design.

4.2 Derated-2D Flow

Since shrinking cell involved in Shrunk-2D M3D design flow is prone to DRC violation, the M3D design flow presented in [6] uses derated resistance and capacitance of wires rather than using shrunk cells for the 3D placement. While using original geometry of cells and wires, it starts with implementing a *Derated-2D design* in the same footprint with the 2D counterpart. The footprint will be reduced by half in later stage of the flow which will be described shortly. Resistance and capacitance of wires are derated by $1/\sqrt{2}$ during *Derated-2D design* implementation, so that the EDA tool incorporates the reduced wire load resulted from the reduced footprint in final M3D while placement optimization.

The cell placement of the resulting *Derated-2D design* is then projected into the footprint of the final M3D design, and partitioned onto two tiers. Since every Manhattan distance between each cell is scaled by $1/\sqrt{2}$ as a result of placement projection, the RC parasitic of the *Derated-2D design* is expected to be the similar to that of the M3D design.

The rest of the steps, including MIV planning and the routing of each tier, are similar to Shrunk-2D design flow, but there are two key differences: (1) Derated-2D M3D design flow optimizes routing optimization by cell resizing and V_T swapping, and (2) supports inter-tier degradation which will be described in the next section in detail.

4.3 Cascade-2D Flow

Unlike the two design flows described above that implement gate-level M3D designs, the M3D design flow in [7] partitions functional blocks in a design into two tiers, implementing block-level M3D designs.

The flow begins by dividing RTL into two groups, top and bottom group, whose elements are functional blocks to be placed on the top and bottom tier (i.e., determine z-location), respectively. The RTL partition is performed in two ways: (1) A pair of modules with tight timing budget is identified based on architectural organization of a design, and assigned to different tiers, so that they benefit from short vertical connection of MIVs. (2) In absence of understanding of architecture of a design, the number of timing paths between a pair of modules is used. We assign modules with high number of timing paths between them onto different tiers, so that the distance between two modules is minimized by placing on top of each other.

The location of the MIVs is determined by pre-designing the top group and bottom group separately in the half footprint of 2D counterpart, placing MIVs near the location of their driving cells, so that wire-length between MIV ports and relevant cells are minimized.

Once the location of MIVs are determined, a *Cascade-2D design* is implemented, which mimics M3D design in a single 2D design using design partition technique and sets of dummy wires and anchor cells to model MIVs. A new die with both tiers placed side-by-side is created with the same total area as the original 2D design. After defining top and bottom partitions on each side of the die, MIV pins are placed on the top metal layer of each partition at the location derived in the previous step.

Using additional metal layers above the top metal layer used

in actual design, MIV pin pairs in the top and bottom partition is routed, creating 'dummy wires', whose function is only to establish logical connection between the two partitions in the physical design without delay and parasitics. Note that since MIV pins are placed on the top metal layer in both partitions, connection between the top metal layer and the bottom metal layer of the top tier is required in order to exactly model MIVs, which connect the top metal layer of the bottom tier and the bottom metal layer of the top tier. This is achieved by 'anchor cells', which model zero-delay virtual connection between a MIV pin and one of the metal layers. After routing MIV pins on the two partition, anchor cells are placed below the MIV pins. These sets of anchor cells and dummy wires effectively act as 'wormholes' which connect the bottom metal layer of the top partition and top metal layer of the bottom partition without delay emulating the behavior of MIVs.

Then we run regular P&R flow, performing placement and routing optimization of both tier simultaneously in a single design, resulting in final M3D design in the single design.

5. TEST CHALLENGES

We present two test challenges for M3D integration: (1) timing variations due to high-density integration and (2) defect analysis and modeling. For each test challenge, we need to analyze and quantify its impact on an M3D IC. For example, in [8, 9, 10], the impact of electrostatic coupling and wafer-bonding defects on delay testing of an M3D IC is analyzed and quantified using detailed simulations. Next, we describe these challenges and quantify their impact on the effectiveness of delay-test patterns. Such an analysis will provide insights into the test methods required to screen defective and marginal M3D ICs.

5.1 Timing Variations

Aggressive scaling of ILD thickness can lead to coupling between two device tiers in an M3D design and result in timing variations. Due to coupling, the threshold voltage of a transistor in the top tier can differ considerably from that of a conventional transistor (SOI or bulk CMOS), and depends on the bias and properties of the back gate. We carried out device simulations to accurately evaluate the impact of coupling on the threshold voltage of a top-tier transistor. Simulations were performed using Silvaco Atlas, a physics-based numerical device simulator.

