Tier Degradation of Monolithic 3-D ICs: A Power Performance Study at Different Technology Nodes

Shreepad Panth, *Member, IEEE*, Sandeep Kumar Samal, *Student Member, IEEE*, Kambiz Samadi, *Member, IEEE*, Yang Du, *Member, IEEE*, and Sung Kyu Lim, *Senior Member, IEEE*

***Abstract*—Monolithic 3-D ICs (M3-D ICs) offer extremely high vertical interconnection density, significantly improving the power-performance envelope when compared to conventional 2-D ICs. However, process limitations lead to one tier having either degraded transistors or interconnects. This paper models the amount of degradation that can be expected at current and future nodes (45, 22, and 10 nm), develops a process development kit using these models to enable evaluation, and presents a block-level M3-D IC RTL-to-GDSII flow that is capable of mitigating some of this degradation. Experimental results indicate that at lower technology nodes, M3-D ICs offer more benefits. Results also indicate that the impact of transistor degradation is diminished at lower technology nodes while the impact of interconnect degradation becomes worse. Overall, perfect M3-D ICs close more than half the gap in the power-performance envelope between 2-D ICs and the "ideal" block-level design. While degraded tiers reduce the benefit of M3-D ICs, our degradation-aware floorplanner gives back up to 17% of the loss, and helps to obtain significant overall benefits compared to 2-D ICs.**

***Index Terms*—Block-level, floorplanning, monolithic 3-D IC (M3-D IC).**

I. Introduction

THREE dimensional integrated circuits (3-D ICs) [1] are emerging as a promising alternative to continue technology scaling. They reduce the length of the average interconnect [2], improving performance and power. Current 3-D ICs are enabled by through-silicon-vias (TSVs), which are large copper pillars that connect the stacked dies [3]. However, these are only useful when there are relatively few connections between the dies, such as memory on logic [4], [5], or stacked FPGAs [6]. TSV-based 3-D IC has matured to the point that reliability and cost has become a concern [7], and researchers have begun looking into cost-aware design flows [8].

An emerging alternative is monolithic 3-D ICs (M3-D ICs) that enables much higher integration densities due to the extremely small size (<100 nm) of the monolithic intertier vias (MIVs). In M3-D ICs, two or more tiers of devices are fabricated sequentially, instead of bonding prefabricated dies. This eliminates the need for die alignment, enabling smaller via sizes. The earliest M3-D process grew a thin layer of amorphous silicon on top of an already fabricated die, and then fabricated thin-film-transistors on top [9]. Next, attempts were made to epitaxially grow silicon on the top tier and crystallize it using lasers [10]. However, truly single crystal silicon was only obtained on the top tier by the wafer-cut process [11], which attaches single crystal silicon to the top of an already processed bottom tier, and then creates devices and interconnects as usual on top of this.

As with any technology, actual design methodologies, not just the fabrication process, heavily influence the final quality of results [12]. Several design styles exist for M3-D ICs. The first is transistor-level design, where transistors within standard cells are split up into multiple tiers [13]. Another design style is gate-level M3-D, where 2-D standard cells are placed on multiple tiers [14]. Finally, block-level design is where functional blocks are 2-D, but are floorplanned onto multiple tiers [15]. Block-level design offers a significant amount of IP reuse, and can be quickly deployed across different technology nodes.

Since M3-D ICs are still an emerging technology, it is crucial to study their potential benefits not only at existing technology nodes, but also at advanced nodes at which they are likely to become mature. However, there has been very limited work on studying the performance of M3-D ICs at advanced nodes [16]. In addition, as will be discussed in Section II, the M3-D IC fabrication process is not perfect, and a tier could have either degraded transistors or interconnects. Although this degradation could play a significant role in the performance of M3-D ICs, no prior work has studied it. The impact of this degradation could be different at lower technology nodes, so it is crucial to study its impact at both current and future nodes.

Therefore, the problem we are trying to solve can be stated as follows—At current and advanced technology nodes, determine the power-performance benefit of block-level M3-D ICs, how tier degradation affects this benefit, and if any loss of performance can be recovered using clever design techniques. In this paper, the current and future nodes considered are 45, 22, and 10 nm. The power-performance benefits are evaluated using an RTL-to-GDSII framework we develop for

Manuscript received July 22, 2016; revised October 11, 2016 and January 1, 2017; accepted February 9, 2017. Date of publication March 10, 2017; date of current version July 17, 2017. This paper was recommended by Associate Editor M. Ozdal.

S. Panth is with Intel Corporation, San Jose, CA 95134 USA (e-mail: shreepad.panth@intel.com).

S. K. Samal and S. K. Lim are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA.

K. Samadi and Y. Du are with Qualcomm Technologies Inc., San Diego, CA 92121 USA.

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.2017.2681064

TABLE I
I_{on} AND I_{off} FOR NORMAL AND DEGRADED TRANSISTORS ACROSS TECHNOLOGY NODES

Device	$I_{on}(mA/\mu m)$			$I_{off}(nA/\mu m)$		
	TT	M10P	M20P	TT	M10P	M20P
45nm-P	0.41 (1.00)	0.37 (0.90)	0.33 (0.80)	5.66	5.59	5.66
45nm-N	1.18 (1.00)	1.07 (0.91)	0.95 (0.80)	0.98	0.97	0.96
22nm-P	0.68 (1.00)	0.62 (0.90)	0.55 (0.80)	10.3	10.2	10.3
22nm-N	1.10 (1.00)	0.99 (0.90)	0.88 (0.80)	14.8	14.8	15.1
10nm-P	1.37 (1.00)	1.23 (0.90)	1.10 (0.80)	0.15	0.15	0.15
10nm-N	1.53 (1.00)	1.37 (0.90)	1.23 (0.81)	0.15	0.15	0.15

TABLE II
VARIOUS INTERCONNECT PARAMETERS. W_0 IS WIDTH

Parameter	Description	Copper	Tungsten
ρ_0	Bulk Resistivity ($\mu\Omega$-cm)	1.68	5.28
u	Line Edge Roughness	$0.4w_0$ [23]	$0.4w_0$ [23]
d	Dist. Betn. Grain Boundaries	w_0 [24], [25]	w_0 [24], [25]
λ	Electron Mean Free Path (nm)	39 [26]	19.1 [27]
p	Sidewall Specularity	0.2 [26]	0.3 [28]
R	Grain Boundary Reflectivity	0.3 [26]	0.25 [28]

block-level M3-D ICs. Tier degradation is handled through modeling the sources of transistor or interconnect degradation, and building a process development kit (PDK) that represents a degraded tier fabrication process. Finally, degradation-aware design is studied through the use of modifications to the basic floorplanner. The contributions of this paper are as follows.

