Ultrahigh Density Logic Designs Using Monolithic 3-D Integration

Young-Joon Lee, *Student Member, IEEE,* and Sung Kyu Lim, *Senior Member, IEEE*

***Abstract*—The nano-scale 3-D interconnects available in monolithic 3-D integrated circuit (IC) technology enable ultrahigh density device integration at the individual transistor level. In this paper, we investigate the benefits and challenges of monolithic 3-D integration technology for ultrahigh density logic designs. We first build a 3-D standard cell library for transistor-level monolithic 3-D ICs and model their timing and power characteristics. Then, we explore various interconnect options for monolithic 3-D ICs that improve design quality. Next, we build timing-closed, full-chip GDSII layouts and perform sign-off iso-performance power comparisons with 2-D IC designs. Based on layout simulations, we compare important design metrics such as area, wirelength, timing, and power consumption of transistor-level monolithic 3-D designs with traditional 2-D, gate-level monolithic 3-D, and TSV-based 3-D designs.**

***Index Terms*—3-D integrated circuit (IC), logic design, low power, monolithic integration.**

I. Introduction

It is believed that in today's logic designs, interconnects dominate the timing and power of circuits; therefore, reducing the interconnect length may improve the timing and power of circuits. By stacking device layers in 3-D using through-silicon-vias (TSVs), not only the footprint is reduced but also the average distance among devices is reduced, leading to a shorter total wirelength and better performance. However, the shortcoming of TSV-based 3-D integrated circuits (ICs) is the area overhead [1] and the minimum keep-out-zone of TSVs because of manufacturing issues such as die alignment precision [2] and mechanical stress [3]. In addition, the parasitic capacitance of TSVs is large (tens-hundreds of fF), which may degrade the timing and power of circuits.

To better exploit the benefits from 3-D die stacking, monolithic 3-D technology is currently being investigated as a next-generation technology. In a monolithic 3-D IC, the device layers are fabricated sequentially, rather than bonding two fabricated dies together using bumps and/or TSVs. When the top layer is attached to the bottom layer, the top layer is a blank silicon. Alignment precision is determined by lithography stepper accuracy, which is around 10 nm today. Also, the top layer can be made very thin, around 30 nm [4]. Thus, monolithic intertier vias (MIVs) for vertical connections are very small—about two orders of magnitude smaller than through-silicon-via (TSV)—with a negligibly small parasitic capacitance (< 0.1 fF). A side view of a typical monolithic 3-D IC is shown in Fig. 1. With these small MIVs, designers can truly exploit the benefit of vertical dimension.

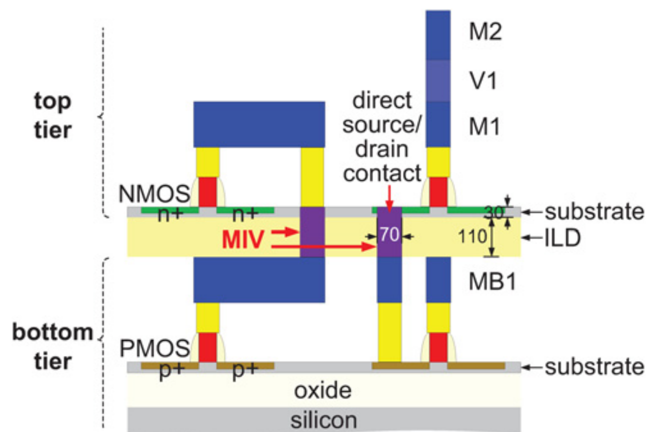

Fig. 1. Side view of a two-tier monolithic 3-D IC [5]. The MIV and ILD stand for monolithic intertier via and interlayer dielectric. On the top tier, only the first two metal layers (M1, M2) are shown. Objects are drawn to scale. Unit is nm.

As discussed in [6] and [7], monolithic 3-D technology enables a very fine-grained 3-D circuit partitioning. We can divide standard cells into pMOS and nMOS parts, place them in different layers, and connect them using MIVs, which we call transistor-level monolithic 3-D integration (T-MI) in this paper. Or, as in TSV-based 3-D ICs, we may place planar cells in different layers and connect them using MIVs, which is called gate-level monolithic 3-D integration (G-MI). In this paper, we focus on T-MI that allows the highest integration density possible. The comparisons among T-MI, G-MI, TSV-based 3-D, and conventional 2-D designs are provided. In addition, we study the power benefit of T-MI based on timing-closed, detailed routing completed GDSII-level layouts and sign-off analysis on timing and power. With our layout-based simulations and in-depth analyses, we demonstrate how to maximize the power benefit of T-MI technology. For fair comparisons between T-MI and 2-D designs, timing is closed on all designs (iso-performance), and power consumption is compared.

Manuscript received January 29, 2013; revised May 2, 2013; accepted June 21, 2013. Date of current version November 18, 2013. This work was supported in part by Intel, Qualcomm, and the Center for Integrated Smart Sensors (CISS) funded by the Korean Ministry of Science, ICT and Future Planning as Global Frontier Project (CISS-2012366054194). This paper was recommended by Associate Editor D. Atienza.

The authors are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30308 USA (e-mail: yjlee@gatech.edu; limsk@ece.gatech.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCAD.2013.2273986

0278-0070 © 2013 IEEE

The major contributions of this paper are as follows.

1) We explain how the 3-D standard cells for T-MI are designed for high-density integration. Various practical layout and design techniques for density and performance are discussed. To the best of our knowledge, this is the first work to characterize the timing and power of the T-MI cells. We extract the internal RC parasitics of our T-MI cells and characterize their timing and power to compare them against 2-D counterparts.
2) We explore interconnect options for T-MI to address the routing congestion problem. The metal layer structures and their dimensions are varied. With layout-based experiments, we provide detailed analysis on wirelength, timing, and power metrics with several benchmark circuits. We also provide wirelength-binning-based analysis to further understand the benefit of T-MI.
3) We present a power benefit study of T-MI. We perform iso-performance comparisons between 2-D and T-MI designs. In addition, we perform layout designs for G-MI and TSV-based 3-D for comparison purposes.

The remainder of this paper is organized as follows. In Section II, we provide background knowledge. In Section III, we present our design methods for T-MI technology in detail. In Section IV, we explore interconnect options for T-MI. In Sections V and VI, we perform iso-performance comparisons between 2-D and T-MI, as well as G-MI and TSV-based 3-D. Finally, we conclude in Section VII.

II. Backgrounds

In this paper, we assume the monolithic 3-D IC fabrication process from CEA/LETI [4]. Key features of their monolithic 3-D process flow are wafer-level molecular bonding with a thin interlayer dielectric and a special salicidation process, under a specific thermal budget.

One of huge benefits of monolithic 3-D technology is the alignment precision between layers. In monolithic 3-D ICs, this alignment between layers only depends on lithographic alignment capability. Batude *et al.* [8] demonstrated high alignment precision in monolithic 3-D ICs ($\sigma \approx 10$ nm) compared with TSV-based 3-D integration ($\sigma \approx 0.5\,\mu$m) [2]. The nano-scale alignment precision and the ultrathin silicon and interlayer dielectric (ILD) layers enable nano-scale 3-D interconnects.

A. Design Styles of Monolithic 3-D ICs

As shown in Fig. 2, we categorize the design styles of monolithic 3-D ICs into two: gate-level (G-MI) and transistor-level (T-MI). As in TSV-based 3-D ICs, in G-MI designs, standard cells are planar (2-D) and each layer contains multiple metal layers. However, in G-MI, device layers are fabricated sequentially, and MIVs are much smaller than TSVs.

The T-MI designs are different from G-MI.

1) Most of the 3-D interconnects are embedded in the 3-D cells.
2) pMOS and nMOS transistors are on different layers, thus manufacturing processes can be optimized separately per layer.

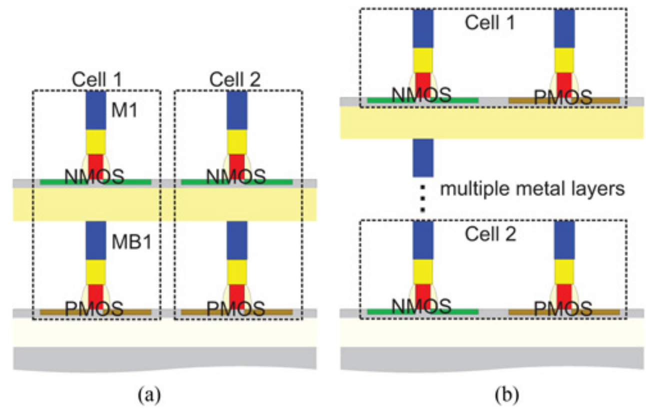

Fig. 2. Design styles of monolithic 3-D ICs. (a) T-MI. (b) G-MI.

3) Physical layout (placement, routing, optimization, etc.) can be performed using existing 2-D electronic design automation (EDA) tools with a little modifications.

In contrast, G-MI or TSV-based 3-D ICs require 3-D-aware physical layout engines. Currently, no commercial EDA tool can handle multiple dies together, especially for optimizations. Thus, previous works [9] and [10] rely on die-by-die optimizations with timing constraints on the die boundary. However, the design quality with this approach is suboptimal because the optimization engine cannot see the whole 3-D paths.[1]

B. Related Works

The monolithic 3-D fabrication technologies were proposed and demonstrated in [4] and [11]. Currently, there are a few related works on the design of monolithic 3-D ICs. Jung *et al.* [12] demonstrated the single-crystal thin-film-based process for their SRAM design, which reduced the SRAM cell footprint by 46.4%. Recently, Golshani *et al.* [13] demonstrated the monolithic 3-D integration of SRAM and image sensor. Also, Naito *et al.* [14] demonstrated the first 3-D FPGA design implementation based on a monolithic 3-D IC technology. These works [12]–[14] were application-specific, meaning that a general design methodology for logic designs was not presented.

Recently, logic design methodologies for monolithic 3-D technology were demonstrated in [6] and [7]. Yet, the presented design techniques and interconnect options did not resolve the routing congestion problem in transistor-level monolithic 3-D designs, which may degrade the design quality much. The routing congestion problem was addressed in our recent work [5]. However, timing was not closed in these works [5]–[7], which makes the timing and power comparisons non-practical and unfair. Since better timing can be traded with lower power consumption, it is essential that all the design options under consideration are timing-closed to allow iso-performance power comparison. In addition, these works assume that the timing and power characteristics of 3-D monolithic gates are the same as 2-D gates and did not demonstrate why that is a reasonable assumption. The authors also did not provide in-depth analyses and discussions on why monolithic

[1]The optimization limitations are presented in Section VI-A.

