Transistor-Level Monolithic 3D Standard Cell Layout Optimization for Full-Chip Static Power Integrity

Bon Woong Ku[†], Taigon Song[§], Arthur Nieuwoudt[§], and Sung Kyu Lim[†]
[†]School of ECE, Georgia Institute of Technology, Atlanta, GA, USA
[§]Synopsys Inc., Mountain View, CA, USA
bwku@gatech.edu, taigon.song@synopsys.com, arthur.nieuwoudt@synopsys.com, limsk@ece.gatech.edu

***Abstract*—Existing transistor-level monolithic 3D (T-M3D) standard cell layouts are based on the folding scheme, in which the pull-down network is simply folded and placed on top of the pull-up network. In this paper, we propose a new layout method, the stitching scheme, targeted towards improved cell performance and power integrity. We perform extensive analysis on each layout scheme and evaluate the timing/power benefits of the stitching scheme. Since the ground and power rails overlap in the T-M3D layouts with the folding scheme, we also present a design methodology for the power delivery network of folding T-M3D ICs to evaluate the impact of the T-M3D cell layout scheme on static power integrity. Compared to 2D ICs at iso-performance, stitching T-M3D ICs show a maximum of 6% power savings, 44% area savings with only 1% more static IR-drop in the 14nm technology node while folding T-M3D ICs undergo serious degradation in static power integrity, causing a reliability issue.**

I. Introduction

Monolithic 3D (M3D) technology, which involves the integration of one or more active layers on top of a prefabricated metal stack in monolithic fashion, is a promising solution in overcoming the physical limitations of logic scaling in advanced technology nodes. The sequential integration of transistors on multiple tiers removes the μm-scale die alignment issue inherent in the through-silicon-via (TSV)-based 3D integration. In addition, it adopts the high precision of advanced lithography technology in the fabrication of *nm*-scale monolithic inter-tier vias (MIVs) for tier-by-tier interconnections [1]. The extremely small size of MIVs minimizes the area and parasitic overhead of 3D routing, allowing us to insert them into any design space. Depending on the granularity of MIV insertion, M3D designs are categorized into the block level (B-M3D), the gate level (G-M3D), and the transistor level (T-M3D). While B-M3D and G-M3D use MIVs to route the 3D nets of blocks or gates placed on multiple tiers, T-M3D uses ultra-dense MIVs inside standard cell designs to connect transistors on separate tiers [2], [3].

One of the major challenges in adopting M3D technology is the high manufacturing cost of multiple device layers and metal stacks [4]. To reduce the cost of M3D wafers, industry must reduce the number of metal layers to the maximum allowable extent. While B-M3D and G-M3D need global and local metal resources for both tiers, T-M3D requires only local interconnects on the bottom tier. Therefore, T-M3D is the most favorable design of the three from an economic perspective.

Another challenge is to maintain the performance of devices and the integrity of interconnects on the bottom tier during the fabrication process of the top tier. To minimize the impact of post-thermal exposure on the bottom tier, the maximum manufacturing thermal budget for the top tier is constrained under 450 °C [5]. In T-M3D, implementing thermally stable devices on the bottom tier is an additional option to controlling variation since T-M3D allows us to place NMOS and PMOS on separate tiers. According to experimental results in [6], [5], the active sheet resistance of NMOS is more susceptible to post-thermal exposure than that of PMOS because of doping deactivation in high concentrations. As a result, existing studies [2], [3] place the pull-down network (PDN) on top of the pull-up network (PUN) in T-M3D standard cell designs.

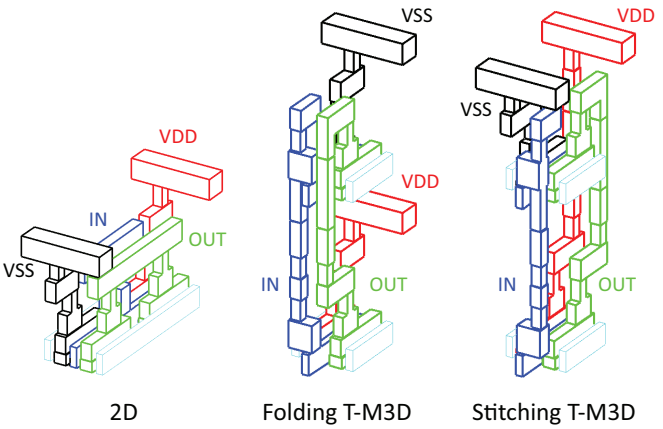

Fig. 1. 2D and T-M3D INV_X1 cell layouts. To create a T-M3D layout, the folding scheme simply folds the pull-down network and places it on top of the pull-up network while retaining the routing topology of a 2D layout. However, The stitching scheme utilizes two MIV tracks on the top and bottom sides of the layout underneath the power and ground rails. This optimizes internal parasitics and improves static power integrity by exposing both VDD and VSS rails to the back end of line directly.

In [2], T-M3D standard cell layouts of the 45nm technology have been designed based on the folding scheme, which retains the same routing topology as 2D layouts while simply folding and placing the PDN on top of the PUN, shown in Figure 1. Although the folding scheme reduces the effort at creating T-M3D standard cell layouts, it leaves a huge margin for layout optimization in T-M3D standard cells. In addition, the ground rail overlaps the power rail, resulting in restricted power connections to each standard cell. More recent studies [3] show the impact of optimized local interconnections in the folding T-M3D cell layouts using the 14nm technology by eliminating dummy poly regions that were originally required in 2D layouts, but they still do not address the power delivery issue.