1) This is the first work to study the power performance benefits of M3-D ICs over a wide technology range.
2) We model and study the system-level impact of interconnect and transistor degradation due to the M3-D fabrication process.
3) We develop an RTL-to-GDSII framework for block-level M3-D ICs.
4) We demonstrate that designs become less sensitive to transistor degradation and more sensitive to interconnect degradation in newer technologies.
5) We develop a degradation-aware floorplanner (DAFP) that can mitigate some of the effects of degradation.

II. Degradation Modeling And Node Scaling

In Section I, we simply stated that the transistors are fabricated on the top tier once it is attached. However, transistor fabrication usually involves a high temperature anneal step, which involves temperatures that damage both the underlying devices and copper interconnects. There are two ways of mitigating this issue: 1) use a low temperature process on the top tier, degrading the transistors or 2) use tungsten for the interconnects on the bottom tier [11], as it has a much higher melting point than copper. Each of these options have their own pros and cons, and modeling of these options are discussed in the following sections. In addition, the relative benefits of each option are bound to change with technology node, so it is essential to obtain models across technologies to accurately compare designs implemented in them. We choose the 45, 22, and 10 nm nodes to compare as they cover a significant technology shrink in uniform steps.

A. Transistor Degradation Modeling

If copper is to be used on the bottom tier, the top tier requires a low-temperature fabrication process. Laser-scan anneal has been proposed as a technique for the dopant activation of the top tier [17], [18]. This results in localized heating in the source/drain regions thereby preventing damage to the underlying devices and interconnects. However, the transistors are usually inferior to those obtained using a high temperature anneal, but the results are mixed. Rajendran *et al.* [18] specifically targeted M3-D IC, but the I_{on} of the pMOS and nMOS degrade by 27.8% and 16.2%, respectively, and there is also minor degradation in I_{off}. Ortolland *et al.* [17] do not specifically targeted M3-D IC, but instead just laser-scan regular HKMG transistors. They actually demonstrate no degradation in either I_{on} or I_{off}, but it is unclear how much this will carry over to the FDSOI transistors created on top of an already existing tier.

To keep this paper as general as possible, we refrain from taking degradation numbers from any specific work. We focus on modeling I_{on} degradation, as it is the primary performance metric. I_{off} only contributes to the leakage power of the chip. The impact of degradation in I_{off} on total leakage power is easier to compute and does not warrant an extra dimension in our analysis. This is particularly true as this paper performs power analysis assuming all blocks are active, so the leakage is a very small component of total power.

We use the ASU PTM transistor models [19] at the different technology nodes, and assume the generic case of either 10% or 20% degradation in the I_{on} of pMOS or nMOS, and designs will be evaluated at all combinations of these (e.g., pMOS = 20% degradation, nMOS = 10% degradation, etc.) we change the transistor parameters in the SPICE model to reflect such degradation, and tabulate the on and off currents for normal and degraded transistors in Table I. In this table, M10P with respect to pMOS implies that the regular pMOS transistors have been degraded by 10%, etc.

B. Interconnect Degradation Modeling

Tungsten is easier to process and has several attractive properties that make it a suitable choice for nano-scale interconnects, but its bulk resistivity is 3.1× that of copper, which has so far prevented its widespread use. When interconnects shrink, the bulk resistivity no longer applies, and resistivity goes up due to several size-related effects [20], [21]. Let w_0 be the width of an interconnect wire, and h_0 be its height (or thickness). Several empirical parameters affect the resistivity of an interconnect, and these are tabulated, along with the relevant literature from where they are obtained in Table II. The equation for size dependent resistivity of an interconnect is given by [22]. Let $\alpha = \lambda R/(dR(1-R))$. Then, the resistivity is given by

$$\rho_{\text{eff}} = \frac{\rho_0}{\sqrt{1-(u/w_0)^2}}\left\{\left[\frac{1}{9}-\frac{\alpha}{6}+\frac{\alpha^2}{3}-\frac{\alpha^3}{3}\ln\left(1+\frac{1}{\alpha}\right)\right]^{-1} + 0.45(1-p)\frac{\lambda}{w_0}\left(\frac{w_0}{h_0}+\frac{1}{1-(u/w_0)^2}\right)\right\}. \quad (1)$$

TABLE III
METAL LAYER DIMENSIONS (nm) AT DIFFERENT TECHNOLOGY NODES

	M1			M2-M3			M4-M6			M7-M8			M9-M10		
	45nm	22nm	10nm	45nm	22nm	10nm	45nm	22nm	10nm	45nm	22nm	10nm	45nm	22nm	10nm
Width	70	35	17.4	70	35	17.4	140	70	35	400	200	100	800	400	200
Spacing	65	32.4	16.2	70	35	17.4	140	70	35	400	200	100	800	400	200
Thickness	130	65	32	140	70	35	280	140	70	800	400	200	1600	800	400
Dielectric Height	120	60	30.5	120	60	30	290	145	72.5	820	410	205	2400	1200	600

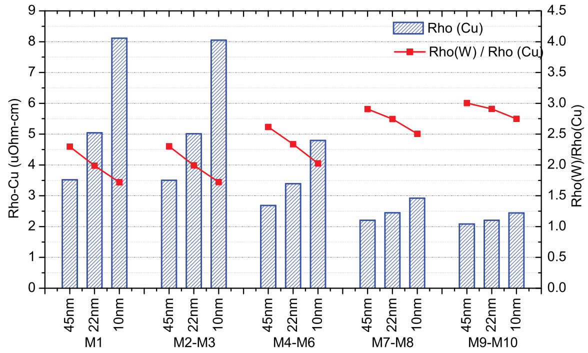


Fig. 1. Copper and tungsten resistivity values for different metal layers across technology generations.

We base our interconnect sizes from the Nangate 45 nm library, and scale them by a factor of 0.7× per technology node. The dimensions assumed are tabulated in Table III. The barrier thickness is the same across all metal layers, and we assume the thickness to be 5.2, 2.6, and 1.3 nm for the 45, 22, and 10 nm technology nodes, respectively. The resistivities obtained by plugging in these dimensions into (1) closely match ITRS predicted data for copper at all technology nodes [29], and the values for tungsten are also in close agreement with measured data from IBM [30].