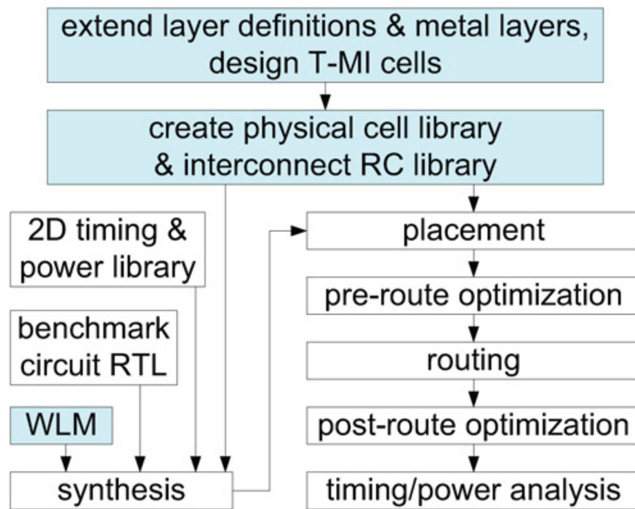

Fig. 3. Overall design and analysis flow for T-MI. Shaded boxes highlight differences in T-MI. The WLM means wire load model.

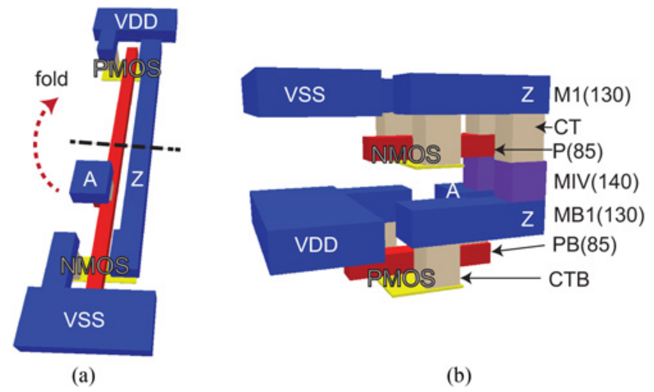

Fig. 4. Layout of an inverter from (a) Nangate 45-nm library and (b) our T-MI library. P, M, and CT represent poly, metal, and contact. The suffix B means the bottom tier. MIV means monolithic intertier via. Top/bottom tier silicon substrate and p/nwells are not shown for simplicity. The numbers in parentheses mean thickness in nm.

3-D technology reduces power consumption and what factors affect the power reduction margin. This knowledge is crucial to maximize the benefit and justify on-going and future research on fabrication and design technologies for monolithic 3-D ICs.

III. Design Methodologies

In this section, we explain our design methods for T-MI technology in detail. Various practical considerations for high density and high performance T-MI designs are discussed.

A. *Overall Design and Analysis Flow*

One of the major benefits of T-MI is that existing 2-D EDA tools can be used, with simple modifications if needed. We extensively use commercial EDA tools in this paper. Our design and analysis flow, summarized in Fig. 3, consists of four parts: 1) library preparations; 2) synthesis; 3) layout; and 4) analysis. In the library preparation part, we prepare T-MI-specific library files. We synthesize the RTL codes of benchmark circuits using Synopsys Design Compiler. In the layout part, we perform placement, routing, and optimizations using Cadence Encounter (v10.12). Finally, we perform static timing analysis and static power analysis.

Our major efforts for T-MI design flow are spent on T-MI cell library construction and characterization, T-MI interconnect structure modeling, and T-MI wire load modeling. We modify the technology files and design rules to account for additional layers on the bottom tier, as well as additional metal layers on the top tier (Section IV-B). Using Cadence Virtuoso, we create our T-MI cells by modifying existing 2-D cells. The cells are then abstracted to create the T-MI physical cell library. We also build interconnect RC libraries using Cadence capTable generator and QRC Techgen. For synthesis, we create the T-MI wire load models that reflect reduced wirelengths with T-MI. The T-MI wire load models guide synthesis optimizations; with shorter wirelengths, the synthesized netlist of T-MI contains weaker cells and less number of buffers than that of 2-D, for the same clock period.

For layout construction, we first run Encounter placer. The tool recognizes T-MI cells as the cells with pins on multiple layers. For routing, we set up Encounter to utilize the additional metal layers on bottom and top tiers. Since our T-MI cells contain routing blockages on the MIV layer, the router avoids routing through the top tier part of the cells. Using our T-MI interconnect library that reflects the T-MI metal layer structures and materials, we perform RC extraction on all the nets in the layout. Our full-chip timing/power optimizations and analyses for T-MI and 2-D are the same, because the entire T-MI design (top/bottom tiers) is captured in a single Encounter session. We perform static power analysis with the switching activity of the primary inputs and sequential cell outputs at 0.2 and 0.1, respectively.

B. *Monolithic 3-D Cell Design*

1) *Cell Design Methodology and Discussions:* We design our T-MI 3-D cells using the (2-D) standard cells in Nangate 45 nm library [15] as our baseline. As shown in Fig. 4, we fold the 2-D standard cells into 3-D and create T-MI 3-D cells. The thicknesses of top/bottom tier silicon substrates and ILD are 30 nm and 110 nm, respectively. The diameter of MIV is 70 nm. Note that by folding, cell pins (A, Z) are on both tiers. We prefer to place the pMOS transistors on the bottom tier and the nMOS on the top tier. In Nangate 45-nm library, p/nMOS transistors show hole/electron mobility skew. To compensate the difference, in Nangate 45-nm library, a pMOS is larger than the corresponding nMOS. Since extra silicon space on the top tier is required for MIVs [not on the bottom tier—see Fig. 4(b)], placing pMOS transistors on the bottom tier balances top/bottom silicon area usage. However, we should also consider manufacturing aspects in deciding the p/nMOS layer assignment.[2]

After folding the cell, VDD and VSS strips are overlapping, as shown in Fig. 4. The power to VDD on the bottom tier can be delivered down through arrays of MIVs, placed apart

[2]In sub-32 nm nodes, due to advanced channel engineering techniques, the hole/electron mobility is about the same.

from the VSS strip. We may need extra space for these VDD MIVs. Yet, power delivery network (PDN) design and IR-drop analysis are outside our scope. Also, since VDD and VSS strips are overlapping, it may act as a small decoupling capacitor. However, in the extracted cell internal RC data for our inverter cell, the coupling capacitance (or *cap*) between VDD and VSS strips is around 0.01 fF, which is small compared with other cell internal parasitic capacitances.

The transistor model in Nangate 45-nm library is PTM 45 nm with bulk silicon technology [16]. In monolithic 3-D technology, because of the structure, top tier transistors are similar to silicon-on-insulator devices [4]. However, in this paper, we assume the same transistor model for T-MI and 2-D cells, because: 1) the original Nangate 45-nm library is based on bulk silicon technology; and 2) if we assume both devices and interconnect structures in T-MI are different from 2-D, it becomes harder to understand which factor contributes to power reduction, by how much.

Our standard cell design method differs from IntraCell Stacking in [6] for three major reasons.

1) We place pMOS transistors on the bottom tier and nMOS transistors on the top. If pMOS is on the top tier as in [6], we need extra space for MIVs, which increases the cell footprint.
2) We apply our cell folding technique on the original 2-D standard cell layouts. Compared with the IntraCell Stacking technique in [6] that requires a complete re-design of internal connections, our method is straight-forward and provides opportunities for reducing internal RC parasitics.
3) We place VDD/VSS strips of standard cells on the bottom side in different tiers. Compared with the intracell stacking in [6] that places power/ground rails on the top/bottom side of the standard cells, our method further reduces the cell footprint because M1/MB1 routing space is even for the top and bottom tiers.[3]

Our T-MI cells preserve the same transistor sizes as in the original 2-D cells. GDSII layouts of some of our T-MI cells are shown in Fig. 5. The T-MI cell height is 0.84 μm, which is 40% smaller than the original 2-D cell height (1.4 μm). Thus, cell footprint reduces by 40%,[4] which is more than the reported values in [6] (about 30%).

When designing T-MI cells, care should be taken to reduce cell internal RC parasitics. As shown in Fig. 4(b), the path from the pMOS on the bottom tier to the nMOS on the top tier consists of CTB, MB1, MIV, CT, M1, then CT to diffusion. This 3-D path may become larger than the original 2-D path and may increase cell internal parasitic RC. Similarly, the path from the PB on the bottom tier to the P on the top tier consists of multiple layers. To reduce cell internal RC parasitics, it is important to minimize the lengths of 3-D paths. To achieve shorter 3-D paths, we should place MIVs close to the connecting transistors. We also need to utilize direct source/drain (S/D) contacts [Fig. 5(c)]. The direct S/D contacts reduce the detour in the 3-D paths and unnecessary RC parasitics.

[3]This may incur small area overhead for PDN to MB1.

[4]The reasons why it is not 50% are: 1) p/nMOS size mismatch incurs extra space on nMOS side, and 2) MIVs require extra space on the top tier.

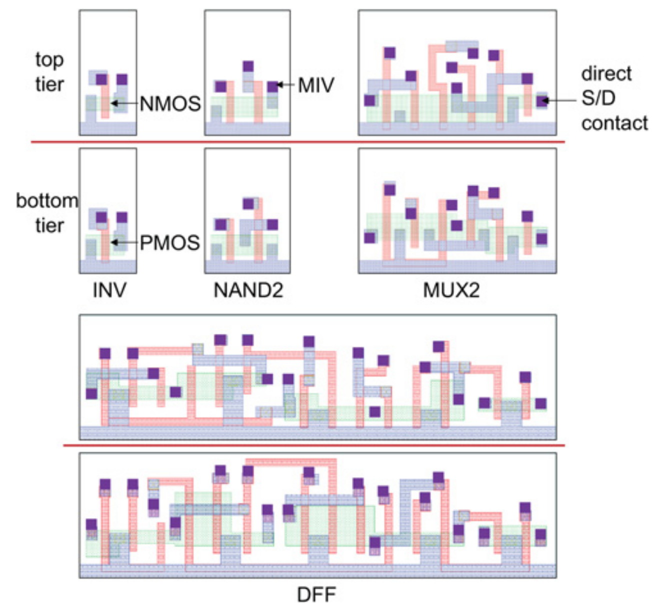

Fig. 5. Layout snapshots of our T-MI cells. The S/D means source/drain. The p/nwell and implants are not shown for simplicity.

2) *Comparison of T-MI and 2-D Cells:* We examine the cell internal RC parasitics of 3-D and 2-D cells and the impact on timing/power. In previous works [5]–[7], the authors assumed that the delay and power of 3-D cells are the same as 2-D cells and used 2-D timing/power library. Batude *et al.* [4] fabricated a transistor-level monolithic 3-D IC and measured the top/bottom transistor performances. They reported that the differences between 3-D transistors and baseline 2-D transistors were negligible. Yet, the delay and power of cells are also affected by cell internal RC parasitics. From Fig. 4(b), we can conjecture that there are coupling capacitances among PB, CTB, MB1, MIV, CT, and M1. Using Mentor Graphics Calibre XRC with electromagnetic-simulation-based extraction rules, we extract these capacitance values, as well as resistances and transistors from our T-MI cell layout. Then, we generate a SPICE netlist of the cell that consists of transistors and parasitic RC components.