In this paper, we propose a new T-M3D standard cell layout optimization method, the stitching scheme, targeted towards improvements in both cell performance and static power integrity. While folding T-M3D cells use only one MIV channel reserved for 3D routing at the bottom side of layouts, the stitching scheme allows two MIV channels at the top and bottom sides at the expense of a small extension of the cell height, and exposes the power and ground rails on the top tier over the MIV tracks. Based on extensive analysis with our 14nm T-M3D technology process design kit, we show that

978-1-5090-6023-8/17/$31.00 ©2017 IEEE

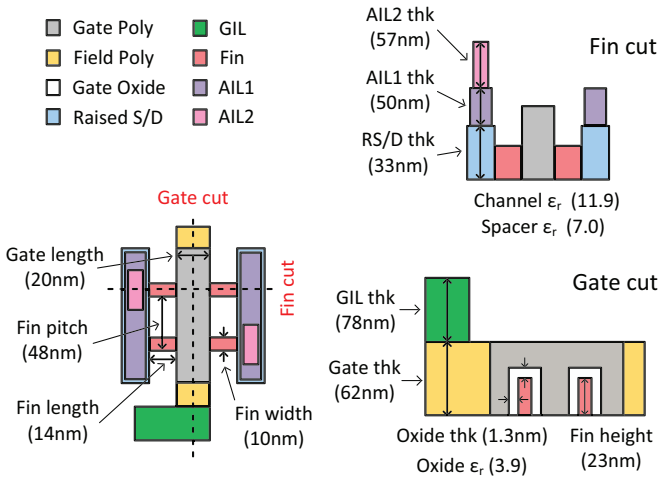

Fig. 2. Technology parameters customized from [7], [10], [12], [11], [8], [9], [13] for the device region of the 14nm T-M3D technology used in this work. GIL stands for a gate interconnect layer and AIL for a active interconnect layer that correspond to the middle-of-line (MOL) layers from [9]. The contacted poly pitch is 80nm, and the sheet resistances of a gate poly and a raised source/drain are 11.0 Ω/sq and 13.0 Ω/sq, respectively. The resistivity of each MOL layer is 0.07 $\Omega \cdot \mu m$, and the supply voltage is 0.8V. The same parameters are assumed for both top and bottom tiers. Thk in the figure indicates the thickness.

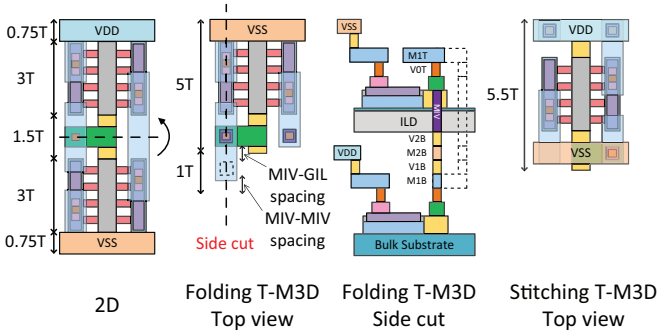

Fig. 3. An example of the folding and stitching T-M3D INV_X1 cell layout to show the requirement of an additional assumption on the MIV layer in the 14nm T-M3D technology. If an MIV makes a connection only between an M2B and an M1T layer, we must extend the height of T-M3D standard cells, resulting in the degradation of area and performance savings in T-M3D cells. Therefore, in our 14nm T-M3D technology, we assume that an MIV can be fabricated underneatch a GILT layer, in which MIVs directly connect an M2B and the GILT layer while not penetrating the active regions of the top tier.

stitching T-M3D cells reduces the wirelength and parasitics for local interconnection because of their two MIV channels. Also, we present that the stitching scheme addresses the intrinsic static power integrity issue of the folding scheme at the full-chip level.

II. 14nm T-M3D Technology Development

This section presents the development of our 14nm T-M3D technology process design kit. Figure 2 shows technology parameters for the device region of the 14nm T-M3D technology used in this work. Since an open-source T-M3D process design kit is not available to the academic community, we customize parameter values derived from 14/16nm Predictive Technology Model (ASU-PTM-MG-HP) [7], 2013 International Technology Roadmap for Semiconductors (ITRS) interconnect technology report [8], NCSU FreePDK15 [9], and foundry information available from the public domain [10], [11], [12].