We plot the resistivity of copper (as bars) for each of the metal layers across technology generations in Fig. 1. As expected, the lower metal layers have higher resistivity, but the difference in resistivity between different metal layers gets quite severe as we scale technology generations. We also plot the resistivity of tungsten divided by copper (as lines) in the same figure. The bulk resistivity of tungsten is approximately three times that of copper, but this resistivity scales better with size. Note that we have assumed that tungsten has an identical diffusion barrier to copper in these charts, which is not strictly necessary. Therefore, all tungsten numbers presented in this paper are a little pessimistic.

III. PDK GENERATION

Once we have the transistor and interconnect models, the next step is to build a PDK so that we can evaluate the impact using real designs. The overall PDK generation flow is shown in Fig. 2. The inputs to the flow are the transistor models and interconnect geometry and resistivity, which has already been discussed. In addition, we require standard cell layouts. We use the Nangate 45 nm layouts [31], and scale them by a factor of 0.7× for each technology generation.

For finFETs, an equivalent number of fins is derived based on the width and fin pitch. Note that there are inaccuracies introduced by this process, as the layout of standard cells is not guaranteed to be the same across technology generations. Aitken *et al.* [32] demonstrated that at lower nodes, the cell layout becomes more regular as design rules become more restrictive. However, there are two main reasons why the trends predicted by our results are valid. First, prior work has demonstrated that scaling such as this correlates well with ITRS predicted data [16]. Next, and perhaps more importantly, Sinha *et al.* [33] demonstrated that the result of increased regularity is an increase in the internal metal parasitics of the standard cell. This implies that trends predicted by our data underestimate the impact of interconnect at lower technology nodes. As will be explained in Section V, this makes our predictions more pessimistic (have more guardband).

The interconnect geometries are used to create several tech files (TF, ICT, and MIPT). These are used in Cadence TechGen, Cadence Encounter, and Mentor xCalibrate to generate tech files used for signoff extraction, capacitance tables for internal place and route stages, as well as parasitic extraction files to extract internal parasitics and generate SPICE netlists of standard cells using Mentor Calibre. These netlists, along with the transistor SPICE models are fed to Synopsys SiliconSmart for characterization to obtain a liberty timing model. Finally, the layout (in GDS format) of standard cells are fed into the "abstract" utility to generate library exchange format files that can be used for place and route.

We plot the normalized delay for various standard cells for both normal and degraded transistors (explained in Table IV) in Fig. 3. We also cover the case where the transistors are

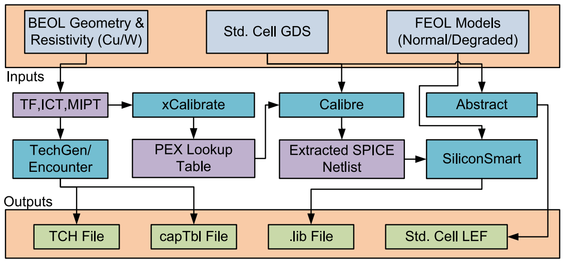


Fig. 2. PDK generation flow.

TABLE IV
NAMING CONVENTION FOR NORMAL AND DEGRADED LIBRARIES

	PMOS	NMOS	Wire
TT	Nominal	Nominal	Copper
TT_W	Nominal	Nominal	Tungsten
TT_PM10P_NM10P	Nominal - 10%	Nominal - 10%	Copper
TT_PM10P_NM20P	Nominal - 10%	Nominal - 20%	Copper
TT_PM20P_NM10P	Nominal - 20%	Nominal - 10%	Copper
TT_PM20P_NM20P	Nominal - 20%	Nominal - 20%	Copper

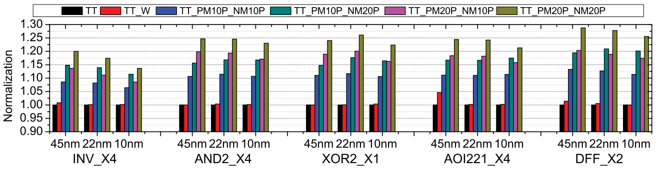


Fig. 3. Normalized delay numbers for regular and degraded standard cells across technology generations.

at nominal, but the intracell metal uses tungsten (TT_W). From this plot, we first observe that degraded transistors have a much bigger impact on the cell performance than degraded interconnect. We also observe a range of sensitivities depending on the cell type. For instance, the flip-flop is far more sensitive to transistor degradation than the inverter. We also observe that certain cells are more sensitive to one type of transistor degradation than the other. For example, in INV4_X1, TT_PM10P_NM20P has worse delay than TT_PM20P_NM10P. However, this trend is reversed in most other cells. In fact, this trend reversal happens within the same cell across technology generations, for instance in AOI221_X4. Finally, we observe a marginal decrease in sensitivity to transistor degradation at lower technology generations.

IV. Design and Analysis Flow

The objective of this paper is to study the impact that tier degradation has on the performance of an M3-D IC at different technology nodes. Such a study can be carried out using different design styles as the baseline, and this section first discusses the choice of gate-level versus block-level M3-D. It then presents the base flow along with modifications that are necessary to make it aware of tier degradation.

A. Choice of Design Style

The normal and degraded PDK that has been developed in the previous two sections can be used to evaluate the impact that tier degradation has on an M3-D IC for *any* design style. A gate-level M3-D design flow has been presented in [14], where Panth *et al.* presented a technique to utilize a commercial tool to design a single M3-D block where both tiers have identical performance. This flow essentially performs an initial placement and optimization step using a commercial tool and "shrunk libraries," followed by a partitioning step to obtain a gate-level M3-D design. This flow does not readily lend itself to degradation-awareness, as the base engine is a commercial tool, and cannot be modified. In addition, the partitioning step relies on correct choice of "bin size," which is a technology dependent parameter. It requires careful tuning at a given technology node, and making comparisons across different technology nodes is subject to tool noise if this parameter is not chosen appropriately. Finally, gate-level M3-D can only be used to evaluate the performance of a single-block, not the performance of an entire system.

These limitations motivate us to choose block-level M3-D for this paper. Floorplanning engines can be readily modified to support different cost functions, and hence lend themselves to degradation-awareness. In addition, the exact same engine can be used for different technology nodes, removing the need for technology-specific tuning and the associated tool noise. Finally, entire systems can be studied, not just a single block.

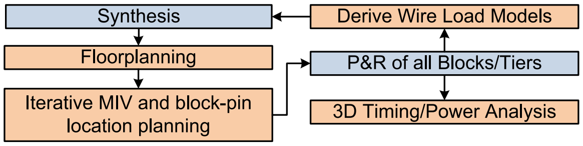


Fig. 4. Block-level RTL-to-GDSII design flow used in this paper. Orange indicates 3-D specific steps.