Since Calibre XRC is designed for 2-D ICs, it can only model one diffusion layer. Due to this tool limitation, top tier diffusion layer can be modeled as either dielectric or conductor. Even though the top tier silicon is doped (low resistivity) and the bodies of top tier trasistors are tied to the ground, we expect that some amount of electric field may penetrate the top tier silicon and coupling among top and bottom tier objects (M1, MB1, P, PB, etc.) may exist. When we assume that the top tier silicon is dielectric, the coupling between top and bottom tier objects would be overestimated; when it is conductor, the coupling would be underestimated. The real case would be between these two extreme cases.

The total cell internal RC values, extracted from the original 2-D cells and our 3-D (T-MI) cells, are shown in Table I.[5] For

[5]In this paper, we assume that copper is used for MB1. In a separate cell characterization run, we also assumed tungsten for MB1 [6]; however, no noticeable difference was found in cell timing and power.

TABLE I

CELL INTERNAL PARASITIC RC VALUES. THE 3D-C MEANS 3-D WITH TOP TIER SILICON MODELED AS A CONDUCTOR

cell	R ($k\Omega$)			C (fF)		
	2D	3D	3D-c	2D	3D	3D-c
INV	0.186	0.107	0.107	0.363	0.368	0.349
NAND2	0.372	0.237	0.237	0.561	0.586	0.547
MUX2	1.133	0.975	0.975	1.823	1.938	1.796
DFF	2.876	3.045	3.045	4.108	5.101	4.740

TABLE II

DELAY AND INTERNAL POWER CONSUMPTION OF CELLS WITH VARIOUS INPUT SLEW AND LOAD CAPACITANCE CONDITIONS. THE LIBRARY USES DIFFERENT INPUT SLEW SETTINGS FOR DFF. THE VALUES IN THE PARENTHESES MEAN THE PERCENTAGE RATIO OF 3-D TO 2-D

cell	delay (ps)		power (fJ)	
	2D	3D	2D	3D
fast case: input slew=7.5ps (5ps for DFF), load cap.=0.8fF				
INV	17.2	16.9 (98.3%)	0.383	0.351 (91.6%)
NAND2	21.2	20.9 (98.6%)	0.616	0.583 (94.6%)
MUX2	59.8	58.2 (97.3%)	2.113	2.060 (97.5%)
DFF	108.8	113.4 (104.2%)	6.341	6.735 (106.2%)
medium case: input slew=37.5ps (28.1ps for DFF), load cap.=3.2fF				
INV	51.1	50.8 (99.4%)	0.362	0.343 (94.8%)
NAND2	56.2	55.9 (99.5%)	0.604	0.581 (96.2%)
MUX2	97.0	95.3 (98.2%)	2.239	2.168 (96.8%)
DFF	142.6	147.0 (103.1%)	6.358	6.756 (106.3%)
slow case: input slew=150ps (112.5ps for DFF), load cap.=12.8fF				
INV	188.3	188.0 (99.8%)	0.449	0.431 (96.0%)
NAND2	195.9	195.5 (99.8%)	0.698	0.675 (96.7%)
MUX2	215.1	212.5 (98.8%)	2.555	2.487 (97.3%)
DFF	237.4	243.3 (102.5%)	7.303	7.659 (104.9%)

the 3-D case, the results with top tier silicon as both dielectric (3-D) and conductor (3D-c) are shown. From the results, we observe the followings.

1) For INV, NAND2, and MUX2, the R values of 3-D are noticeably smaller than 2-D counterparts because we reduce the length of poly and metal lines inside the cells, using 3-D interconnects.
2) The C values of 3-D are comparable with those of 2-D—the 2-D value is between 3-D and 3D-c.
3) For DFF, both R and C of 3-D are larger than 2-D counterparts. Due to the complex internal connections, we could not create a 3-D cell layout that matches RC parasitics of 2-D. In summary, depending on the cell layout complexity, the internal RC ratio between 3-D and 2-D may vary.

Yet, the delay and power of the cells are more important metrics. We perform cell timing/power characterizations using commercial softwares. The SPICE netlists obtained from the previous RC extractions are fed into Cadence Encounter Library Characterizer, which runs SPICE simulations to characterize delay and power of cells under various input slew and load capacitance conditions. The delay/power of 3-D and 2-D cells are shown in Table II. The values are obtained from the data tables in the characterized Liberty library. The delay is the cell internal delay including load effect, and the power is the dynamic power consumed within cell boundary (including short circuit power and power for gate/parasitic capacitances). We observe that for INV, NAND2, and MUX2, the delay and power of 3-D are slightly better than 2-D, whereas for DFF, they are a little worse. In addition, as the input slew and load capacitance condition changes from fast to slow case, the difference between T-MI and 2-D becomes smaller. Note that depending on cell design quality and manufacturing technology, the results may change. We believe that with proper cell designs, the delay and power of 3-D cells could be similar to 2-D counterparts.

C. Full-Chip Physical Layout

With the libraries built for T-MI, we proceed to full-chip layout experiments. Using Synopsys Design Compiler, we synthesize the benchmark circuits based on our T-MI standard cells and benchmark design constraints. These benchmark circuits are summarized in Table III. Next, we build physical layouts of the circuits using Cadence Encounter. Starting from floorplaning, we perform power delivery network planning, timing-driven placement of cells, clock synthesis, and timing-driven routing. Since a T-MI cell contains both the top and the bottom tier parts and MIVs as a single unit, the placer places the cells in a 2-D fashion without any overlap between cells. The T-MI cells have pins on the first metal of both the bottom and the top tiers [MB1 and M1 in Fig. 8(b)].

Unlike the metal layer assumption in [6], we allow our router to use the metal layer on the bottom tier [MB1 in Fig. 8(b)] for routing as well [5]. In this setup, the timing-driven router in Encounter chooses which pin on which layer to connect to, based on routing congestion and timing information.

After routing is finished, we perform RC extraction of nets, which is required for timing and power analysis. Once the RC information and the netlist are available, static timing analysis (STA) engine handles the entire top and bottom tiers at once, providing true 3-D STA results. Using Synopsys PrimeTime PX, we perform static power analysis. We assume certain switching activity values at the primary input pins and the flip-flop outputs (0.2 and 0.1, respectively). Then, the tool propagates switching activity information to the rest of the circuit. Based on the switching activity and library information, power calculation is performed.

Layout snapshots of AES (Table III) are shown in Fig. 6. In the zoom-in shots, cells, signal nets, and power rails are shown. For the top tier, only the first two metals (M1 and M2) are shown. We observe that Encounter places and routes T-MI cells without any problem. Note that MIVs used in net routing are placed in the white spaces between cells, avoiding any contact. Since we use the state-of-the-art EDA software for layout, the quality of placement and route is very good.

IV. EXPLORATION OF METAL LAYER OPTIONS

The metal layer structure of T-MI is dramatically different from conventional 2-D or TSV-based 3-D. In this section, we explore the metal layer options for T-MI that enable ultrahigh density integration. For this exploration, we use the benchmark circuits in Table III. Note that in this section,

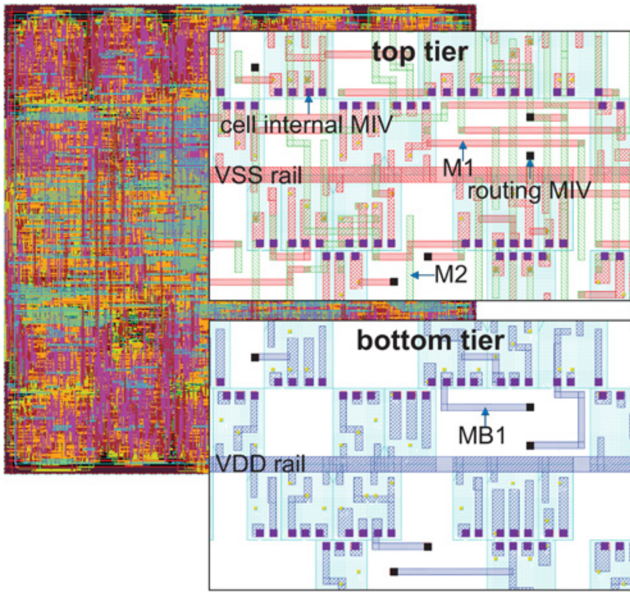

Fig. 6. Layout snapshots of the benchmark circuit AES. On the right, zoomed-in view shots of the top and the bottom tier are shown. Black and purple squares indicate the MIVs used for net routing and cell internal connections, respectively.

TABLE III
BENCHMARK CIRCUITS USED FOR METAL LAYER OPTION EXPLORATION

	AES	VGA	DES	JPEG	FFT
#cells	19,719	68,318	76,088	297,028	582,621
#nets	20,146	74,696	78,608	381,548	751,399
average fanout	2.131	2.307	2.034	1.850	2.130
clock period (ns)	0.5	0.5	0.5	3.0	0.6

TABLE IV
PIN DENSITY OF THE BENCHMARK CIRCUITS. CELL AREA AND PIN DENSITY (= #CELL PINS / CELL AREA) ARE SHOWN IN μm^2 AND $pins/\mu m^2$, RESPECTIVELY

		AES	VGA	DES	JPEG	FFT
#cell pins		63,068	247,015	238,488	1,087,390	2,351,692
cell area	2D	20,964	129,977	102,840	639,677	1,357,493
	T-MI	12,578	33,728	61,704	383,806	814,496
pin density	2D	3.01	1.90	2.32	1.70	1.73
	T-MI	5.01	3.17	3.87	2.83	2.89

we do not perform layout optimizations yet, to highlight the timing/power differences between interconnect options. Also, the same synthesized netlist is used for all design options.

A. Routing Congestions in T-MI Designs

Our preliminary study shows that routing congestion is a major problem in T-MI designs. Since our T-MI cells occupy 40% smaller footprints than the original 2-D cells, the overall chip footprint is reduced by about 40%. Yet, the number of cell pins to connect stays the same. As shown in Table IV, the pin density of T-MI becomes much higher than that of 2-D. For instance, the pin density of the T-MI design for AES is 66% higher than that of the 2-D design. The nets need to be routed within 40% smaller footprint, which means increased routing demand per unit area (or routing tile). The additional metal layer on the bottom tier of T-MI (MB1) can be used only for local interconnects because the MB1 strips inside cells (internal wires and pins) block cell-to-cell routing. Thus, the routing capacity (#routing tracks per routing tile) of T-MI per routing tile (= a tile in $N \times N$ grid for global routing) is almost the same as that of 2-D and cannot satisfy the much increased routing demand. To satisfy the high routing demand, we need to increase the routing capacity.

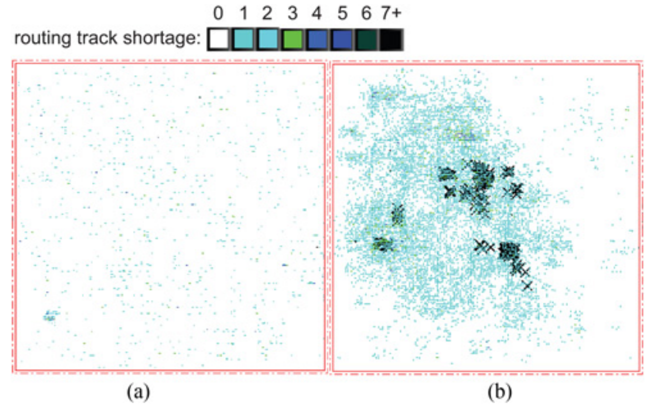

Fig. 7. Routing congestion map of VGA with (a) 2-D and (b) T-MI. Black X marks show design rule violations due to routing congestions.