We simulated two device structures using Atlas: (1) an asymmetric double-gate SOI transistor to study the impact of coupling in transistor-level M3D integration, and (2) an asymmetric double-gate SOI transistor in which the front- and back-gates are independent of each other to study the impact of coupling in gate-level M3D integration. The values of various parameters for the devices in our simulations were obtained from [19]. In our simulations, we have included physical models such as Shockley-Read-Hall recombination, Lombardi's mobility, and Selberherr's impact ionization. To get realistic results, we calibrated the model parameters using experimental values of on- and off-currents reported in [19].

We analyze the change in threshold voltage ($|\Delta V_{th}|$) from that of a conventional SOI transistor due to coupling. The threshold voltage values are calculated from $I_D - V_{FG}$ plots using the constant current threshold voltage extraction method [20] for a drain-source bias of 100 mV. In this method, threshold voltage of a device is the gate-source voltage at which the drain current is equal to a constant current (I_{D0}) multiplied by the ratio of gate width to gate length. The value of I_{D0} is process dependent, and it typically ranges between 50 and 500 nA. In this paper, the value of I_{D0} was chosen to be 300 nA for N-channel and 70 nA for P-channel transistors. These values for obtained from fab measurements for 32 nm partially-depleted SOI technology [20].

Figure 2(a) shows the $|\Delta V_{th}|$ of a double-gated device created to simulate a top-tier transistor in the case of transistor-level M3D integration. For both N-channel and P-channel double-gate transistors, the $|\Delta V_{th}|$ was obtained to be 51 mV for an ILD thickness of 23 nm. However, for an ILD thickness of 100 nm, the $|\Delta V_{th}|$ reduces to 7 mV and 9 mV for N-channel and P-channel transistors, respectively. Therefore, we conclude that for ILD thicknesses greater than 100 nm, the impact of the bottom tier on threshold voltage of a top-tier transistor in transistor-level M3D is minimal. For ILD thickness less than 50 nm, the impact of coupling on threshold voltage of a top-tier transistor is significant, and cannot be ignored.

Figure 2(b) shows the $|\Delta V_{th}|$ of a double-gated device created to simulate a top-tier transistor in the case of gate-level M3D integration. For the N-channel double-gate transistor, the $|\Delta V_{th}|$ was obtained to be 44 mV and 65 mV for a back-gate (metal line from the bottom tier) voltage of 0V and 1V, respectively. However, for an ILD thickness of 100 nm, the $|\Delta V_{th}|$ reduces to 9 mV and 13 mV for a back-gate voltage of 0 V and 1 V, respectively. Similarly, for the P-channel double-gate transistor, we observe that the impact of a metal line from the bottom tier on the threshold voltage of a top-tier transistor is minimal for ILD thickness greater than 100 nm. Therefore, we conclude that for ILD thickness greater than 100 nm, coupling between the top and bottom tiers is minimal. For ILD thickness less than 50 nm, the impact of coupling on threshold voltage of a top-tier transistor is significant, and cannot be ignored.

5.2 Defect Analysis and Modeling

Defects that are new to M3D have to be analyzed and modeled for yield learning. For example, defects that arise during the wafer-bonding step can impact the top-tier transistors, as well as the MIVs. The condition of the bonding surfaces plays a crucial role in achieving a defect-free bond during the wafer-bonding step in the fabrication of an M3D design. Oxide layers, which are the prime candidates for bonding surfaces, act as the ILD. It is therefore important to understand and analyze wafer-bonding defects, and develop methods to test for these defects.

The defects during the wafer-bonding process occur at the bond interface. Defects can often be attributed to voids, delaminations, and foreign particles [21]. The size of these defects can be in the mm to nm range. The major factors that affect the quality of the bond can be summarized as follows: (1) the root-mean-square value of surface roughness of the wafers to be bonded must be less than 0.5 nm; (2) the total thickness variation of the bonding surfaces needs to be less than 5 μm; (3) the warpage of the wafers prior to bonding has to be less than 50 μm; (4) accidental particle contamination or process-related residuals (e.g., from the CMP step), can degrade the bond strength and result in large unbonded areas; (5) equipment wear-out can lead to a variation in the initiation of contact between two wafers and the bond initiation force. Initiation of contact at more than one point will lead to the propagation of multiple bonds fronts, thereby resulting in voids.