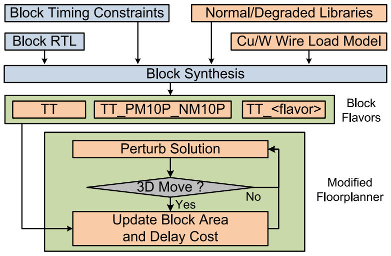


Fig. 5. Degradation Aware Floorplanning Methodology.

B. Overall Block-Level Design Flow

An overview of the flow is shown in Fig. 4. In this figure, orange boxes indicate 3-D specific steps. Once the design is synthesized, it is sent to our floorplanner (described in Section IV-C), which gives us the outlines of all the blocks in the 3-D space. Next, we perform MIV planning (described in Section IV-D) to determine all the MIV locations. With these locations, each block and tier is placed and routed (P&R) separately in Cadence Encounter. At this stage, we dump wire-load models and go back to synthesis to get a better synthesis result. Once the P&R is complete again, we proceed to 3-D timing and power analysis (described in Section IV-E).

C. Degradation-Aware Floorplanning

Several floorplanning works for TSV-based 3-D ICs exist in [34]–[37], each using a slightly different model to represent the 3-D wirelength or floorplan. However, [38] uses a wirelength model that is the closest representation of the final routed wirelength, so we base our floorplanner on this. This is a sequence-pair based floorplanner, and the main difference in our base engine is that ours is timing driven, achieved by weighting each interblock net by the longest path delay (LPD) through it. If we assume that both tiers have identical performance, the cost function for the floorplanner is simply the weighted sum of wirelength and footprint area. We now discuss one technique to make such a floorplanner aware that one tier has degraded performance.

In most designs, each block will have a different architecture or function, and hence will have different characteristics. The overall chip frequency will be decided by a few blocks that will be timing critical. As long as the floorplanner ensures that these critical blocks do not operate with slower transistors or interconnects, the chip can still meet timing.

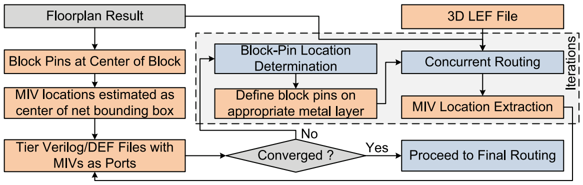

Floorplan Result
3D LEF File
Block Pins at Center of Block
Block-Pin Location Determination
Concurrent Routing
Iterations
MIV locations estimated as center of net bounding box
Define block pins on appropriate metal layer
MIV Location Extraction
No
Tier Verilog/DEF Files with MIVs as Ports
Converged ?
Yes
Proceed to Final Routing

Fig. 6. Iterative process to obtain both block-pin and MIV locations.

Some prior works have addressed process variations in a block-level TSV-based 3-D IC, although none have directly provided a variation or DAFP that can be used. Garg and Marculescu [39] presented the impact of process variation on critical path delay, but did not present any techniques to mitigate their impact. A simple, manual die-ordering mechanism to reduce the thermal impact in presence of process variation was provided in [40], but this algorithm was entirely manual, and was just changing the order of the dies. Building awareness (such as thermal-awareness) into a 3-D IC floorplanner can be done by modifying the cost function [41], but none exist that address degradation or variation. This is the first work to truly consider block degradation during 3-D IC floorplanning.

An overview of our DAFP is shown in Fig. 5. We first synthesize different versions of each block: one for the nominal corner, and one for each of the degraded libraries. In the case of tungsten interconnects, we also modify the resistivity of the wire load models to accurately drive synthesis. For each version of the block, we measure the area and LPD through it.

Given that both tiers have different performance, a block will have a different area and LPD depending on the tier in which it lies. If LPD(b_i) is the tier-dependant LPD of a block b_i, we define the modified cost function of the floorplanner as

$$\text{Cost}_{\text{DA}} = \alpha.\text{WL} + \beta.\text{Area} + \gamma \sum_{i=1}^{N_{\text{Block}}} \text{LPD}\ (b_i). \tag{2}$$

In the above equation, WL refers to the wirelength. The area of a block is also dependent on its tier. Therefore, whenever a 3-D move is made, we update the area of all the blocks that have changed their tier. The third term in the above equation will try to place the timing critical blocks in the faster tier, and push the nontiming critical blocks to the slower tier.

D. MIV Planning

The output of the floorplanner is block outlines in a 3-D space. Once we have these, we need to insert MIVs into the design. Existing TSV-based insertion algorithms cannot be used as TSVs are very large, require whitespace manipulation [37], careful manipulation of TSV-locations [36], or even simultaneous buffering, and TSV insertion due to their large capacitance [35]. MIVs are extremely small, and we can assume that whitespace is always available to insert them.

Given a set of hard blocks, Panth *et al.* [15] provided a methodology to determine MIV locations by tricking a 2-D router to perform 3-D routing. However, this requires the block-pin locations to be predetermined. In the case of soft blocks, the block pin locations are determined only after floorplanning is finished, so this methodology cannot be used directly. In this paper, we present an iterative approach to simultaneously determine both the block-pin and MIV locations, and it is outlined in Fig. 6.

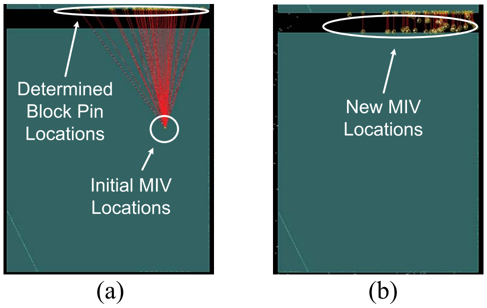


Fig. 7. Layout screenshots of our MIV planning methodology. (a) Initial estimated MIV locations (all MIVs overlap each other). (b) After one iteration of MIV planning.

First, given a floorplan result, where only the outlines of the blocks are known, we assume that all the block pins are in the center of each block. Next, for each 3-D net, we construct a 3-D bounding box of all the pins that it connects to, and compute the MIV to be in the center of the net bounding box, which could lead to MIV overlap, as shown in Fig. 7(a) ("initial MIV locations").

The next step is to create verilog and DEF files for each tier, such that the MIVs are represented as ports in the netlist. We then open the netlist of each tier in a commercial tool (Cadence Encounter), and use the tool's internal 2-D block-pin assignment capabilities to determine block pin locations based on the tier netlist and MIV locations. Essentially, the MIV locations defined as ports guide the selection of block pin locations. This is shown in Fig. 7(a) as "determined block pin locations."