Routing congestion maps of the 2-D and the T-MI design for a benchmark circuit are shown in Fig. 7. It is evident that T-MI shows more severe routing congestions than 2-D.[6] Because of metal layer changes and detours to deal with routing congestions, the timing and power quality of T-MI is also degraded. In addition, we observe that the routing congestion becomes more severe with timing optimization because the optimizer inserts buffers and breaks a complex cell into a group of simpler cells to improve timing, which in turn increases pin density considerably.

This routing congestion problem is unique in T-MI technology; it does not happen when the technology node is scaled down, because local metal dimensions and cells shrink at about the same rate. It does not happen for G-MI or TSV-based 3-D ICs either, because enough metal layers are available on each tier and the routing demand is satisfied.

To enable high density and high performance designs in T-MI technology, the routing congestion problem needs to be mitigated. Increasing the footprint of T-MI designs to reduce routing congestion is not a good idea because this reduces device density. In our study, we consider two kinds of metal interconnect modifications: 1) adding more metal layers, and 2) reducing metal dimensions.

B. Impact of Additional Metal Layers

1) *Additional Metal Layer Options:* Adding more local metal layers is an effective way to increase routing capacity and reduce congestion. The most area-efficient way is to add local metal layers because of the small pitch. We believe that

[6]The overall over-congestion rate (reported by Encounter, calculated from metal layers with maximum shortage) is 0.30% for 2-D case and 4.36% for T-MI.

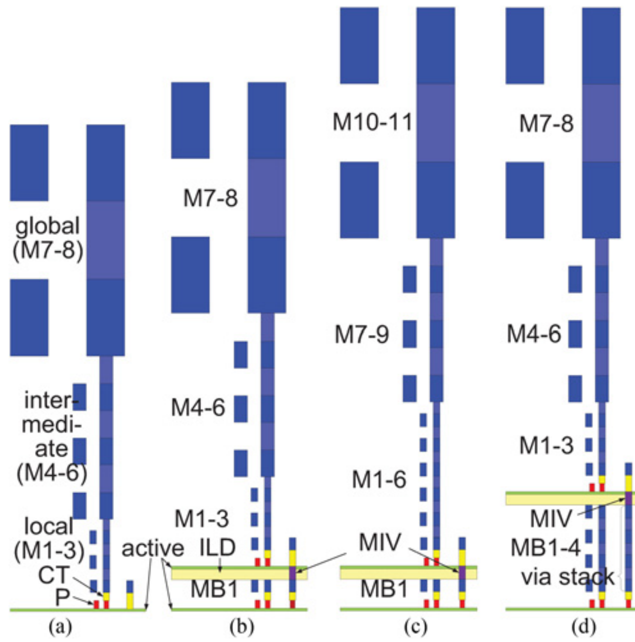

Fig. 8. Metal layer stack options. (a) 2-D, (b) baseline T-MI. (c) 3 local metal layers added to the top tier, (d) 3 local metal layers added to the bottom tier. ILD stands for interlayer dielectric between the top and the bottom tier. The bottom tier substrate and ILD for metal layers are not shown for simplicity. Objects are drawn to scale.

TABLE V
SUMMARY OF METAL LAYERS IN THE 2-D DESIGN OPTION. WE USE EIGHT OUT OF TEN METAL LAYERS IN THE NANGATE 45 NM LIBRARY. UNIT IS NM

level	metal layers	width	spacing	thickness
global	2D: M7-8	400	400	800
intermediate	2D: M4-6	140	140	280
local	2D: M2-3	70	70	140
first	2D: M1	70	65	130

more investment will be made to allow additional metal layers on the top and/or the bottom tier of monolithic 3-D ICs if there is a clear evidence that they improve the design quality of T-MI significantly. The baseline metal layer dimensions are summarized in Table V. As shown in Fig. 8, we now have three metal layer stack options for T-MI.

1) *1BM*: This is our baseline T-MI layer stack with one bottom tier metal layer.
2) *3TM*: We add three additional (local) metal layers to the top tier. As a result, we have total six local metal layers on the top tier.
3) *4BM*: We add three metal layers to the bottom tier. As a result, we have total four local metal layers on the bottom tier.

Due to manufacturing issues (low thermal budget), Bobba *et al.* [6] suggested tungsten is suitable for bottom tier metal. However, in this paper, we assume copper because a copper-based manufacturing process may be developed. Besides, MB1 is mostly used for short interconnects such as within cells or short nets. From our layout simulations, we found that the wirelength of MB1 for net routing is usually less than 1% of the total wirelength. Thus, the impact of MB1 material on the timing and power of a whole circuit is minimal. When tungsten is used, IR-drop on the VDD strips could be an issue, which is outside our scope.

TABLE VI
COMPARISON OF TIMING AND POWER OF A CELL WITH AND WITHOUT VIA STACK RC. THE VALUES ARE FROM THE TIMING/POWER TABLES OF THE CHARACTERIZED LIBRARIES

load	delay			power		
cap (fF)	without RC (ps)	with RC (ps)	diff. (%)	without RC (fW)	with RC (fW)	diff. (%)
0.4	28.4	31.2	9.86	1.15	1.33	15.65
0.8	33.1	35.8	8.16	1.40	1.52	8.57
1.6	42.8	45.4	6.07	1.86	1.98	6.45
3.2	62.4	64.9	4.01	2.81	2.99	6.41
6.4	100.3	103.0	2.69	4.78	4.93	3.14
12.8	175.8	179.9	2.33	8.54	8.74	2.34
25.6	330.0	330.6	0.18	16.17	16.33	0.99

2) *Via Stack Modeling for 4BM:* In the 4BM case, as shown in Fig. 8(d), the connections from a pMOS on the bottom tier to an nMOS on the top tier are made through metal and via layers on the bottom tier (MB1-4, VB1-3) and MIVs, which we call via stack in this paper. The physical size of a via stack is considerably larger than that of a single MIV. In addition, there could be metal interconnects surrounding a via stack, which may increase its coupling capacitance. Thus, we investigate the impact of RC parasitics of these via stacks on the timing/power of 4BM cells.

Using Synopsys Raphael, the capacitance of a via stack is extracted [5]. The capacitance of a via stack (C_{vs}) reported by Raphael is 0.123 fF. The resistance of a via stack (R_{vs}) is dominated by the resistances of local vias (VB1-3) and the MIV. From the values in the technology definition file, the calculated R_{vs} is 20 Ω, which includes contact resistances.

A lumped RC model of a via stack is incorporated into the SPICE netlist of each standard cell to characterize its timing/power behavior. The C_{vs} and R_{vs} of via stacks are inserted at the corresponding SPICE nodes. Then, we run Cadence Encounter Library Characterizer to characterize the timing and power of the modified standard cell for the 4BM case.

In Table VI, we compare the timing and power of a buffer cell with or without via stack RC. The delay includes both the cell intrinsic delay and load-dependent delay, and the power is the cell internal power, excluding wire switching and leakage power. In general, when the load capacitance of a cell is small, the impact of via stack RC on timing and power is large; the impact becomes smaller with larger load capacitance. This trend is observed in most of the cells. If a driving net is very short and has a small load capacitance, the timing and power of the driver may degrade by about 10%. From layout simulations, we found that the overall degradation of timing and power of the entire circuit is about 2%–3%, which is significant. Thus, we incorporate via stack RC in all of our 4BM designs.

TABLE VII

COMPARISON BETWEEN 2-D AND MONOLITHIC 3-D DESIGNS. #ROUTING MIVS MEANS THE NUMBER OF MIVS USED IN NET ROUTING, EXCLUDING THE MIVS USED INSIDE THE MONOLITHIC CELLS. THE WL, LPD, AND TNS MEAN WIRELENGTH, LONGEST PATH DELAY, AND TOTAL NEGATIVE SLACK, RESPECTIVELY. TOTAL POWER INCLUDES CELL INTERNAL, SWITCHING, AND LEAKAGE POWER. CLOCK POWER INCLUDES THE POWER OF CLOCK BUFFERS AND WIRES. THE VALUES IN PARENTHESES SHOW THE PERCENTAGE RATIO TO THE 2-D DESIGNS

circuit name	design type	footprint (μm^2)	total silicon area (μm^2)	total WL (m)	clk WL (mm)	#routing MIVs	LPD (ns)	TNS (μs)	total power (mW)	wire power (mW)	clock power (mW)
AES	2D	174x172	29,948	0.271 (100)	2.125	0	1.310 (100)	0.226 (100)	13.7 (100)	3.31 (100)	3.89 (100)
	1BM	135x134	35,938	0.209 (77.2)	1.819	1,070	1.260 (96.2)	0.202 (89.4)	13.6 (99.3)	2.93 (88.4)	4.26 (109)
	3TM	135x134	35,938	0.209 (76.9)	1.696	897	1.165 (88.9)	0.190 (83.9)	12.8 (93.4)	2.41 (72.7)	3.72 (95.5)
	4BM	135x134	35,938	0.214 (78.8)	1.866	3,266	1.226 (93.6)	0.207 (91.4)	13.7 (100)	2.64 (79.6)	4.20 (108)
VGA	2D	432x430	185,682	1.623 (100)	1.489	0	2.173 (100)	15.29 (100)	43.5 (100)	13.23 (100)	7.61 (100)
	1BM	334x333	222,822	1.284 (79.1)	1.236	2,349	1.954 (89.9)	13.01 (85.1)	41.8 (96.1)	11.59 (87.6)	7.48 (98.4)
	3TM	334x333	222,822	1.281 (78.9)	1.243	2,357	1.632 (75.1)	10.64 (69.6)	40.1 (92.2)	9.79 (74.0)	7.33 (96.3)
	4BM	334x333	222,822	1.363 (84.0)	1.179	18,020	1.843 (84.8)	11.27 (73.7)	43.8 (101)	11.22 (84.9)	8.24 (108)
DES	2D	384x382	146,916	0.849 (100)	32.77	0	1.086 (100)	0.581 (100)	134.9 (100)	24.81 (100)	66.1 (100)
	1BM	297x297	176,298	0.659 (77.6)	24.16	3,152	0.968 (89.1)	0.527 (90.8)	131.1 (97.2)	20.36 (82.1)	64.0 (96.8)
	3TM	297x297	176,298	0.654 (77.0)	24.54	3,121	0.923 (85.0)	0.503 (86.5)	126.1 (93.5)	19.36 (78.1)	60.1 (90.9)
	4BM	297x297	176,298	0.682 (80.3)	25.41	11,300	1.000 (92.1)	0.557 (95.8)	130.7 (96.9)	20.30 (81.8)	64.1 (97.0)
JPEG	2D	957x955	913,825	5.148 (100)	163.9	0	6.053 (100)	10.514 (100)	314.9 (100)	51.31 (100)	53.20 (100)
	1BM	741x740	1,096,592	4.032 (78.3)	126.4	16,502	5.422 (89.6)	2.999 (28.5)	300.6 (95.5)	42.92 (83.6)	46.70 (87.8)
	3TM	741x740	1,096,592	3.997 (77.6)	121.2	17,148	5.096 (84.2)	2.642 (25.1)	296.6 (94.2)	39.20 (76.4)	45.70 (85.9)
	4BM	741x740	1,096,592	4.160 (80.8)	127.9	71,944	5.967 (98.6)	4.018 (38.2)	312.6 (99.3)	41.56 (81.0)	50.10 (94.2)
FFT	2D	1394x1392	1,939,278	12.93 (100)	629.0	0	5.958 (100)	340.0 (100)	1469.2 (100)	295.9 (100)	1053.8 (100)
	1BM	1079x1079	2,327,134	10.41 (80.5)	463.6	30,407	4.250 (71.3)	299.0 (87.9)	1431.2 (97.4)	248.9 (84.1)	1025.3 (97.3)
	3TM	1079x1079	2,327,134	10.26 (79.4)	462.7	31,478	3.593 (60.3)	250.0 (73.5)	1345.4 (91.6)	226.9 (76.7)	948.1 (90.0)
	4BM	1079x1079	2,327,134	10.75 (83.2)	492.1	163,833	3.810 (63.9)	287.0 (84.4)	1535.8 (105)	245.2 (82.9)	1114.4 (106)