A. Technology Assumption

We narrow the scope of our research down to the impact of layout optimization in two-tier T-M3D standard cells. Therefore, we assume that device models for the top and bottom tier of T-M3D cells are equivalent to the model of 2D cells. We use 14nm ASU-PTM-MG-HP [7] for the transistor model and employ the middle-of-line (MOL) structure from NCSU FreePDK15 [9] for both top and bottom tiers. The MOL structure has a gate interconnect layer (GIL) for the gate contact, an active interconnect layer-1 (AIL1) for the connection between individual fins of a device, and an active interconnect layer-2 (AIL2) between a AIL1 and a Metal1 (M1) [9]. Both the GIL and AIL2 layers have a connection to the M1 layer through the Via0 layer (V0). The thickness of each MOL layer is modified from the original structure in FreePDK15 to reflect the device parameters of our 14nm T-M3D technology, but the total height of the MOL structure is retained. We provide T-M3D cells with two local routing layers on the bottom tier (M1B, M2B) and the top tier (M1T, M2T). The M1 pitch is 80nm, which is the same as the contacted poly pitch (CPP), and the M2 pitch is 64nm. Both M1 and M2 are 60nm thick, and the aspect ratio is 1.875. The low-K dielectric constant ($\varepsilon_r = 2.55$) and the conductor resistivity ($\rho = 0.0451\Omega \cdot \mu m$) are both derived from [8].

Since the inter-layer dielectric (ILD) should electrically separate device regions on the top tier from closely spaced interconnect lines on the bottom tier, we assume that the thickness of the ILD layer is 100nm to prevent the threshold voltage of devices on the top tier from changing over 5% [13].

An MIV should make a connection between the top and bottom tier without significant area overhead. Previous studies [2] used the 45nm planar CMOS technology, which allows the insertion of MIVs without increasing the height of standard cells since PMOS is larger than NMOS. However, the 14nm FinFET CMOS technology requires the same number of fins in PMOS and NMOS, and the areas of the devices are also the same. If an MIV makes a connection only between an M2B and an M1T layer, we must extend the height of T-M3D standard cells because of the spacing rule between the MIV and the GIL layer on the top tier (GILT), or between the MIV and MIVs in neighboring cells. An example of the folding T-M3D layout of the INV_X1 cell in Figure 3 clearly demonstrates this problem. If we assume that the minimum width and the spacing of an MIV is 32nm, the height of T-M3D cells should be six metal tracks (6T) high, resulting in only 33% of area savings in T-M3D cells over 9T 2D layouts. Moreover, the wirelength of local interconnects is lengthened because 3D routing detours the device region on the top tier, degrading the performance and power savings of T-M3D cells. Therefore, we assume that an MIV can be fabricated underneath a GILT layer, in which MIVs directly connect an M2B and the GILT layer while not penetrating the active regions of the top tier. Now, we have two MIVs in our 14nm T-M3D technology. One is an MIVM that connects an M2B and an M1T, and the other is an MIVG that connects an M2B and a GILT. With these two MIVs, folding T-M3D cells achieve 44% of area savings over 9T 2D cells. Taking the ILD thickness (100nm) into account, the aspect ratio of an MIVG is 5, and 7.5 for an MIVM.

B. Technology Characterization

Characterization flow starts from creating a technology file (.tf) that defines every layer in the developed T-M3D technology. Based

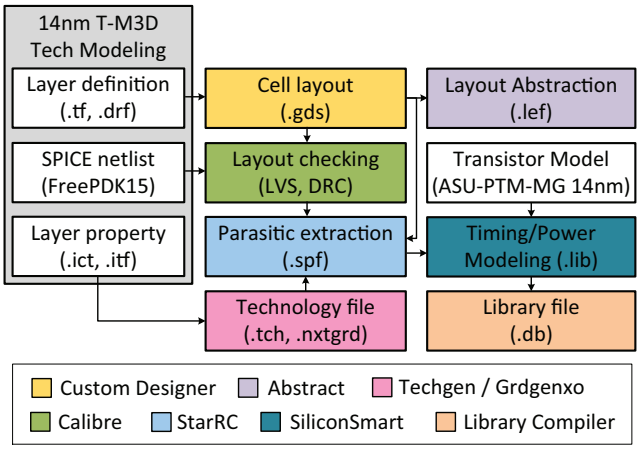

Fig. 4. PDK generation flow of our 14nm T-M3D technology.

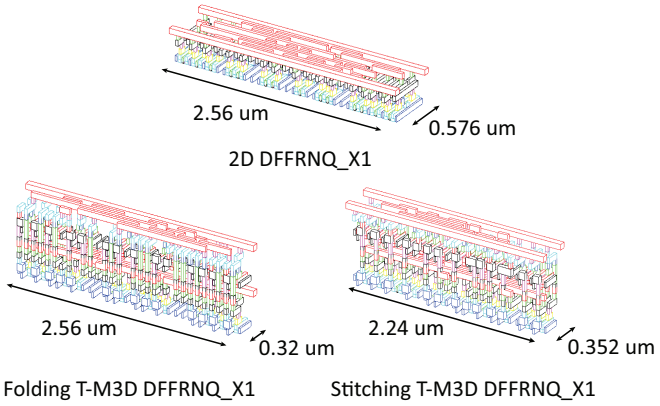

Fig. 5. An example of a DFF cell with 2D, folding T-M3D, stitching T-M3D layouts. Compared with the 2D and the folding T-M3D counterpart, well-designed stitching T-M3D D flip-flop (DFF) layout reduces the cell width by 320*nm* (12.5%), resulting in 4% of more layout footprint savings over folding T-M3D DFF, and 52% of savings over 2D DFF.