The next step is similar to [15] in that we use a commercial 2-D router to mimic routing between the blocks in a 3-D space to determine new MIV locations based on the determined block-pin locations. This entire process can be repeated to refine the MIV locations. However, we observe that in one or two iterations, no further change in the wirelength is observed. A snapshot of the MIV and block pin locations after one iteration are shown in Fig. 7(b). Once the MIV locations are finalized, each block and tier can be P&R separately in Cadence Encounter.

E. 3-D Timing and Power Analysis

Once we have the P&R netlists of all the blocks and tiers, we load them into Synopsys PrimeTime. For each cell, depending on the tier in which it lies, we pick the appropriate std. cell library (normal or degraded). We also use the appropriate interconnect extraction tech file for a block depending on whether it is on a tier that uses tungsten or copper interconnects. These parasitics are also loaded into Synopsys PrimeTime. We also create a top-level netlist and parasitic file to represent the MIV connectivity and parasitics. According to [11], if the intertier oxide thickness is greater than or equal to 100 nm, there is negligible intertier coupling. Therefore, we ignore any such coupling in this paper. Once all

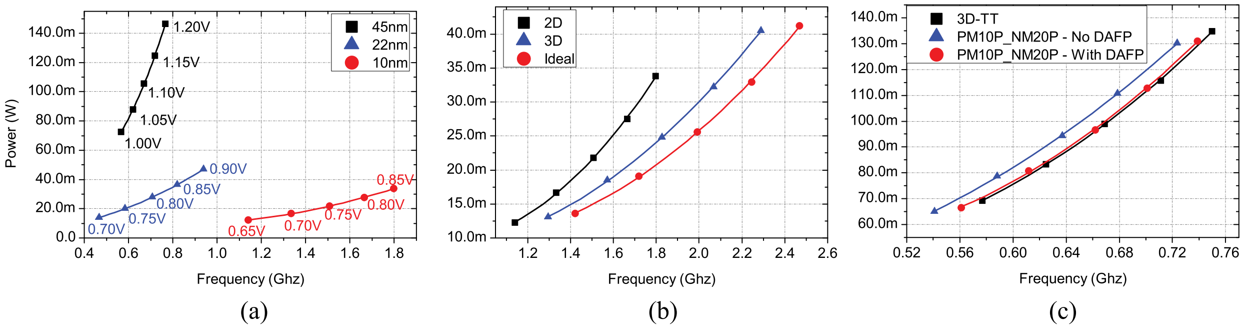


Fig. 8. Power-performance envelopes for the des3 benchmark. (a) 2-D design across technology nodes. (b) 2-D versus 3-D versus ideal at the 10 nm node. (c) Impact of degradation-aware floorplanning for the case, where the top-tier is at the TT_PM10P_NM20P corner.

the netlists and parasitics are loaded, we perform 3-D timing analysis, and statistical power analysis using PrimeTime.

V. Experimental Results

We pick one benchmark from the OpenCores benchmark suite (des3—55 blocks, 60k gates, and 6k interblock nets), one from the IWLS benchmark suite (b19—55 blocks, 80k gates, and 14k interblock nets), and design one custom 128-bit integer multiplier (63 blocks, 250k gates, and 12k interblock nets). “des3” and “b19” are timing critical from both an intrablock and interblock perspective, while “mul128” has long timing paths mainly within the blocks.

A. Deriving Power-Performance Envelopes

In this paper, we over-constrain the frequency during the design to obtain the fastest possible implementation, and then analyze it at several different voltages to obtain the power-performance envelope. To do this, in addition to characterizing the PDK at the nominal voltage for given technology, we also characterize four additional voltages in increments of ±50 mV. A given design is then analyzed at all five voltages, giving five points in the power versus frequency curve. A quadratic curve is fit to these five points to obtain the power-performance envelope.

An example curve for the 2-D implementation of des3 across all technology generations is shown in Fig. 8(a). The data points obtained by Primetime timing and power analysis are shown as points, while the fitted quadratic curve is shown as a line in between the points. The actual voltages involved in analyzing the power/performance are also shown. From these curves, it is clear that the 22 nm process node primarily offers a power saving, while the 10 nm process node offers a performance boost. Note that these curves are very sensitive to the actual transistor *I*–*V* curves, and will change with different transistor models.

B. Benefits of M3-D With Perfect Fabrication Process

Since 3-D ICs only reduce the interblock wirelength and do not significantly affect the intrablock power and performance, we also define an “ideal” block-level implementation to compare against. This implementation is obtained by assuming that all the interblock nets have zero length and parasitics. During the block implementation, we set the output load of the blocks to be zero and the inputs to be driven by ideal drivers. This is the theoretical lower bound on any block-level implementation of this design, given the same set of blocks that are all implemented in 2-D.

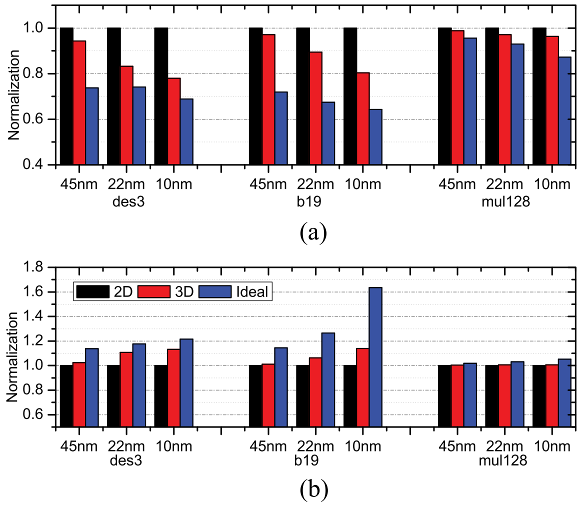
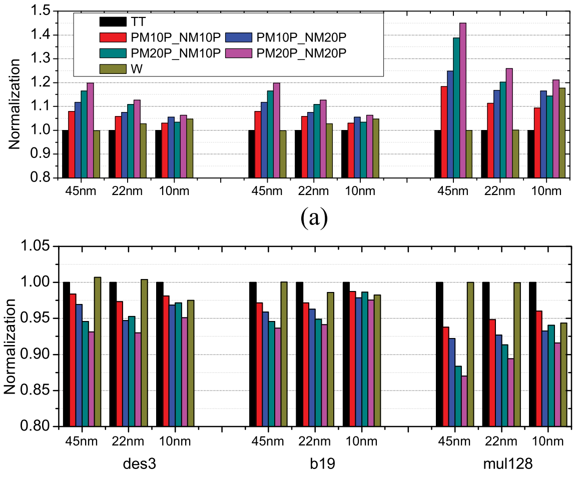


Fig. 9. 2-D versus M3-D versus ideal block-level implementation across different technology generations. (a) Power at same frequency. (b) Frequency at same power.