The presence of a void at the bond interface impacts the threshold voltage of top-tier transistors. If a void is present in the back-gate dielectric of a double-gate transistor, the back-gate capacitance will be lower compared to the defect-free case, since the dielectric constant of air is lower than that of a dielectric material. The effective back-gate dielectric capacitance can be expressed as the capacitance of the void in series with the capacitance of the dielectric layers above and below the void:

$$\frac{1}{C_{effective}} = \frac{t'_{box}}{\epsilon_{ox}} + \frac{t_{void}}{\epsilon_{void}} + \frac{t''_{box}}{\epsilon_{ox}} \quad (2)$$

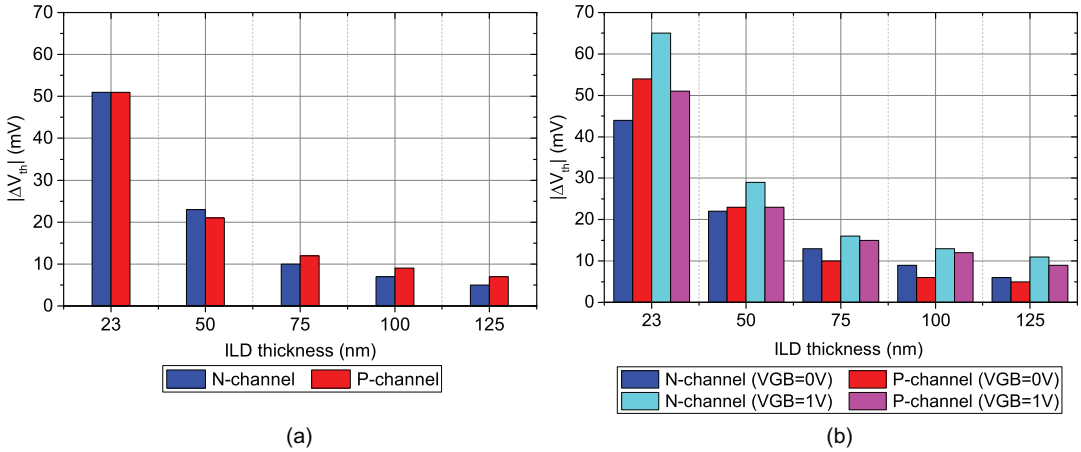

Figure 2: Threshold voltage change from that of a conventional SOI transistor for: (a) transistor-level M3D, and (b) gate-level M3D.

where t_{box} is the thickness of the back-gate dielectric in the absence of a void, t'_{box} and t''_{box} are the thicknesses of the dielectric layers above and below the void, respectively, t'_{void} is the thickness of the void, and ϵ_{ox} and ϵ_{air} are the dielectric constants of the ILD and void, respectively. From (2), we can conclude that the effective back-gate dielectric capacitance ($C_{effective}$) will be less than the capacitance of the back-gate dielectric in the absence of a void (t_{box}/ϵ_{ox}) since $\epsilon_{void} < \epsilon_{ox}$.

5.3 Impact on Delay-Test Patterns

Timing-aware automatic test pattern generation (ATPG) tools can be used to target the longest sensitizable path for each delay fault. Such ATPG tools rely on circuit timing information to generate path profiles. Electrostatic coupling and wafer-bonding defects significantly impact the path profiles in a gate-level-integrated M3D IC. If these changes are not considered, the ATPG tool will propagate faults through paths that are much shorter than the long paths. This will decrease the effectiveness of patterns used to screen delay defects, resulting in lower test quality.

We analyze the impact on the effectiveness of delay-test patterns for an IWLS'05 benchmark. This benchmark was synthesized and scan-inserted using the Nangate open cell library for the 45 nm technology node. We assumed that each benchmark design constitutes the top layer in an M3D IC. In order to carry out this analysis, we developed a tool flow based on conventional library characterization, timing analysis, and ATPG tools: Synopsys SiliconSmart, Synopsys PrimeTime, and Synopsys TetraMax, respectively.

We obtained the SDQL values for 50 instances of vga_lcd. The results are presented in Figure 3. We normalized these SDQL values with respect to the SDQL value obtained using the nominal timing library for the standard cells in the benchmark. For most defective M3D instances of these benchmarks, we observe that the SDQL values obtained are greater than those obtained using the nominal timing library. Such an outcome is possible if there is an increase in the range of defect sizes for most faults in the benchmark that are not tested by test patterns generated using the nominal timing library. For each fault, the range of untested defect size is defined as the difference between the slack on the longest-sensitizable path and the slack on the path being sensitized by the test patterns. These results show that coupling and wafer-bonding defects increase the range and size of defects that are not tested by timing-aware ATPG patterns, and therefore, the likelihood of delay-defect test escapes. For some circuit instances, we also observe a decrease in SDQL. Such an outcome is possible if for most