on the layer definition of technology file, we create T-M3D standard cell layouts in the GDS format, and carry out LVS and DRC check with rule files compatible with our 14nm T-M3D technology. The LVS and DRC rule files are modified from FreePDK15 [9], and DRC check is done on the top and the bottom tier independently. Next, we prepare interconnect files (.ict, .itf) that describe all technology parameters presented in the previous section, and then transform these files into the parasitic database (.tch, .nxtgrd). Since T-M3D layout architecture presents vertically coupled situations that are unpredicted in the 2D extraction rule, we leverage a field-solver engine in the commercial tool to accurately calculate the parasitics of T-M3D standard cell layouts. We also utilize FinFET modeling capability of the commercial extraction engine in this step. However, since the tool is designed for 2D ICs, FinFET modeling for the top tier device is not available. Therefore, we manually ignore the double-counted internal parasitics of device layers that the transistor model already contains. The extraction results generate SPICE netlists with all the parasitics (.spf), and the netlist is used for modeling timing/power of T-M3D cells (.lib, .db). Finally, we abstract the cell layout in the layout exchange format (.lef), and complete our 14nm T-M3D technology PDK generation flow. Figure 4 summarizes the overall generation flow.

C. 14nm 2D Standard Cell Library

Although 15*nm* 2D Open Cell Library (Nangate 15) [14] is available, cell layouts are twelve metal tracks (12T) high. The number of fins for each transistor is seven, and the CPP is 64*nm*, which is not compatible with the industry standard [3], [11], [15]. We create our own 14*nm* 2D standard cell library targeting 9T cell height with four fins per device. This makes them compatible with the industry-grade 2D layouts which use the CPP as 80*nm* and the metal pitch as 64*nm*. The template-based unidirectional routing scheme [16] is used in the cell layouts. We use 11 routing templates to create our 41 standard cells: (INV, BUF)_X(1, 2, 4, 8), (NAND, NOR, AND, OR) (2, 3, 4)_X(1, 2), (AOI, OAI) (21, 22)_X(1, 2), DFFRNQ_X1. Our 2D layouts are composed of 3T for each device region, 0.75T for the power / ground rails, and 1.5T for the gate contact region.

III. Impact of T-M3D Cell Layout Scheme

This section analyzes advantages of the stitching scheme over the folding scheme in the T-M3D standard cell layouts.

A. Footprint Analysis

While the folding scheme provides T-M3D cells with only one MIV channel at the bottom side of the layouts, the stitching scheme allows two MIV channels at the top and the bottom edge of the layouts. Adding one more MIV channel in the stitching T-M3D cell layouts leads to the extension of cell height by 0.5T compared with folding T-M3D cells. This is under the assumption that we have to honor the minimum width and spacing rule of the MIV layer to prevent overlap with neighboring cells and for reliable MIV fabrication. As a result, folding and stitching T-M3D cells are 5T and 5.5T high, turning into 44.4% and 38.8% of the height of 2D 9T layouts respectively.

Although stitching T-M3D cells are taller than folding T-M3D cells, we should account for the impact of each scheme on the width of cell layouts to evaluate the final footprint savings. This is because two MIV channels in stitching T-M3D cells simplify the routing topology in the large standard cells that suffer from complex local interconnection. For example, compared with the 2D and the folding T-M3D layouts, the stitching T-M3D D flip-flop (DFF) layout decreases the width of the cell by 320*nm* (12.5%) shown in Figure 5. Therefore, the stitching T-M3D DFF achieves 4% lower layout footprint than that of the folding T-M3D DFF, and 52% lower than that of the 2D DFF.

As we observe in the example of a DFF cell, the footprint savings of stitching T-M3D cells depend on the level of routing optimization using two MIV channels. Folding T-M3D cells with simple internal routing has better footprint savings than stitching T-M3D cells, but the large and complex T-M3D cells achieve the footprint savings from the stitching scheme because of two MIV channels since they have much room for the routing optimization.

B. Parasitic Analysis

We analyze the parasitics of T-M3D layouts based on the extraction results of the simplest layout. Table I shows net coupling capacitances in the INV_X1 layout of the 2D, folding T-M3D, and stitching T-M3D cells. We first address the noticeable changes in the parasitics of the folding T-M3D layout. Compared to the coupling capacitance between the IN and OUT net of the 2D parasitics, that of the folding T-M3D INV_X1 layout increases 45.4% (49aF). After all, like the 2D layout, the folding scheme retains the routing of the IN and

OUT nets in parallel, while vertical interconnection with an MIV lengthens the routing interconnects of the IN and OUT nets. Both the IN and OUT nets in the folding T-M3D layout go through every metal stack, including M1B, V1B, M2B, V2B, (MIVG, MIVM), (GILT, M1T), V0T, and M1T. Therefore, the sum of area facing each other between the IN and OUT nets in the folding T-M3D layout is $0.023\mu m^2$ while it is $0.017\mu m^2$ for M1 of the 2D layout. Analysis of layer-by-layer capacitance clearly explains the increase in coupling capacitance between the IN and OUT nets. Coupling between layers in the parallel pillars for the IN and OUT nets contribute to 21aF capacitance. The GILT layer of the IN net and the MIVM layer of the OUT net are also very close (12*nm*) because of the one MIV channel in the folding scheme. Hence, coupling between these two layers contributes another 10aF capacitance. A decrease of 81.3% (26aF) in the VDD and VSS net coupling is noteworthy because the VDD and VSS nets do not face each other in the folding T-M3D layout, and ILD electrically separates the coupling between the top and bottom tiers.