Power-performance envelope curves for 2-D, 3-D, and ideal for des3 at the 10 nm technology node are shown in Fig. 8(b). We clearly see that 3-D improves the envelope from 2-D, and closes the gap to the ideal implementation quite significantly. To get a more quantitative analysis, we measure the power at the same frequency as well as the frequency at the same power for all designs across technologies, and plot it in Fig. 9. From these graphs, we clearly see that 3-D offers more benefit at lower technology generations. In a design where the interblock wires are more dominant than intrablock wires, we expect that a reduction in the interblock wirelength will have a greater impact on the total power or performance of the design, as seen from this graph.

C. Tier-Degradation Induced Power/Performance Loss

We now take the existing 3-D floorplan and analyze the power-performance envelope due to both transistor and interconnect degradation. For degraded transistors, we simply



(a)

(b)

Fig. 10. Tier-degradation induced power/performance loss of an M3-D IC. (a) Power at same frequency. (b) Frequency at same power.

replace the libraries in PrimeTime. However, for degraded interconnects, this requires re-extraction of the entire degraded tier with the tungsten extraction tech file. These results are plotted in Fig. 10.

The overall frequency reduction due to a degraded tier depends on what percentage of the critical path lies on that tier, and the proportion of cell delay and wire delay. From Fig. 10, we observe that for implementation in lower technologies, the system performance is affected less by degraded transistors and more by degraded interconnects. This follows the expected trend as the transistors get better while the interconnects worse. In general, tungsten interconnects are the better option at the 45 and 22 nm technology nodes, and is only the better option at 10 nm if both the pMOS and nMOS degradation cannot be contained to less than 20%. Finally, we observe that the sensitivity to pMOS/nMOS degradation changes with different technology nodes, likely due to strain engineering that has sought to equalize the pMOS and nMOS characteristics. In the 45 nm node, it is always favorable to have nMOS degradation over pMOS degradation. In the 22 nm node, results are mixed, and in the 10 nm node, the sensitivities are roughly equal.

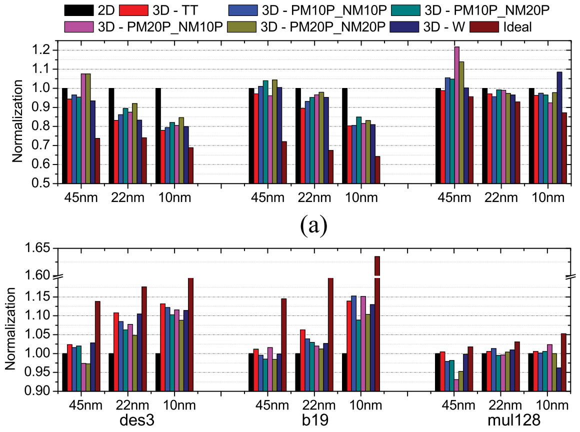
Now, if we wish to have the same frequency for nominal and degraded circuits, the only option is to boost the system voltage, which will lead to power increase roughly proportional to V^2. The power increase trend roughly follows the frequency reduction one, except for the fact that magnitudes are increased due to the square relationship with voltage. If standard cell are designed specifically for lower technology nodes with increased regularity, etc., this trend will only be emphasized. Note that the reason why the mul128 power increase is much higher than other benchmarks is because the critical path is completely within a block. Quite a large voltage increase is needed to bring the speed of the degraded blocks back to the nominal one.

D. Impact of Degradation-Aware Floorplanning

We now generate 3-D floorplans using the technique presented in Section IV-C to mitigate some of this degradation.

TABLE V
NORMALIZED IMPROVEMENT IN THE POWER-PERFORMANCE ENVELOPE BY DEGRADATION-AWARE FLOORPLANNING

Flavor	Power at Same Freq.			Freq. at Same Power		
	45nm	22nm	10nm	45nm	22nm	10nm
des3						
PM10P_NM10P	0.982	0.989	0.979	1.008	1.005	1.011
PM10P_NM20P	0.932	0.982	0.986	1.028	1.010	1.007
PM20P_NM10P	0.977	0.967	0.968	1.009	1.019	1.017
PM20P_NM20P	0.953	0.977	0.975	1.020	1.013	1.013
W	1.009	1.025	0.984	0.996	0.990	1.008
b19						
PM10P_NM10P	0.968	0.989	0.926	1.013	1.006	1.033
PM10P_NM20P	0.962	0.995	1.012	1.015	1.005	0.994
PM20P_NM10P	0.865	0.984	0.938	1.060	1.010	1.025
PM20P_NM20P	0.910	0.982	0.978	1.038	1.010	1.010
W	1.037	1.048	0.983	0.987	0.978	1.006
mul128						
PM10P_NM10P	0.908	0.892	0.911	1.039	1.062	1.040
PM10P_NM20P	0.867	0.885	0.847	1.059	1.066	1.076
PM20P_NM10P	0.889	0.861	0.835	1.048	1.083	1.086
PM20P_NM20P	0.825	0.826	0.844	1.087	1.111	1.084
W	1.016	0.992	0.974	0.993	1.005	1.010



(a)

(b)

Fig. 11. Power-performance comparisons between 2-D and 3-D with and without tier-degradation, and the ideal block-level implementation. (a) Power at same frequency. (b) Frequency at same power.

We tabulate the results of this floorplanning in Table V, and also plot a sample of the improvement achieved in Fig. 8(c).

This table shows that in general, we can expect a few percentage of improvement by degradation-aware floorplanning. However, in the case of mul128, the critical paths are entirely within a block, so the amount of improvement obtained by simply moving these blocks to a nondegraded tier is quite significant, and we see up to a 17.5% improvement in the power at the same frequency. We also note that degradation-aware floorplanning is not always successful, and mainly fails to give a better result when the amount of degradation is not severe in the first place, for example, tungsten interconnects in the 45 nm technology node.

E. Overall Comparisons

In this section, we compare the 2-D design, all flavors of the 3-D designs, and the ideal implementation across technology nodes. Basic floorplan comparisons of wirelength, footprint area, and the number of MIVs used are tabulated in Table VI.