3) *Design and Analysis Results:* For a cell driving a net and the sink cells on the net, the delay (D) is

$$D_{total} = D_{cell} + D_{net} \tag{1}$$

$$D_{cell} = D_{intrinsic} + D_{load-dependent} \tag{2}$$

$$D_{load-dependent} = f_d(C_{load}, input\ slew) \tag{3}$$

$$C_{load} = C_{wire} + C_{pin}. \tag{4}$$

The $D_{intrinsic}$ is the intrinsic delay of the cell. The $D_{load-dependent}$ is a function of C_{load} and the signal slew at the cell input pin. Compared with 2-D designs, wires are shorter in T-MI designs, which in turn reduces C_{wire}, C_{load}, and $D_{load-dependent}$. The D_{net} also reduces as wires become shorter. However, the overall delay improvement may not keep up with wirelength reduction. If C_{pin} is larger than C_{wire}, the C_{load} may not decrease significantly because C_{pin} is not reduced. Moreover, $D_{intrinsic}$ also contributes to D_{cell}. Thus, depending on the circuit characteristics and layouts, the delay improvement of T-MI may vary. Meanwhile, the power consumption (P) of a cell is

$$P_{total} = P_{internal} + P_{switching} + P_{leakage} \tag{5}$$

$$P_{internal} = f_p(C_{load}, input\ slew) \tag{6}$$

$$P_{switching} \propto switching\ activity \times C_{load}. \tag{7}$$

The $P_{internal}$ is the power consumed for the objects within the cell boundary, which weakly depends on C_{load} and the cell input slew. When the input slew is larger, $P_{internal}$ increases. With our standard cell library (based on Nangate 45 nm library), $P_{leakage}$ is usually much smaller than $P_{internal}$ and $P_{switching}$. The $P_{switching}$ is proportional to both the switching activity and C_{load}. Assuming that the switching activity is the same for 2-D and T-MI designs, the reduction of C_{load} in T-MI designs is the main reason for the total power reduction. Note that if: 1) C_{pin} is more dominant than C_{wire}, or 2) $P_{internal}$ is more dominant than $P_{switching}$, the total power reduction of T-MI designs caused by wirelength reduction may not be significant.

The design and analysis results for 2-D and T-MI design options are summarized in Table VII. Placement utilization of all designs is 70%. Compared with 2-D designs, the footprints of T-MI designs are 40% smaller, while the total silicon areas are 20% larger. Compared with 2-D, the total wirelength and clock wirelength of all three T-MI design types are reduced by about 20%. The total number of MIVs used in routing is about the same for 1BM and 3TM, while 4BM utilizes considerably more MIVs because the bottom tier metals are highly utilized for routing.

The timing improvement of 3TM is the best among the T-MI design types. For the largest circuit (FFT), the longest path delay improvement of 3TM over 2-D is 39.7%. Note that this timing improvement can be used toward power reduction during the timing/power optimization; for the same target clock speed, 3TM may use more power-efficient (slower) cells to reduce power. However, the total power reduction of T-MI designs is less significant than timing improvement. The power reduction of T-MI designs over 2-D design is mostly from reduced wire power. However, wire power is only a small fraction of the total power. For instance, the wire power of JPEG for 3TM is 39.2 mW, which is only 13.2% of the total power. Depending on the quality of Encounter clock tree synthesis (CTS) results, the clock tree power may decrease. We observe that CTS usually produces the best results for 3TM among T-MI designs, because the CTS quality is related to the routing quality. The timing and power of 4BM designs are generally worse than 1BM and 3TM designs mainly because of the RC effect of via stacks inside cells.

TABLE VIII
IMPACT OF ADDITIONAL METAL LAYERS FOR 2-D

circuit name	design type	total wire-length (m)	longest path delay (ns)	total power (mW)
AES	2D	0.271	1.310	13.7
	2DM	0.272	1.250	13.6
	3TM	0.209	1.165	12.8
VGA	2D	1.623	2.173	43.5
	2DM	1.621	2.520	41.9
	3TM	1.281	1.632	40.1
DES	2D	0.849	1.086	134.9
	2DM	0.839	0.965	130.5
	3TM	0.654	0.923	126.1
JPEG	2D	5.148	6.053	314.9
	2DM	5.148	5.344	307.8
	3TM	3.997	5.096	296.6
FFT	2D	12.93	5.958	1469.2
	2DM	12.95	5.617	1441.6
	3TM	10.26	3.593	1345.4

TABLE IX
MINIMUM WIDTH/SPACING OF METAL LAYERS WITH VARIED METAL DIMENSION REDUCTION RATIO. FIRST METAL MEANS THE LOWEST METAL LAYER OF THE TOP/BOTTOM TIER. UNIT IS *nm*

reduction ratio (%)	0	10	20	30	40
global	400/400	360/360	320/320	280/280	240/240
intermediate	140/140	126/126	112/112	98/98	84/84
local	70/70	63/63	56/56	49/49	42/42
first	70/65	63/59	56/52	49/46	42/39

TABLE X
UNIT LENGTH RESISTANCE AND CAPACITANCE OF LOCAL METALS WITH VARIED METAL DIMENSION REDUCTION RATIO. THE C_{high} AND C_{low} ARE THE MAX/MIN TOTAL WIRE CAPACITANCE PER UNIT LENGTH, DEPENDING ON THE SURROUNDING WIRES

reduction ratio (%)	0	10	20	30	40
R ($\Omega/\mu m$)	3.57	4.41	5.59	7.29	9.93
C_{high} ($fF/\mu m$)	0.163	0.175	0.153	0.166	0.173
C_{low} ($fF/\mu m$)	0.104	0.105	0.107	0.108	0.111

C. Additional Metal Layers for 2-D

To see if the additional metal layers in 3TM majorly contributed to design quality improvement over 2-D, we add three metal layers to 2-D as well. *2DM*: We add three local metal layers to 2-D. The number of metal layers in 2DM is the same as that of top tier metal layers in 3TM. As shown in Table VIII, compared with 2-D, the additional metal layers in 2DM do not reduce total wirelength. Compared with 2-D, the timing and power of 2DM improved a little, because the additional metal layers reduced congestions and coupling capacitances. However, the timing and power of 3TM are still much better than those of 2DM. Thus, we conclude that the design quality improvement of 3TM over 2-D is mainly because of reduced footprint and wirelength.

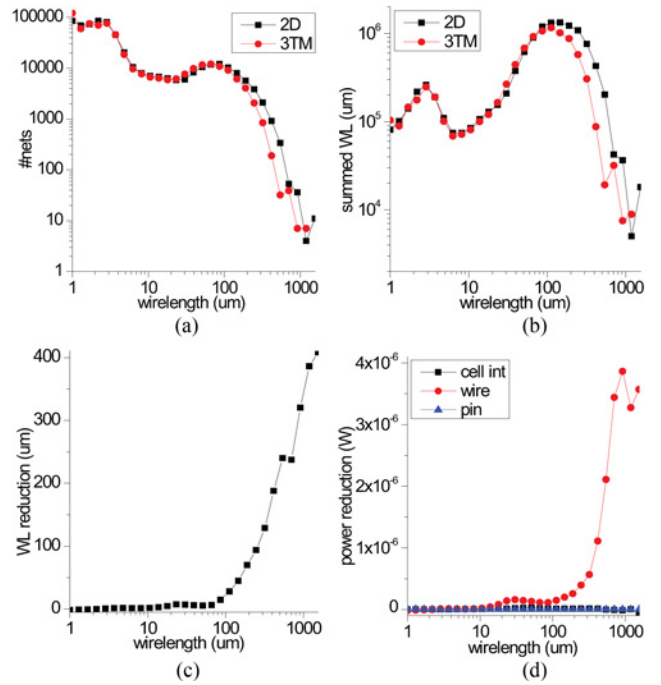

Fig. 9. Wirelength binning analysis for FFT: (a) wirelength distribution, (b) summed wirelength, (c) wirelength reduction, (d) power reduction. The x-axis is in log scale and represents wirelength bins.

D. Wirelength-Binning-Based Analysis

To further understand the timing and power improvement of T-MI, we plot the wirelength distribution. In Fig. 9(a), the wirelength distribution of 2-D and 3TM designs for the FFT circuit is shown. Yet, the wirelength distribution does not show which kinds of nets (short/medium/long) provide how much wirelength or power reduction. To answer this question, we perform a wirelength-binning-based analysis. From the layouts of 2-D and 3TM, we gather the metrics on each net such as wirelength, wire/pin cap and power, and driving cell power. We create wirelength bins by dividing wirelength range in log scale. Depending on the wirelength of the net, we assign the net into the corresponding wirelength bin. Then, we compare the improvement of 3TM over 2-D for the wirelength bins. Note that the improvement is calculated per net; for instance, for the same net the wire cap of 3TM is compared with that of 2-D.

From Fig. 9(b), we observe that the total wirelength per wirelength bin is the longest for the medium length nets (around 100 μm). Also, although there are only a few long nets, the summed wirelengths of long nets are significant. Note that for medium-long nets, the summed wirelength of 3TM is much shorter than that of 2-D, as shown in Fig. 9(c). As a result, the wire power reduction is larger for medium-long nets, as shown in Fig. 9(d). Since long nets tend to be on the critical path, reducing the wirelengths of long nets improves the critical path delay significantly.

For the majority of the nets, the wirelengths are very short ($< 10\,\mu$m). For short nets, the pin cap (C_{pin}) is dominant over the wire cap (C_{wire}). Thus, reducing the wirelengths of short nets will not improve the timing and power much. It is clear that wire power benefit is mostly from medium/long nets.