With regard to the parasitics of the stitching T-M3D layout, we observe only an 8% (9aF) increase in coupling capacitance between the IN and OUT nets. The stitching T-M3D layout does not suffer from huge coupling capacitance between the IN and OUT nets. This is because each net uses an independent MIV channel. However, a small increase comes from the short distance between the MIVG layer of the IN net and the diffusion layer of the OUT net. Since the VDD and OUT nets share the same MIV channel in the stitching T-M3D layout, coupling between the parallel pillars for the OUT and VDD nets result in a 107.1% (30aF) increase in coupling capacitance between the VDD and OUT nets.

The long wirelength of the vertical interconnections with MIVs also significantly impact the ground capacitance and the resistance of the 3D net. In the 2D layout, the ground capacitance of the IN net is only 4aF, but the ground capacitances of the IN net in the folding and stitching T-M3D are 18aF and 17.2aF, respectively. On an average, the resistance values of the IN and OUT nets of the folding and stitching T-M3D layouts are 13% more than that of the 2D layout. An increase in the resistance of the VDD net in the stitching T-M3D layout is noticeable. While the resistance of the VDD net in the 2D layout is 15.4Ω, it becomes 43.4Ω in the stitching T-M3D layout.

In summary, parasitic analysis shows that T-M3D layouts suffer from additional parasitics. They do not fully turn the huge footprint savings into the power/performance savings. This is because the wirelength of a net that becomes 3D by using vertical interconnection with MIVs is longer than that of the net in the 2D layout. However, T-M3D layouts can improve parasitics when the net in the 2D layout is long enough for routing optimization using multiple MIVs.

C. Cell Power and Performance Analysis

Based on the results of timing/power model (.lib) of the 41 standard cells of the 2D, folding T-M3D, and stitching T-M3D layouts, Table II shows the best, worst, and average savings in the timing and power metrics of T-M3D cells normalized to 2D metrics. On an average, folding T-M3D cells show a small degradation in both delay and power consumption compared to 2D cells. On the other hand, The stitching T-M3D cells show benefits in power consumption. In terms of the best timing and power savings compared to the 2D metrics, the stitching T-M3D NOR4_X1 cell reduces power by 6.12% while the folding T-M3D NAND4_X2 cell reduces power by 4.44%. The improvement ratio in the timing metrics of stitching T-M3D layouts is also higher than that of folding T-M3D layouts. Although the stitching scheme has a disadvantage in footprint savings, its two MIV channels reduce coupling between vertical routings. This leads to better timing and power than the folding scheme.

TABLE I

COMPARISON OF NET COUPLING CAPACITANCE FOR INV_X1. COMPARED TO THE 2D LAYOUT, THE COUPLING CAPACITANCE BETWEEN THE IN AND OUT NETS INCREASES 45.4% IN THE FOLDING T-M3D LAYOUT WHILE IT IS ONLY AN 8% INCREASE IN THE STITCHING T-M3D LAYOUT.

2D (fF)	IN	OUT	VDD	VSS
IN	-	0.108	0.057	0.058
OUT	0.108	-	0.027	0.027
VDD	0.057	0.027	-	0.032
VSS	0.058	0.027	0.032	-
Folding T-M3D (fF)	IN	OUT	VDD	VSS
IN	-	**0.157**	0.055	0.061
OUT	**0.157**	-	0.028	0.022
VDD	0.055	0.028	-	**0.006**
VSS	0.061	0.022	**0.006**	-
Stitching T-M3D (fF)	IN	OUT	VDD	VSS
IN	-	**0.117**	0.056	0.064
OUT	**0.117**	-	**0.058**	0.019
VDD	0.056	**0.058**	-	0.022
VSS	0.064	0.019	0.022	-

TABLE II

THE BEST, WORST, AND AVERAGE SAVINGS IN THE TIMING AND POWER METRICS OF T-M3D CELLS NORMALIZED TO THE 2D METRICS. Δ% INDICATES THE SAVINGS COMPARED WITH 2D. IN THE BEST CASES, STITCHING T-M3D CELLS OUTPERFORM THE FOLDING T-M3D CELLS.

Folding	Best Cell	Δ%	Worst Cell	Δ%	Ave (Δ%)
Fall Delay	NOR3_X1	0.50	AND2_X2	-5.11	-1.68
Rise Delay	NOR3_X1	0.37	AND2_X2	-5.74	-1.82
Fall Slew	INV_X8	0.78	BUF_X8	-4.62	-1.61
Rise Slew	NOR3_X1	0.14	BUF_X8	-7.17	-2.06
Fall Power	NAND4_X2	4.44	AND2_X2	-10.31	-0.72
Rise Power	NOR3_X1	1.33	OR3_X2	-9.20	-4.03
Stitching	Best Cell	Δ%	Worst Cell	Δ%	Ave (Δ%)
Fall Delay	NOR4_X2	1.97	BUF_X8	-3.84	-0.21
Rise Delay	NOR4_X1	1.40	BUF_X8	-5.39	-1.20
Fall Slew	NOR4_X2	1.09	BUF_X8	-6.42	-0.84
Rise Slew	NOR4_X1	1.09	BUF_X8	-10.60	-2.31
Fall Power	OR4_X1	3.17	AND2_X2	-6.17	-0.84
Rise Power	NOR4_X1	6.12	AND2_X2	-9.31	1.21