TABLE VI
NORMALIZED FLOORPLAN COMPARISONS BETWEEN 3-D WITH AND WITHOUT TIER-DEGRADATION, AND THE IDEAL BLOCK-LEVEL IMPLEMENTATION

Flavor	45nm			22nm			10nm		
	FP Area	Total WL	MIV $\times 10^3$	FP Area	Total WL	MIV $\times 10^3$	FP Area	Total WL	MIV $\times 10^3$
des3									
TT	0.53	0.85	3.1	0.55	0.84	3.4	0.57	0.81	3.1
PM10P_NM10P	0.55	0.82	3.3	0.54	0.86	3.9	0.59	0.81	3.9
PM10P_NM20P	0.56	0.82	3.3	0.58	0.86	3.1	0.60	0.80	4.4
PM20P_NM10P	0.55	0.82	3.2	0.55	0.91	4.6	0.57	0.80	3.9
PM20P_NM20P	0.52	0.91	4.6	0.52	0.84	3.8	0.57	0.82	4.7
W	0.51	0.87	3.8	0.52	0.88	4.2	0.54	0.79	4.1
Ideal	-	0.64	-	-	0.67	-	-	0.63	-
b19									
TT	0.47	0.92	14.5	0.54	1.02	13.8	0.50	1.04	14.6
PM10P_NM10P	0.50	0.92	14.1	0.54	1.01	13.4	0.55	1.05	14.5
PM10P_NM20P	0.46	0.94	14.3	0.54	1.00	12.8	0.50	1.08	14.6
PM20P_NM10P	0.49	0.90	14.2	0.55	1.00	13.3	0.52	1.05	14.6
PM20P_NM20P	0.47	0.92	14.1	0.57	1.01	12.9	0.49	1.05	13.9
W	0.47	0.92	14.1	0.53	1.01	13.8	0.49	1.06	14.0
Ideal	-	0.62	-	-	0.69	-	-	0.74	-
mul128									
TT	0.48	0.95	8.7	0.50	0.91	8.4	0.49	0.91	10.1
PM10P_NM10P	0.53	0.99	4.5	0.52	0.93	4.5	0.51	0.92	4.5
PM10P_NM20P	0.49	0.98	4.5	0.52	0.93	4.5	0.51	0.92	4.5
PM20P_NM10P	0.49	0.97	4.5	0.50	0.93	4.5	0.53	0.93	4.5
PM20P_NM20P	0.49	0.98	4.5	0.51	0.92	4.5	0.52	0.92	4.5
W	0.49	1.01	4.9	0.52	0.93	4.8	0.54	0.94	4.7
Ideal	-	0.76	-	-	0.74	-	-	0.76	-

We also plot the improvement in the frequency and power in Fig. 11. Note that there is some tool noise introduced in this plot as the degradation-aware floorplanning is not always consistent across floorplans due to the random nature of simulated annealing. We observe that at 45 nm, 3-D without degradation always offers a benefit over 2-D ICs, but this benefit may go away if there is excessive tier-transistor degradation. Interconnect degradation is the preferred choice at this technology node. At 22 nm, any option for tier-degradation still gives designs that are better than 2-D ICs, but the benefit gets reduced with more degradation. Similar trends hold true for 10 nm as well, but as the baseline improvement from 2-D to 3-D is much higher and the sensitivity to degradation is lower, tier transistor degradation does not significantly erode the benefits offered by 3-D. However, at this node, transistor degradation is preferred to interconnect degradation, as tungsten interconnects at 10 nm sometimes even fail to produce designs better than 2-D ICs.

VI. CONCLUSION

Due to an imperfect manufacturing process, M3-D ICs will have either degraded transistors or interconnects in one tier. This paper models the amount of degradation that can be expected at current and future nodes (45, 22, and 10 nm), develops a PDK using these models to enable evaluation, and presents a block-level M3-D IC RTL-to-GDSII flow that is capable of mitigating this degradation. Experiments indicate that at lower technology nodes, M3-D ICs offer more benefits and are also more resistant to transistor degradation. However, they become more susceptible to interconnect degradation. At 45 nm, it is preferable to go with degraded interconnects. However, at lower nodes, especially at 10 nm, degraded transistors are the better choice. The DAFP can help recover up to 17% of the loss in the power-performance envelope, and overall, M3-D ICs can close more than half the gap in the power-performance envelope between 2-D and the ideal block-level implementation.

REFERENCES

[1] W. R. Davis *et al.*, "Demystifying 3D ICs: The pros and cons of going vertical," *IEEE Des. Test Comput.*, vol. 22, no. 6, pp. 498–510, Nov./Dec. 2005.

[2] R. Venkatesan, J. A. Davis, K. A. Bowman, and J. D. Meindl, "Optimal n-tier multilevel interconnect architectures for gigascale integration (GSI)," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 9, no. 6, pp. 899–912, Dec. 2001.

[3] S. Borkar, "3D integration for energy efficient system design," in *Proc. ACM Design Autom. Conf.*, San Diego, CA, USA, Jun. 2011, pp. 214–219.

[4] P. D. Franzon, W. R. Davis, T. Thorolfsson, and S. Melamed, "3D specific systems: Design and CAD," in *Proc. IEEE Asian Test Symp.*, New Delhi, India, Nov. 2011, pp. 470–473.

[5] R. G. Dreslinski *et al.*, "Centip3De: A 64-core, 3D stacked, near-threshold system," in *Proc. IEEE Hot Chips Symp.*, Cupertino, CA, USA, Aug. 2012, pp. 1–30.

[6] T. Thorolfsson and P. D. Franzon, "System design for 3D multi-FPGA packaging," in *Proc. IEEE Elect. Perform. Electron. Packag.*, Atlanta, GA, USA, Oct. 2007, pp. 171–174.

[7] R. Topaloglu, "Applications driving 3D integration and 1 corresponding manufacturing challenges," in *Proc. ACM Design Autom. Conf.*, San Diego, CA, USA, Jun. 2011, pp. 220–223.

[8] Q. Zou, J. Xie, and Y. Xie, "Cost-driven 3D design optimization with metal layer reduction technique," in *Proc. Int. Symp. Qual. Electron. Design*, Santa Clara, CA, USA, Mar. 2013, pp. 294–299.

[9] T. Naito *et al.*, "World's first monolithic 3D-FPGA with TFT SRAM over 90nm 9 layer Cu CMOS," in *Proc. IEEE Int. Symp. VLSI Technol.*, Honolulu, HI, USA, Jun. 2010, pp. 219–220.