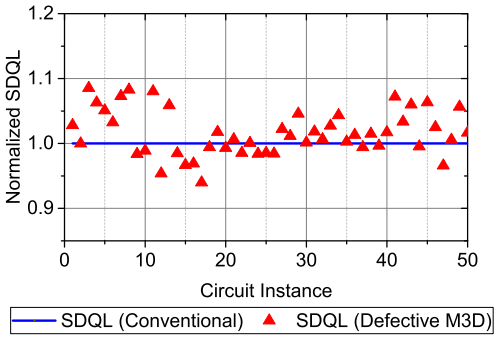

Figure 3: Normalized SDQL values for vga_lcd.

faults in the benchmark, the range of untested defect size decreases. For a fault, the range of untested defect size decreases if the slack on the path being sensitized by the test pattern decreases.

6. INTER-TIER VARIATION ISSUES

One of the most challenging issues with M3D fabrication is the limited thermal budget during top-tier fabrication process. Since Cu wire is able to withstand up to 450 °C, fabricating top-tier cells with normal process with more than 800 °C might impair the wires on the bottom tier. To prevent this, either low temperature process for fabricating top-tier cells or W or Cobalt (Co) interconnect which endures higher temperature are utilized for the bottom-tier wires. However, low temperature process degrades the performance of the top-tier cells, and W or Co interconnect on the bottom tier not only degrades overall performance due to their high resistivity.

While the method presented in [5] assumes same performance between top and bottom-tier cells, new methodologies are proposed to incorporate the top-tier cell degradation as well as high resistivity of wires in the bottom tier [6, 22].

In [6], the authors use cell-slack sorting technique during the tier partitioning step and degradation aware MIV planning to resolve the inter-tier variation challenges. During tier partitioning step, it first extract slack of cells in the Derated-2D design, which represents the timing criticality of each cell. Cells with small cell slack, which indicates that the cells have tight timing budget, is assigned to the bottom tier to avoid top-tier cell degradation, and cells with high cell slack is assigned to the top tier since they have enough timing budget, which can compensate the degradation.

While MIV planning, the flow utilizes a full metal stack which contains all metal layers in both top and bottom tier, so that it can perform routing optimization by cell resizing and V_T swapping, considering the the material difference between the top and bottom tier (i.e., W or Co wires in the bottom tier and Cu interconnect in the top tier). With the full metal stack, timing-driven routing is performed, so that the design utilizes top metal layers more than bottom metal layers, and makes it possible to minimize the timing degradation due to high resistivity of the bottom-tier metal layers.

On the other hand, the authors in [22] use path-based tier partitioning method to tackle the bottom-tier degradation due to W interconnect using the fact that slack distribution of timing paths in a design is wide, and there are many timing paths which have high slack. By forcing cells in the timing paths with tight timing budget in the top tier while constraining area skew and the number of MIVs, they ensure that the timing critical paths are not affected by the high resistance interconnects.

The study also propose a net-based partitioning method to reduce the metal layer usage in the bottom tier, so that minimize the cost of a design. Cells connected to long wires are confined on the top tier during the tier partitioning, and the other shorter nets use the bottom-tier metal layers resolving congestion on the bottom tier. In this way, the routing resource usage of the bottom tier is minimized, allowing room for reducing the number of metal layers on the bottom tier.

7. OPEN PROBLEMS

Thermal is one of the well known challenges in M3D integration. Although M3D designs have lower power consumption compared to 2D, the increased power density due to reduced footprint worsens the temperature of M3D designs. Unlike TSV-based 3D ICs, M3D utilizes extremely thin Inter Layer Dielectrics (ILD). Thus, the lateral thermal conductivity is extremely small, resulting in considerable thermal coupling between tiers. Some initial effort is made to optimize the floorplan of M3D designs [23], but more aggressive thermal-aware design methods and cooling methodologies are required.

Power delivery remains a challenge in M3D simply due to the fact that we need to serve more devices in all tiers with fewer C4 bumps on the bottom. Studies have shown that the tier that is farther away from C4s suffer more IR-drop and thus requires more resources [24]. Unfortunately, M3D is already crowded with extremely dense inter-tier connections. Thus, aggressive strategies are called for to strike a balance between IR-drop, routing congestion, and PPA overhead.

Research has shown [25] that the more clock TSVs used, the more wirelength and thus power consumption saving is achieved in TSV-based 3D ICs. The same argument holds in M3D. However, new issues in M3D clock distribution network (CDN) arises: clock MIV is extremely small thus under the danger of manufacturing defects. Besides, due to the closer proximity between the devices in M3D, inter-tier coupling may degrade the quality of clock signals delivered to FFs. Research is required to address these reliability and low power issues in M3D.