IV. PDN DESIGN METHODOLOGY FOR FOLDING T-M3D ICS

In the stitching T-M3D cell layout, VDD and VSS rails are located on the top and bottom edges of the cell boundary in the same way of the traditional 2D layout. Therefore, full-chip T-M3D ICs with the stitching T-M3D standard cells (Stitching T-M3D ICs) are designed with existing 2D CAD engines without any problems. However, full-chip T-M3D ICs with the folding T-M3D standard cells (Folding T-M3D ICs) are in need for a novel power delivery network (PDN) design methodology. The reason is that the ground and power rails overlap in the folding T-M3D cell layout. To evaluate the impact of the T-M3D layout scheme on the full-chip static power integrity, this section presents a PDN design methodology for the folding T-M3D ICs.

When the power is delivered from the VDD ring around the folding T-M3D ICs as proposed in [3], it does not guarantee the tolerable static IR-drop at the center of design. Therefore, we periodically make VDD landing spaces by disconnecting the top VSS rails. After we finish the floorplanning and route bottom VDD stripes only,

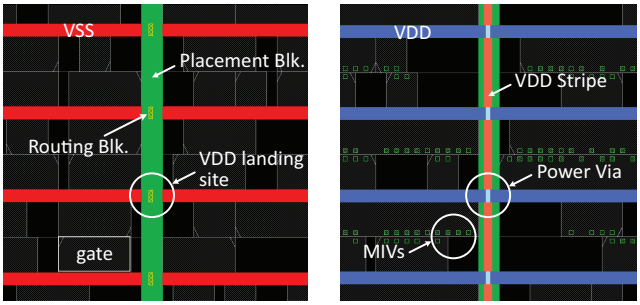

Fig. 6. Die shots of the PDN in the folding T-M3D IC. (a) VSS rail routing on the top tier, (b) VDD rail routing on the bottom tier

we create placement and routing blockages underneath the expected VDD landing sites. Placement blockages keep the space empty during subsequent design stages, so that VDD landing sites are not occupied by standard cells. Next, we finish the placement step, and route the VSS stripes. Due to the routing blockages that we have created before, ground rails are disjoint. Then we modify the initial routing blockages to place them only on top of the VSS rails as shown in Figure 6(a). This step allows the routing engine to make use of M2 layer effectively during the detail routing step. Now we create the power grid mesh from M3 layer (M3T) in the same way as in 2D ICs. In this stage, VDD is not completely connected but only routing blockages are located at the VDD landing sites. After Clock Tree Synthesis (CTS), routing, and post-route optimization are done, we route the bottom VDD rail to connect power grid mesh above the M3T layer while removing the existing routing blockages. Figure 6(b) shows the final VDD network on the bottom tier.

V. Impact of Full-Chip T-M3D ICs

To evaluate the impact of T-M3D layout optimization on the full-chip design, we choose Triple Data Encryption Standard (DES3), Advanced Encryption Standard (AES), and Low-Density Parity-Check (LDPC) circuit benchmarks from open source hardware benchmark suites. For each benchmark, the core size is set by 70% of the final placement utilization, and the clock period is determined by the worst negative slack which is less than 20ps. We set the clock period of DES3 as 0.4*ns*, 0.5*ns* for AES, and 1.2*ns* for LDPC. PDN metal usage is determined under the condition that the die static IR-drop is less than 2% (16mV) of 0.8V supply voltage for each benchmark. 10% of M5 with 64*nm* width, 10% of M6 with 128*nm* width, and 20% of M7, M8 with 384*nm* width are used for PDN design in every benchmark. Under the same settings, folding and stitching T-M3D ICs are designed by respective design flows, and voltage sources are placed on every cross-section between M8 and M7 layers. In this work, we do not take the packaging IR-drop into account.

A. Area and Wirelength Results

Table III shows the final design result. Area savings in the T-M3D cell layout have a significant impact on the footprint of the full-chip design. It is worth noting that the design footprint savings in the stitching T-M3D ICs vary from 41% to 44%. Since the D flip-flops (DFF) composition in each benchmark varies, the design footprint savings are higher than the cell-level footprint savings (38%). Because the stitching T-M3D DFF cell has simplified routing topology as mentioned in Section III, its area is 4% lower than that of the folding T-M3D DFF cell. Therefore, the more DFF-dominant a circuit is, the better the footprint savings are in the stitching T-M3D

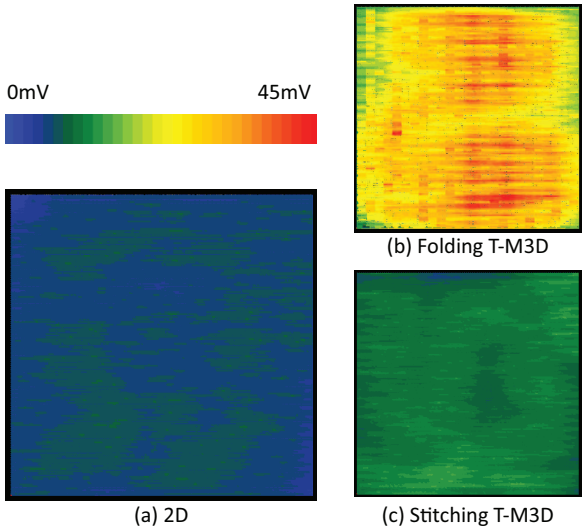

Fig. 7. Static IR-drop map of DES3. (a) 2D (max = 8.05mV), (b) Folding T-M3D (max = 44.68mV), (c) Stitching T-M3D (max = 13.2mV).

designs. DES3 is an example of the DFF-dominant circuit. Out of 57K gates, DFF has 15% of cells as DFF, so stitching T-M3D DES3 design achieves 44% of footprint savings, which is in the comparable level of the footprint savings (45%) from the folding T-M3D ICs.