TABLE XI
TOTAL WIRELENGTH, LONGEST PATH DELAY, AND TOTAL POWER OF AES, VGA, DES, AND FFT WITH REDUCED METAL DIMENSIONS

		total WL (m)					LPD (ns)					total power (mW)				
reduction ratio (%)		0	10	20	30	40	0	10	20	30	40	0	10	20	30	40
AES	2D	0.271	-	-	-	-	1.310	-	-	-	-	13.7	-	-	-	-
	1BM	0.209	0.214	0.205	0.204	0.200	1.260	1.204	1.207	1.146	1.172	13.6	13.1	12.9	12.8	12.7
	3TM	0.209	0.208	0.203	0.197	0.196	1.165	1.206	1.147	1.152	1.133	12.8	12.8	12.9	12.8	12.8
	4BM	0.214	0.213	0.208	0.205	0.200	1.226	1.226	1.252	1.170	1.164	13.7	13.4	13.3	13.2	13.1
VGA	2D	1.623	-	-	-	-	2.173	-	-	-	-	43.5	-	-	-	-
	1BM	1.284	1.278	1.254	1.256	1.233	1.954	2.161	2.346	2.530	2.781	41.8	41.8	40.8	40.5	39.9
	3TM	1.281	1.255	1.254	1.242	1.236	1.632	2.007	1.728	2.522	2.586	40.1	39.4	39.3	39.1	38.9
	4BM	1.363	1.275	1.250	1.251	1.231	1.843	1.968	2.364	2.435	2.548	43.8	42.5	42.0	41.8	41.4
DES	2D	0.849	-	-	-	-	1.086	-	-	-	-	134.9	-	-	-	-
	1BM	0.659	0.656	0.647	0.644	0.639	0.968	0.927	0.947	0.924	0.941	131.0	128.6	130.6	132.6	126.2
	3TM	0.654	0.652	0.640	0.638	0.638	0.923	0.918	0.951	0.932	0.916	126.2	127.8	125.8	125.4	125.1
	4BM	0.682	0.657	0.646	0.638	0.637	1.000	0.969	0.972	1.030	0.953	136.6	136.2	132.0	132.0	131.7
FFT	2D	12.93	-	-	-	-	5.958	-	-	-	-	1469	-	-	-	-
	1BM	10.41	10.28	10.14	9.99	10.00	4.250	4.390	4.509	4.621	4.085	1431	1463	1361	1406	1344
	3TM	10.26	10.18	10.07	9.95	9.99	3.593	3.934	4.166	4.286	3.931	1345	1436	1431	1426	1327
	4BM	10.75	10.29	10.16	10.06	10.06	3.810	4.231	4.471	4.425	4.045	1536	1490	1524	1477	1469

E. Impact of Reduced Metal Dimensions

Another interconnect modification option to mitigate the routing congestion problem is to reduce the width, spacing, and thickness of metal layers. The local metal width/spacing is close to the minimum feature size of the technology node. However, if scaling down the metal dimensions brings large benefits in design quality, process engineers are willing to invest efforts toward it. Thus, the purpose of this metal dimension reduction study is to explore the interconnect design space for maximizing the benefit of T-MI; extreme scalings (>20%) may not be manufacturable with the technology node due to lithography limitations, chemical mechanical polishing issues, etc. For all T-MI cases (1BM, 3TM, and 4BM), we reduce the minimum metal width, spacing, and thickness of all metal layers up to 40% by 10% step. The diameters of vias and MIVs are also reduced to match the corresponding metal layers. Table IX summarizes the reduced metal width/spacing. Note that to keep the aspect ratio, the thickness of metal layers is also reduced, which is not shown in Table IX. Per each reduced metal dimension setting, the interconnect-related libraries such as capacitance table are rebuilt. Note that we do not modify the cell internal wires.

The unit length resistance and capacitance of local metal layers with reduced metal dimensions are summarized in Table X. As the width and thickness of a metal layer reduce, the unit length resistance of the metal layer increases. In constrast, the unit length capacitance of the metal layer does not change much. Note that depending on the surrounding wires, the unit length capacitance changes significantly (C_{high} versus C_{low}), mainly due to the difference in coupling capacitance. With reduced metal dimensions, more routing tracks are available. Thus, the router has a better chance for improving timing by carefully routing metal wires to reduce coupling capacitance. However, if the reduction ratio is too high, the metal resistance may increase the net delay and signal slew considerably.

Various design metrics of the JPEG circuit with varied metal dimension reduction ratio are shown in Fig. 10. The wirelength generally reduces as metal dimensions reduce, because of less routing congestion and detour. The number of clock buffers generally increases slowly when the reduction ratio increases. The reason is that as the metal dimensions decrease, the metal unit length RC increases, and the clock signal slew degrades. To meet the clock skew/slew specifications, the CTS engine inserts more buffers. For the LPD, the sweet spot of 1BM and 4BM cases is at the 30% reduction, while that of 3TM is 10%. Moreover, the LPD improvement of 4BM at the sweet spot over the default setting (=0% reduction) is larger than 1BM and 3TM cases. The wire power generally decreases with the reduced metal dimensions. However, we see that the cell internal power increases, which is also related to the signal slew degradation with reduced metal dimensions. As a result, the total power of 3TM and 4BM is minimum when the reduction ratio is 30%.

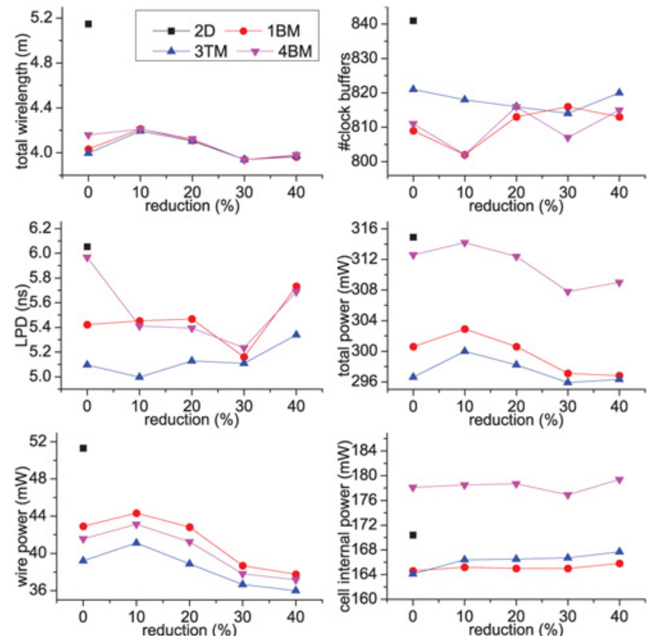

Fig. 10. Various results of JPEG with reduced metal dimensions.

TABLE XII
BENCHMARK CIRCUITS AND SYNTHESIS RESULTS

	FPU	AES	LDPC	DES	M256
target clock period (ns)	1.8	0.8	2.4	1.0	2.4
#cells	9,694	13,891	38,289	51,162	202,877
cell area (μm^2)	19,123	16,756	60,590	85,526	293,636
#nets	11,345	14,218	44,153	54,724	222,569
average fanout	2.35	2.40	2.38	2.33	2.23

The total wirelength, longest path delay, and total power of the other benchmark circuits are shown in Table XI. For total wirelength, the same trend as with JPEG is observed. The maximum wirelength reduction is 27.7% for AES with 3TM and 40% reduced metal dimensions. However, depending on the circuit characteristics, reducing metal dimensions may not translate to longest path delay reduction (see VGA and FFT results). In general, 3TM provides the most power improvement over 2-D designs. We observe that the maximum power reduction is 9.7% with 3TM and 40% reduced metal dimensions for FFT circuit. Note that depending on the benchmark circuit, the sweet spot changes.

From the simulation results in this section, we conclude that 3TM (=T-MI with three additional metal layers on the top tier) is the best option for T-MI. The reduced metal dimensions may further improve the design quality, however considering the increased cost and difficulties for manufacturing, it may not be a good option. Thus, in the following sections, we focus on 3TM without metal dimension reduction.

V. POWER BENEFIT STUDY

In this section, we study the power benefit of T-MI. We perform iso-performance comparison: under the same target clock period, the timing is closed for all design options and the power consumption is compared.

A. Benchmark Circuits and Synthesis Results

Our benchmark circuits and synthesis results are summarized in Table XII. The FPU is a double precision floating point unit. The AES and the DES are encryption engines. The LDPC is a low-density parity-check engine for the IEEE 802.3an standard. And the M256 is a simple partial-sum-add-based 256-bit integer multiplier. The circuits are in different sizes. We use Synopsys Design Compiler (ver. F-2011.09) for synthesis. The synthesis results are from 2-D results. All synthesized designs (2-D and T-MI) met target clock periods.

B. Layout Simulation Results

The layout simulation results are summarized in Table XIII. With T-MI, the footprint reduces by 40.9%–43.4%, which is larger than the cell footprint reduction rate, 40%. With T-MI, timing is better because of shorter wirelengths, and the optimizer may downsize cells and use less number of buffers while still meeting the target clock period. Thus, the footprint of the whole T-MI design could be further reduced than the individual cell footprint reduction rate. With T-MI, total wirelength reduces by 21.5%–33.6%. We observe that the circuit with a larger wirelength reduction rate tends to show a larger power reduction rate. All designs met the timing.

TABLE XIII
SUMMARY OF LAYOUT RESULTS. THE VALUES REPRESENT THE PERCENTAGE DIFFERENCE OF T-MI OVER 2-D

circuit name	footprint	total wirelen.	power			
			total	cell	net	leakage
FPU	-41.7%	-26.3%	-14.5%	-9.4%	-19.5%	-11.1%
AES	-42.4%	-23.6%	-10.9%	-7.6%	-13.9%	-9.5%
LDPC	-43.2%	-33.6%	-32.1%	-12.8%	-39.2%	-21.7%
DES	-40.9%	-21.5%	-4.1%	-1.6%	-7.7%	-1.4%
M256	-43.4%	-28.4%	-17.5%	-10.7%	-22.2%	-12.9%

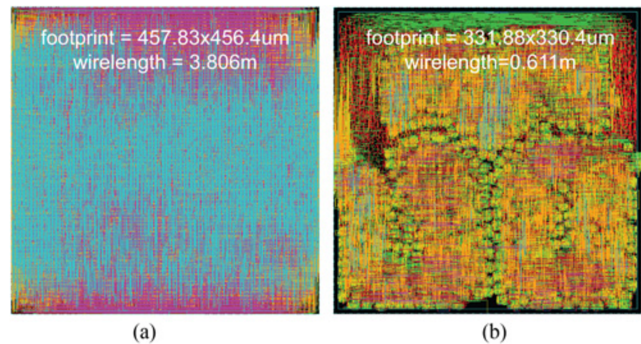

Fig. 11. Snapshots of routing results for T-MI designs. Cyan and magenta lines are global metal layers, whereas red, yellow, and green are local layers.