M3D suffers from insufficient understanding of test issues and the lack of design-for-test (DfT) solutions. We must develop effective tests that target new defects in M3D. In addition, significant rethinking in test generation is needed in order to screen delay defects in the presence of inter-tier timing variations that area unique in M3D. In order to enable defect isolation and yield enhancement, low-cost DfT solutions also need to be developed for M3D ICs. DfT solutions for TSV-based 3D ICs are ineffective for M3D simply because the number of MIVs is expected to be orders of magnitude higher than TSVs [4].

8. REFERENCES

[1] S. Kannan *et al.*, "Device performance analysis on 20nm technology thin wafers in a 3D package," in *IRPS*, April 2015, pp. 4C.4.1–4C.4.5.

[2] International Technology Roadmap for Semiconductors, "http://www.itrs.net".

[3] K. Arabi *et al.*, "3D VLSI: A Scalable Integration Beyond 2D," in *ISPD*, 2015, pp. 1–7.

[4] P. Batude *et al.*, "3-D Sequential Integration: A Key Enabling Technology for Heterogeneous Co-Integration of New Function With CMOS," *JETCAS*, vol. 2, no. 4, pp. 714–722, Dec 2012.

[5] S. A. 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.

[6] B. W. Ku *et al.*, "Physical Design Solutions to Tackle FEOL/BEOL Degradation in Gate-level Monolithic 3D ICs," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2016.

[7] 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.

[8] A. Koneru *et al.*, "Impact of Electrostatic Coupling and Wafer-Bonding Defects on Delay Testing of Monolithic 3D Integrated Circuits," *JETC*, vol. 13, no. 4, pp. 54:1–54:23, July 2017.

[9] A. Koneru and K. Chakrabarty, "Analysis of electrostatic coupling in monolithic 3D integrated circuits and its impact on delay testing," in *ETS*, May 2016, pp. 1–6.

[10] A. Koneru *et al.*, "Impact of wafer-bonding defects on Monolithic 3D integrated circuits," in *EPEPS*, Oct 2016, pp. 91–94.

[11] T. Yonehara, "Epitaxial layer transfer technology and application," in *S3S*, Oct 2015, pp. 1–3.

[12] L. Brunet *et al.*, "(Invited) Direct Bonding: A Key Enabler for 3D Monolithic Integration," *ECS Transactions*, vol. 64, no. 5, pp. 381–390, 2014.

[13] I. Radu *et al.*, "Novel low temperature 3D wafer stacking technology for high density device integration," in *ESSDERC*, Sept 2013, pp. 151–154.

[14] P. Batude *et al.*, "3DVLSI with CoolCube process: An alternative path to scaling," in *VLSIT*, June 2015, pp. T48–T49.

[15] D. C. Sekar and Z. Or-Bach, "Monolithic 3D-ICs with single crystal silicon layers," in *3DIC*, Jan 2012, pp. 1–2.

[16] B. Rajendran *et al.*, "Pulsed laser annealing: A scalable and practical technology for monolithic 3d ic," in *3DIC*, Oct 2013, pp. 1–5.

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

[18] ——, "Impact and Design Guideline of Monolithic 3-D IC at the 7-nm Technology Node," *IEEE Trans. on VLSI Systems*, vol. 25, no. 7, pp. 2118–2129, July 2017.

[19] P. Batude *et al.*, "Advances, challenges and opportunities in 3D CMOS sequential integration," in *IEDM*, Dec 2011, pp. 7.3.1–7.3.4.

[20] A. L. Loke *et al.*, "Constant-current threshold voltage extraction in HSPICE for nanoscale CMOS analog design," in *Synopsys Users Group Conference*, 2010, pp. 1–19.

[21] J. Rubin *et al.*, "Essential edge protection techniques for successful multi-wafer stacking," in *S3S*, Oct 2015, pp. 1–2.

[22] 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.

[23] ——, "Fast and accurate thermal modeling and optimization for monolithic 3D ICs," in *Proc. ACM Design Automation Conf.*, 2014.

[24] ——, "Full chip impact study of power delivery network designs in monolithic 3D ICs," in *Proc. IEEE Int. Conf. on Computer-Aided Design*, 2014.

[25] X. Zhao *et al.*, "Low-Power and Reliable Clock Network Design for Through Silicon Via based 3D ICs," *IEEE Trans. on Components, Packaging, and Manufacturing Technology*, vol. 1, no. 2, 2011.