Wirelength savings follow the huge footprint savings. Folding T-M3D LDPC design achieves 24% lesser wirelength at its best. Since LDPC is a wire-dominant circuit based on the average net length, the impact of the footprint reduction on the wirelength savings is higher than other benchmarks. Stitching T-M3D LDPC has relatively small footprint savings than the folding T-M3D design because only 2% of gates are DFFs. As a result, folding T-M3D LDPC shows more wirelength reduction than the stitching T-M3D design does.

B. Power and IR-drop Results

Even though the huge wirelength savings result in the wire capacitance savings in T-M3D ICs, the switching power of T-M3D designs shows degradation in AES and DES3. This is because the sum of wire capacitance and gate pin capacitance determines the total switching power of a design. Therefore, the impact of wire capacitance savings on the switching power depends on whether the circuit characteristic is wire-dominant or gate-dominant. Moreover, T-M3D standard cells have larger input and output pin capacitance than 2D on average. For gate-dominant DES3 and AES, wire capacitance reduction is not enough to compensate the pin capacitance increase, resulting in the increase of total capacitance. However, the stitching T-M3D cells reduce the increase of pin capacitance with the help of layout optimization. The switching power derived from pin capacitance is relatively smaller than that of the folding T-M3D designs, and shows more total switching power savings.

Folding T-M3D designs also have more internal power than respective 2D designs due to the degraded cell characteristic. Gate-dominant circuits have more impact from the internal power increase of the folding T-M3D cells. However, stitching T-M3D designs reduce the internal power than 2D in the case of every benchmark. As a result, folding T-M3D designs have 2.5% of total power increase in gate-dominant DES3 and AES, but 7% of savings in wire-dominant LDPC. On the other hand, stitching T-M3D designs always show total power savings. They are 1% of power savings for DES3, 4% for AES, and 6% for LDPC. Folding T-M3D designs have more power savings in

TABLE III
FULL-CHIP DESIGN METRIC, POWER, AND IR-DROP COMPARISON. EACH BENCHMARK IS TIMING-CLOSED AT THE SAME TARGET FOR ISO-PERFORMANCE COMPARISONS.

Design Style	Footprint (μm^2)	Placement utilization(%)	Wirelength (m)	Netlength (μm)	Wire Cap / Pin Cap	Switching Power (mW)	Internal Power (mW)	Total Power (mW)	Static IR-drop (mV)
DES3 circuit, 2.5GHz									
2D	32596	75.7	0.265	4.63	23:77	61.4	89.0	151.1	8.05
Folding T-M3D	17789 (-45%)	74.3	0.218 (-18%)	3.91	20:80	64.0 (+4%)	90.5 (+2%)	155.1 (+3%)	44.68 (x5.55)
Stitching T-M3D	18141 (-44%)	75.2	0.215 (-19%)	3.77	21:79	61.9 (+1%)	87.1 (-2%)	149.7 (-1%)	13.28 (x1.65)
AES circuit, 2.0GHz									
2D	60981	72.4	0.824	6.32	30:70	80.2	93.9	175.5	6.43
Folding T-M3D	33410 (-45%)	72.2	0.651 (-21%)	5.02	25:75	80.5 (+0%)	95.2 (+1%)	177.1 (+2%)	24.67 (x3.84)
Stitching T-M3D	35014 (-43%)	72.2	0.671 (-19%)	5.15	26:74	78.7 (-2%)	90.9 (-3%)	170.9 (-4%)	9.43 (x1.47)
LDPC circuit, 833MHz									
2D	32585	69.7	1.099	14.07	54:45	68.5	34.9	104.2	11.47
Folding T-M3D	18020 (-45%)	65.7	0.837 (-24%)	11.28	47:52	62.4 (-9%)	34.0 (-3%)	97.0 (-7%)	62.58 (x5.46)
Stitching T-M3D	19372 (-41%)	65.01	0.896 (-18%)	12.12	50:49	63.9 (-7%)	33.3 (-5%)	97.9 (-6%)	15.57 (x1.36)

the wire-dominant circuit than stitching T-M3D designs do since the 45% of guaranteed footprint reduction leads to more wire capacitance savings.

For the die static IR-drop, folding T-M3D designs do not avoid the huge degradation in the static power integrity because of the limited VDD connections. While folding T-M3D designs show a maximum die IR-drop of 8%, stitching T-M3D designs guarantee less than 2%. Since improving static IR-drop by increasing the PDN metal usage increases the routing congestion and the number of placement blockages because of the VDD landing sites, the power optimization of the folding T-M3D design is limited.