The power reduction was the largest in LDPC, 32.1%, whereas in DES, only 4.1%. The snapshots of routing results for these two circuits are shown in Fig. 11. In LDPC, the net power is much larger than the cell power, thus a large net power reduction with T-MI leads to a large total power reduction. We also observe that with T-MI, not only net power but also cell power reduces; with a better timing, cells are downsized and less number of buffers are used, to reduce cell power. In DES layout, there are many small regions where cells are tightly connected inside but not so much to outside. For these short nets, pin capacitances dominate wire capacitances, thus reducing wirelength does not reduce net power as much.

The detailed layout simulation results are shown in Table XIV, which supplements Table XIII. We set the final utilization (after all optimizations) to around 80%, which is a common practice in industry designs. Since we observed severe wire congestions in LDPC [Fig. 11(a)], the target utilization was lowered to about 33%; the 2-D design was barely routable with this setting. We also observed significant wire congestions in M256, thus the target utilization was lowered to 68%.

C. Comparison With Previous Work

Our results and the results from a previous work [6] are summarized in Table XV.[7] Both works use Nangate

[7]Note that the purpose of this paper is not to directly compare the design quality of ours to the previous works; due to various reasons (floorplan setup, design and analysis flow, optimization methods, target clock period, switching activity factors, etc.), it is not possible to provide fair comparisons.

TABLE XIV

LAYOUT RESULTS OF 2-D AND 3-D DESIGNS. THE 3-D MEANS OUR T-MI WITH 3TM METAL LAYER OPTION. THE #CELLS MEAN TOTAL NUMBER OF CELLS, AND #BUFFERS MEAN THE NUMBER OF INVERTING/NONINVERTING BUFFERS. THE #CELLS INCLUDE #BUFFERS. THE UTILIZATION MEANS FINAL CELL PLACEMENT DENSITY, AFTER ALL OPTIMIZATIONS. THE WL AND WNS MEAN WIRELENGTH AND WORST NEGATIVE SLACK, RESPECTIVELY. POSITIVE WNS VALUE MEANS TIMING IS MET WITH A POSITIVE SLACK. THE VALUES IN PARENTHESES SHOW THE PERCENTAGE RATIO TO THE 2-D DESIGNS

circuit name	design type	footprint (μm^2)	#cells	#buffers	utili-zation (%)	total WL (m)	WNS (ps)	total power (mW)	cell power (mW)	net power (mW)	leakage (mW)
FPU	2D	24,839 (100)	10,959	1,644 (100)	80.4	0.202 (100)	+6	8.44 (100)	3.98 (100)	4.21 (100)	0.25 (100)
	3D	14,476 (58.3)	9,922	1,240 (75.4)	79.5	0.149 (73.7)	+4	7.22 (85.5)	3.61 (90.6)	3.39 (80.5)	0.23 (88.9)
AES	2D	25,375 (100)	19,577	4,952 (100)	79.9	0.260 (100)	+30	13.69 (100)	6.36 (100)	6.94 (100)	0.40 (100)
	3D	14,613 (57.6)	18,996	5,157 (104.1)	79.7	0.199 (76.4)	+25	12.20 (89.1)	5.87 (92.4)	5.97 (86.1)	0.36 (90.5)
LDPC	2D	208,954 (100)	47,017	13,374 (100)	32.6	3.806 (100)	0	54.79 (100)	14.17 (100)	39.78 (100)	0.85 (100)
	3D	118,758 (56.8)	42,831	6,868 (51.4)	32.4	2.528 (66.4)	+12	37.22 (67.9)	12.36 (87.2)	24.20 (60.8)	0.66 (78.3)
DES	2D	109,652 (100)	54,402	8,436 (100)	79.9	0.611 (100)	+24	63.88 (100)	36.17 (100)	26.68 (100)	1.03 (100)
	3D	64,830 (59.1)	53,534	8,170 (96.8)	80.5	0.479 (78.5)	+32	61.24 (95.9)	35.60 (98.4)	24.62 (92.3)	1.02 (98.6)
M256	2D	478,077 (100)	245,935	62,970 (100)	68.2	6.647 (100)	0	194.6 (100)	74.73 (100)	115.2 (100)	4.70 (100)
	3D	270,748 (56.6)	216,956	48,125 (76.4)	67.3	4.760 (71.6)	0	160.5 (82.5)	66.70 (89.3)	89.66 (77.8)	4.10 (87.1)

TABLE XV

SUMMARY OF DESIGN RESULTS IN OUR WORK AND A PREVIOUS WORK. THE [6]-3D MEANS THEIR INTRACEL METHOD WITH TIMING DRIVEN + IPO, WHICH CORRESPONDS TO TRANSISTOR-LEVEL MONOLITHIC 3-D DESIGN

circuit name	design type	total wire-length (m)	longest path delay (ns)	total power (mW)
LDPC	ours-2D	3.806	2.400	54.79
	ours-3D	2.528 (-33.6%)	2.388	37.22 (-32.1%)
	[6]-2D	1.83	2.461	1,554
	[6]-3D	1.60 (-12.6%)	2.421	1,461 (-6.0%)
DES	ours-2D	0.611	0.976	63.88
	ours-3D	0.479 (-21.6%)	0.968	61.24 (-4.1%)
	[6]-2D	0.671	1.132	620.2
	[6]-3D	0.581 (-13.4%)	0.971	608.2 (-1.9%)

45 nm library as baseline 2D. The footprint reduction rate of 3-D over 2-D in this paper and [6] are about 42.3% and 30%, respectively. This footprint reduction rate mostly affects overall design quality of 3-D designs, because the timing and power reduction in the monolithic 3-D designs is from reduced footprint and wirelength. Our results show larger wirelength reduction than the previous work. In [6], they intentionally chose small target clock periods, thus timing was not closed. Note that power values in different works vary by much. For LDPC, our results show larger power reduction rate than the previous work. Interestingly, in both works, the power reduction rates for DES circuit are low (only 2%–4%).

VI. COMPARISON WITH G-MI AND TSV-BASED 3-D

In this section, we compare the design quality of T-MI designs with G-MI and TSV-based 3-D designs (TSV-3D). The layer structures of our G-MI and TSV-3D are shown in Fig. 12. Note that we assume two layers for G-MI and TSV-3D designs. For G-MI designs, we use six metal layers on the bottom tier and eight on the top. The reason why we use only six metal layers on the bottom tier is that the MIV pitch is determined by the top metal pitch on the bottom tier. If we use all eight metal layers because the minimum pitch of metal eight wires is large, the density of MIV becomes

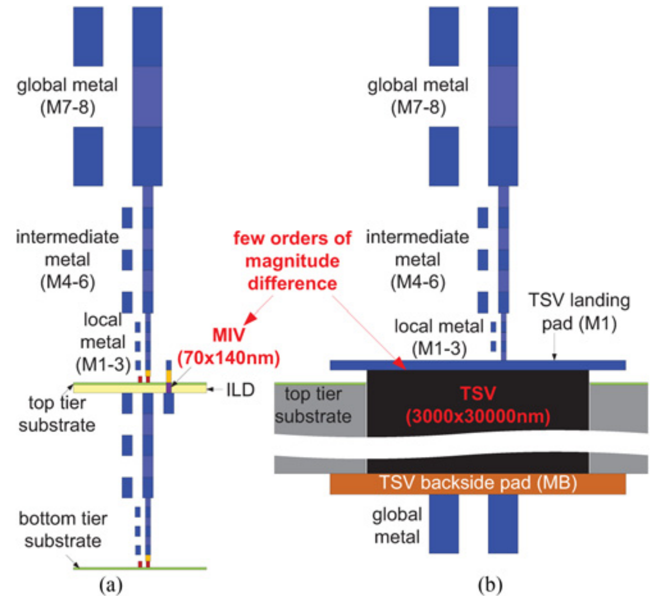

Fig. 12. Layer structures of (a) G-MI and (b) TSV-3D ICs. For simplicity, in (b), only the top metal layer of the bottom tier is shown.

small. For TSV-3D designs, we use eight metal layers on both top and bottom tiers because TSVs are large. The diameter and height of our TSV are 3 μm and 30 μm. Based on our physical assumptions such as TSV oxide liner thickness and doping concentration, using the parasitic RC models for TSVs [17], we determine that the resistance and capacitance of our TSVs are 1 Ω and 31.1 fF.

A. Design Flow and Its Limitation

Our design flows for G-MI and TSV-3D ICs are similar to [10]. Since today's commercial EDA tools cannot handle multiple dies together, we use on our in-house 3-D partitioner/placer [9] and timing-constraint-based iterative optimization method [10]. After the synthesis, we perform circuit partitioning.[8] We place the gates on Die 0/1 and

[8]As suggested in [9], we vary XY/Z-cut sequences to find the best layout results in terms of final timing and power.

TABLE XVI
LAYOUT RESULTS OF G-MI AND TSV-3D DESIGNS. THE VALUES IN PARENTHESES SHOW THE PERCENTAGE RATIO TO THE 2-D DESIGNS IN TABLE XIV

circuit name	design type	footprint (μm^2)	#cells	#buffers	util. (%)	total WL (m)	WNS (ps)	total power (mW)	cell power (mW)	net power (mW)	leak. (mW)
FPU	G-MI	12,100 (48.7)	11,532	2,048	84.2	0.195 (96.6)	+23	11.52 (102.9)	5.79 (101.1)	5.48 (105.0)	0.26
	TSV-3D	20,736 (83.5)	12,057	2,441	65.0	0.263 (130.1)	-28	14.25 (127.3)	6.61 (115.5)	7.29 (139.7)	0.35
AES	G-MI	12,544 (49.4)	17,618	4,135	77.6	0.219 (84.0)	-8	13.96 (96.9)	6.29 (102.0)	7.26 (92.7)	0.41
	TSV-3D	20,736 (81.7)	20,282	5,586	64.6	0.331 (127.4)	-334	19.18 (133.2)	7.67 (124.3)	11.0 (140.4)	0.51
LDPC	G-MI	108,900 (52.1)	52,705	19,043	35.1	3.089 (81.2)	+26	93.8 (98.3)	32.4 (131.2)	60.2 (86.1)	1.20
	TSV-3D	211,600 (101.3)	57,723	22,879	34.3	4.725 (124.1)	-939	138.0 (144.6)	36.6 (148.2)	99.6 (142.5)	1.82
DES	G-MI	58,564 (53.4)	60,666	12,250	78.7	0.645 (105.7)	+24	63.91 (109.6)	32.8 (108.6)	29.9 (110.3)	1.21
	TSV-3D	72,900 (66.5)	68,280	15,451	70.0	0.887 (145.3)	-26	74.38 (127.5)	34.7 (114.9)	38.3 (141.3)	1.38
M256	G-MI	260,100 (54.4)	281,320	78,320	68.5	6.657 (100.1)	-493	233.5 (118.9)	103.5 (118.8)	123.5 (118.1)	6.45
	TSV-3D	372,100 (77.8)	336,493	98,637	68.9	8.058 (121.2)	-908	282.17 (143.7)	116.5 (133.8)	158.0 (151.1)	7.67

MIVs/TSVs on Die 0 (= top tier), followed by a 3-D STA to generate the timing constraints on the die boundary ports (MIVs or TSVs). Then, per each die, we perform preroute optimizations, followed by a 3-D STA and timing constraint generation. As suggested in [10], we perform several iterations of optimizations to improve timing. After routing, we perform post-route optimizations in multiple iterations. Last, we perform the final 3-D STA and power analysis.