To summarize, folding T-M3D designs show a maximum footprint savings of 45% and power savings of 7%. However, the severe degradation in static power integrity reduces their reliability. On the other hand, our stitching T-M3D designs guarantee maximum power savings of 6% at static IR-drop less than 2% while achieving design footprint savings greater than 41%.

VI. CONCLUSION

In this study, we proposed a new layout optimization method, the stitching scheme, for the transistor-level monolithic 3D (T-M3D) standard cell design. The stitching scheme addresses the static power integrity issue inherent in the folding scheme for T-M3D cell layouts. It also minimizes the timing/power degradation caused by parasitics originating from the unique T-M3D layout architecture. We developed the 14*nm* T-M3D technology process design kit and designed 41 standard cells in the form of 2D, folding T-M3D, and stitching T-M3D layouts. We proved that the stitching scheme outperforms the folding scheme in terms of timing and power metrics at the expense of the increase in the cell height by only 0.5 metal tracks. We also presented a design methodology for a power delivery network in folding T-M3D ICs, and performed sign-off IR-drop analysis in both folding and stitching T-M3D ICs. Lastly, we found that the folding scheme cannot be applied to commercial grade layouts because of its severe IR-drop. However, compared to 2D ICs, the stitching T-M3D ICs experience only 6mV increase in maximum IR-drop while reducing the footprint by up to 44% and power consumption by 6%.

VII. ACKNOWLEDGEMENT

The authors would like to thank Tao Peng, Alexei Svizhenko, and Arindam Chatterjee at Synopsys StarRC R&D group for the helpful advice and guidance regarding extracting and analyzing the parasitics of T-M3D standard cells.

REFERENCES

[1] C. C. Yang *et al.*, "Footprint-efficient and power-saving monolithic IoT 3D IC constructed by BEOL-compatible sub-10nm high aspect ratio (AR>7) single-grained Si FinFETs with record high Ion of 0.38 mA/μm and steep-swing of 65 mV/dec. and Ion/Ioff ratio of 8," in *Proc. IEEE Int. Electron Devices Meeting*, 2016, pp. 9.1.1–9.1.4.

[2] 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, pp. 1–10.

[3] J. Shi *et al.*, "On the Design of Ultra-High Density 14nm Finfet Based Transistor-Level Monolithic 3D ICs," in *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2016, pp. 449–454.

[4] B. W. Ku, P. Debacker, D. Milojevic, P. Raghavan, and S. K. Lim, "How much cost reduction justifies the adoption of monolithic 3d ics at 7nm node?" in *Proc. IEEE Int. Conf. on Computer-Aided Design*, 2016, pp. 87:1–87:7.

[5] C. Fenouillet-Beranger *et al.*, "FDSOI bottom MOSFETs stability versus top transistor thermal budget featuring 3D monolithic integration," *Solid-State Electronics*, vol. 113, pp. 2 – 8, 2015.

[6] P. Batude *et al.*, "3-D Sequential Integration: A Key Enabling Technology for Heterogeneous Co-Integration of New Function With CMOS," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2012.

[7] S. Sinha, G. Yeric, V. Chandra, B. Cline, and Y. Cao, "Exploring Sub-20Nm FinFET Design with Predictive Technology Models," in *Proc. ACM Design Automation Conf.*, 2012, pp. 283–288.

[8] "International Technology Roadmap for Semiconductors 2013," 2013.

[9] K. Bhanushali and W. R. Davis, "FreePDK15: An Open-Source Predictive Process Design Kit for 15Nm FinFET Technology," in *Proceedings of the 2015 Symposium on International Symposium on Physical Design*, ser. Proc. Int. Symp. on Physical Design, 2015, pp. 165–170.

[10] L. Liebmann, J. Zeng, X. Zhu, L. Yuan, G. Bouche, and J. Kye, "Overcoming scaling barriers through design technology CoOptimization," in *IEEE Int. Symposium on VLSI Technology, Systems, and Applications*, June 2016, pp. 1–2.

[11] K. Miyaguchi *et al.*, "Modeling FinFET metal gate stack resistance for 14nm node and beyond," in *International Conference on IC Design Technology (ICICDT)*, 2015, pp. 1–4.

[12] Y. Sasaki *et al.*, "Novel junction design for NMOS Si Bulk-FinFETs with extension doping by PEALD phosphorus doped silicate glass," in *Proc. IEEE Int. Electron Devices Meeting*, 2015, pp. 21.8.1–21.8.4.

[13] Y. S. Yu, S. Panth, and S. K. Lim, "Electrical Coupling of Monolithic 3-D Inverters," *IEEE Transactions on Electron Devices*, vol. 63, pp. 3346–3349, 2016.

[14] M. Martins *et al.*, "Open Cell Library in 15Nm FreePDK Technology," in *Proc. Int. Symp. on Physical Design*, 2015, pp. 171–178.

[15] R. Aitken *et al.*, "Physical Design and FinFETs," in *Proc. Int. Symp. on Physical Design*, 2014, pp. 65–68.

[16] B. Chava *et al.*, "Standard cell design in N7: EUV vs. immersion," in *Proc. SPIE*, 2016, pp. 94 270E–94 270E–9.