The most serious problem with die-by-die optimizations is the optimization quality. We cannot perform many effective optimizations in die-by-die optimization approach. The main reasons are: 1) the optimization engine cannot see the whole path; 2) it is not allowed to violate the logic equivalency at die boundary ports (MIVs or TSVs); 3) it is not allowed to move gates across the die boundary; and 4) it is not allowed to add/remove die boundary ports. For instance, when two buffers (one buffer on each die) were inserted for a two-pin 3-D net, we may convert the buffer pair to an inverter pair to reduce delay and power. However, since it would violate the logic equivalence check at the die boundary port, Encounter optimization engine cannot perform this conversion. In addition, the timing-constraint-based die-by-die optimization tends to use more buffers/inverters than necessary [18]. These limitations in optimizations degrade the timing and power of G-MI and TSV-3D designs.

B. Layout Simulation Results

The detailed layout simulation results for G-MI and TSV-3D designs are shown in Table XVI. The footprints are determined so that design is routable. Note that for TSV-3D cases, the footprints need to be increased significantly to accomodate TSVs. Comparing G-MI and TSV-3D results, we observe that in all aspects (wirelength, #buffers, timing, and power) G-MI is better than TSV-3D. This is mainly because MIVs are much smaller than TSVs in terms of physical dimensions and RC parasitics.

Comparing the G-MI and TSV-3D results with the T-MI results in Table XIV, we observe that the design quality of G-MI and TSV-3D is worse than that of T-MI. Possible reasons for this trend are as follows.

1) Placement quality of our 3-D placer is not as good as commercial 2-D EDA tool. Note that the wirelength of G-MI is much longer than that of T-MI.
2) As mentioned in Section VI-A, layout optimization quality in our G-MI and TSV-3D design flow is not as good as in T-MI or 2-D design flow.

Note that for many cases, we could not close the timing. Especially, when there are lots of long 3-D nets, the timing of G-MI or TSV-3D became worse than that of T-MI or 2D. These two reasons support the claim that T-MI produces better designs than G-MI or TSV-3D. In addition, for G-MI or TSV-based 3-D designs, we need true 3-D placement and optimization engines that can handle multiple dies together.

VII. CONCLUSION

In this paper, we investigated the benefits and challenges of monolithic 3-D IC technology. We identified the routing congestion problem that reduces the benefit of monolithic 3-D technology and studied interconnect options to overcome it. In transistor-level monolithic 3-D ICs, reduced footprints lead to shorter wirelengths, better performances, and lower power consumptions. With carefully designed transistor-level monolithic 3-D cells, we performed layout simulations and demonstrated up to 32.1% total power reductions. In contrast, because of the limitations in 3-D net optimizations, gate-level monolithic 3-D and TSV-based 3-D designs did not produce promising results. True 3-D EDA tools are necessary.

REFERENCES

[1] D. H. Kim, K. Athikulwongse, and S. K. Lim, “A study of through-silicon-via impact on the 3-D stacked IC layout,” in *Proc. IEEE Int. Conf. Comput.-Aided Design*, 2009, pp. 674–680.

[2] A. W. Topol, D. C. La Tulipe, L. Shi, S. M. Alam, D. J. Frank, S. E. Steen, J. Vichiconti, D. Posillico, M. Cobb, S. Medd, J. Patel, S. Goma, D. DiMilia, M. T. Robson, E. Duch, M. Farinelli, C. Wang, R. A. Conti, D. M. Canaperi, L. Deligianni, A. Kumar, K. T. Kwietniak, C. D’Emic, J. Ott, A. M. Young, K. W. Guarini, and M. Ieong, “Enabling SOI-based assembly technology for three-dimensional (3D) integrated circuits (ICs),” in *Proc. IEEE IEDM*, 2005, pp. 352–355.

[3] C. L. Yu, C. H. Chang, H. Y. Wang, J. H. Chang, L. H. Huang, C. W. Kuo, S. P. Tai, S. Y. Hou, W. L. Lin, E. B. Liao, K. F. Yang, T. J. Wu, W. C. Chiou, C. H. Tung, S. P. Jeng, and C. H. Yu, “TSV process optimization for reduced device impact on 28 nm CMOS,” in *Proc. Symp. VLSI Technol.*, 2011, pp. 138–139.

[4] P. Batude, M. Vinet, A. Pouydebasque, C. Le Royer, B. Previtali, C. Tabone, J.-M. Hartmann, L. Sanchez, L. Baud, V. Carron, A. Toffoli, F. Allain, V. Mazzocchi, D. Lafond, O. Thomas, O. Cueto, N. Bouzaida, D. Fleury, A. Amara, S. Deleonibus, and O. Faynot, "Advances in 3-D CMOS sequential integration," in *Proc. IEEE IEDM*, 2009, pp. 1–4.

[5] Y.-J. Lee, P. Morrow, and S. K. Lim, "Ultrahigh density logic designs using transistor-level monolithic 3-D integration," in *Proc. IEEE Int. Conf. Comput.-Aided Design*, 2012, pp. 539–546.

[6] S. Bobba, A. Chakraborty, O. Thomas, P. Batude, T. Ernst, O. Faynot, D. Z. Pan, and G. D. Micheli, "CELONCEL: Effective design technique for 3-D monolithic integration targeting high performance integrated circuits," in *Proc. Asia South Pacific Des. Autom. Conf.*, 2011, pp. 336–343.

[7] C. Liu and S. K. Lim, "A design tradeoff study with monolithic 3-D integration," in *Proc. Int. Symp. Quality Electronic Des.*, 2012, pp. 531–538.

[8] P. Batude, M. Vinet, A. Pouydebasque, L. Clavelier, C. LeRoyer, C. Tabone, B. Previtali, L. Sanchez, L. Baud, A. Roman, V. Carron, F. Nemouchi, S. Pocas, C. Comboroure, V. Mazzocchi, H. Grampeix, F. Aussenac, and S. Deleonibus, "Enabling 3-D monolithic integration," *ECS Trans.*, vol. 16, no. 8, pp. 47–54, Aug. 2008.

[9] M. Pathak, Y.-J. Lee, T. Moon, and S. K. Lim, "Through-silicon-via management during 3-D physical design: When to add and how many?" in *Proc. IEEE Int. Conf. Comput.-Aided Design*, 2010, pp. 387–394.

[10] Y.-J. Lee and S. K. Lim, "Timing analysis and optimization for 3-D stacked multicore microprocessors," in *Proc. IEEE Int. Conf. 3-D Syst. Integr.*, 2010, pp. 1–7.

[11] B. Rajendran, R. S. Shenoy, D. J. Witte, N. S. Chokshi, R. L. DeLeon, G. S. Tompa, and R. F. W. Pease, "CMOS transistor processing compatible with monolithic 3-D integration," in *Proc. VLSI Multi Level Interconnect Conf.*, 2005, pp. 76–82.

[12] S.-M. Jung, J. Jang, W. Cho, J. Moon, K. Kwak, B. Choi, B. Hwang, H. Lim, J. Jeong, J. Kim, and K. Kim, "The revolutionary and truly 3-dimensional $25F^2$ SRAM technology with the smallest S^3 (stacked single-crystal Si) cell, $0.16um^2$, and SSTFT (stacked single-crystal thin film transistor) for ultrahigh density SRAM," in *Proc. Symp. VLSI Technol.*, 2004, pp. 228–229.

[13] N. Golshani, J. Derakhshandeh, R. Ishihara, C.I.M Beenakker, M. Robertson, and T. Morrison, "Monolithic 3-D integration of SRAM and image sensor using two layers of single grain silicon," in *Proc. IEEE Int. Conf. 3-D Syst. Integr.*, 2010, pp. 1–4.

[14] T. Naito, T. Ishida, T. Onoduka1, M. Nishigoori, T. Nakayama, Y. Ueno, Y. Ishimoto, A. Suzuki, W. Chung, R. Madurawe, S. Wu, S. Ikeda, and H. Oyamatsu, "World's first monolithic 3D-FPGA with TFT SRAM over 90 nm 9 layer Cu CMOS," in *Proc. Symp. VLSI Technol.*, 2010, pp. 219–220.

[15] Nangate. (2008, Mar.). Nangate 45 nm Open Cell Library [Online]. Available: http://www.nangate.com/openlibrary

[16] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm design exploration," in *Proc. Int. Symp. Quality Electronic Des.*, 2006, pp. 717–722.

[17] G. Katti, M. Stucchi, K. D. Meyer, and W. Dehaene, "Electrical modeling and characterization of through silicon via for three-dimensional ICs," *IEEE Trans. Electron Devices*, vol. 57, no. 1, pp. 256–262, Jan. 2010.

[18] Y.-J. Lee, I. Hong, and S. K. Lim, "Slew-aware buffer insertion for through-silicon-via-based 3-D ICs," in *Proc. IEEE Custom Integr. Circuits Conf.*, 2012, pp. 1–8.

Young-Joon Lee (S'09) received the B.S. and M.S. degrees from Seoul National University, Seoul, Korea, in 2002 and 2007, respectively, and the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, USA, in 2013.

His current research interests include monolithic 3-D integrated circuit (IC) design automation, timing optimization, and low-power design techniques for through-silicon-via-based 3-D ICs, and co-optimization of traditional metrics and reliability metrics on 3-D ICs.

Sung Kyu Lim (S'94–M'00–SM'05) received the B.S., M.S., and Ph.D. degrees from the Computer Science Department, University of California, Los Angeles (UCLA), CA, USA, in 1994, 1997, and 2000, respectively.

From 2000 to 2001, he was a Post-Doctoral Scholar at UCLA, and a Senior Engineer at Aplus Design Technologies, Inc., Los Angeles, CA, USA. He joined the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2001, where he is currently a Full Professor. His research on 3-D integrated circuit (IC) reliability is featured as a research highlight in the communication of the ACM in 2013. He is the author of *Practical Problems in VLSI Physical Design Automation* (Springer, 2008). His current research interests include the architecture, circuit design, and physical design automation for 3-D ICs.

Dr. Lim received the National Science Foundation Faculty Early Career Development (CAREER) Award in 2006. He was on the Advisory Board of the ACM Special Interest Group on Design Automation from 2003 to 2008 and received the Distinguished Service Award in 2008. He was an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS from 2007 to 2009. He has served technical program committees of several premier conferences in electronic design automation. He received the Best Paper Award from TECHCON'11, TECHCON'12, and ATS'12. His work is nominated for the Best Paper Award at ISPD'06, ICCAD'09, CICC'10, DAC'11, DAC'12, and ISLPED'12. He was a member of the Design International Technology Working Group of the International Technology Roadmap for Semiconductors. He led the Cross-Center Theme on 3-D Integration for the Focus Center Research Program, Semiconductor Research Corporation, from 2010 to 2012.