[10] S.-M. Jung, H. Lim, K. H. Kwak, and K. Kim, "A 500-MHz DDR high-performance 72-Mb 3-D SRAM fabricated with laser-induced epitaxial c-Si growth technology for a stand-alone and embedded memory application," *IEEE Trans. Electron Devices*, vol. 57, no. 2, pp. 474–481, Feb. 2010.

[11] P. Batude *et al.*, "3-D sequential integration: A key enabling technology for heterogeneous co-integration of new function with CMOS," *IEEE J. Emerg. Sel. Topic Circuits Syst.*, vol. 2, no. 4, pp. 714–722, Dec. 2012.

[12] J. Burns *et al.*, "Design, CAD and technology challenges for future processors: 3D perspectives," in *Proc. ACM Design Autom. Conf.*, San Diego, CA, USA, Jun. 2011, p. 212.

[13] S. Bobba *et al.*, "CELONCEL: Effective design technique for 3-D monolithic integration targeting high performance integrated circuits," in *Proc. Asia South Pac. Design Autom. Conf.*, Yokohama, Japan, Jan. 2011, pp. 336–343.

[14] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Design and CAD methodologies for low power gate-level monolithic 3D ICs," in *Proc. Int. Symp. Low Power Electron. Design*, Aug. 2014, pp. 171–176.

[15] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "High-density integration of functional modules using monolithic 3D-IC technology," in *Proc. Asia South Pac. Design Autom. Conf.*, Yokohama, Japan, Jan. 2013, pp. 681–686.

[16] K. Chang *et al.*, "Power benefit study of monolithic 3D IC at the 7nm technology node," in *Proc. Int. Symp. Low Power Electron. Design*, Rome, Italy, Jul. 2015, pp. 201–206.

[17] C. Ortolland *et al.*, "Laser-annealed junctions with advanced CMOS gate stacks for 32nm node: Perspectives on device performance and manufacturability," in *Proc. IEEE Int. Symp. VLSI Technol.*, Honolulu, HI, USA, Jun. 2008, pp. 186–187.

[18] B. Rajendran *et al.*, "Low thermal budget processing for sequential 3-D IC fabrication," *IEEE Trans. Electron Devices*, vol. 54, no. 4, pp. 707–714, Apr. 2007.

[19] *ASU Predictive Technology Models*. Accessed on Jan. 2016. [Online]. Available: http://ptm.asu.edu/

[20] W. Steinhögl, G. Schindler, G. Steinlesberger, and M. Engelhardt, "Size-dependent resistivity of metallic wires in the mesoscopic range," *Phys. Rev. B, Condens. Matter*, vol. 66, Aug. 2002, Art. no. 075414.

[21] S. R. Nassif, "Modeling and forecasting of manufacturing variations," in *Proc. Asia South Pac. Design Autom. Conf.*, Yokohama, Japan, 2001, pp. 145–149.

[22] G. Lopez, "The impact of interconnect process variations and size effects for gigascale integration," Ph.D. dissertation, School Elect. Comput. Eng., Georgia Inst. Technol., Atlanta, GA, USA, 2009.

[23] W. Steinhögl, G. Schindler, G. Steinlesberger, M. Traving, and M. Engelhardt, "Impact of line edge roughness on the resistivity of nanometer-scale interconnects," *Microelectron. Eng.*, vol. 76, nos. 1–4, pp. 126–130, Oct. 2004.

[24] G. Schindler *et al.*, "A morphology study of copper and aluminum interconnects," in *Proc. Adv. Metallization Conf.*, 2003, pp. 213–217.

[25] G. Schindler, M. Meyer, G. Steinlesberger, M. Engelhardt, and E. Zschech, "Grain orientation of copper nano interconnects," in *Proc. Adv. Metallization Conf.*, 2003, pp. 205–211.

[26] J. J. Plombon, E. Andideh, V. M. Dubin, and J. Maiz, "Influence of phonon, geometry, impurity, and grain size on copper line resistivity," *Appl. Phys. Lett.*, vol. 89, no. 11, 2006, Art. no. 113124.

[27] D. Choi *et al.*, "Electron mean free path of tungsten and the electrical resistivity of epitaxial (110) tungsten films," *Phys. Rev. B, Condens. Matter*, vol. 86, Jul. 2012, Art. no. 045432.

[28] W. Steinhogl *et al.*, "Tungsten interconnects in the nano-scale regime," *Microelectron. Eng.*, vol. 82, nos. 3–4, pp. 266–272, 2005.

[29] *International Technology Roadmap for Semiconductors*. Accessed on Jan. 2016. [Online]. Available: http://www.itrs.net/

[30] J. C. Cabral, Jr., *et al.*, "Metallization opportunities and challenges for future back-end-of-the-line technology," in *Proc. Adv. Metallization Conf.*, Oct. 2010, pp. 136–137.

[31] *Nangate Open Cell Library*. Accessed on Jan. 2016. [Online]. Available: http://www.nangate.com/

[32] R. Aitken *et al.*, "Physical design and finfets," in *Proc. Int. Symp. Phys. Design (ISPD)*, Petaluma, CA, USA, 2014, pp. 65–68.

[33] S. Sinha, B. Cline, G. Yeric, V. Chandra, and Y. Cao, "Design benchmarking to 7nm with FinFET predictive technology models," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, Redondo Beach, CA, USA, 2012, pp. 15–20.

[34] J. Knechtel, I. L. Markov, and J. Lienig, "Assembling 2-D blocks into 3-D chips," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 2, pp. 228–241, Feb. 2012.

[35] X. He, S. Dong, Y. Ma, and X. Hong, "Simultaneous buffer and interlayer via planning for 3D floorplanning," in *Proc. Int. Symp. Qual. Electron. Design*, San Jose, CA, USA, Mar. 2009, pp. 740–745.

[36] M.-C. Tsai, T.-C. Wang, and T. T. Hwang, "Through-silicon via planning in 3-D floorplanning," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 8, pp. 1448–1457, Aug. 2011.

[37] X. Tang, R. Tian, and M. D. R. Wong, "Optimal redistribution of white space for wire length minimization," in *Proc. Asia South Pac. Design Autom. Conf.*, vol. 1. Shanghai, China, Jan. 2005, pp. 412–417.

[38] D. H. Kim, R. O. Topaloglu, and S. K. Lim, "Block-level 3D IC design with through-silicon-via planning," in *Proc. Asia South Pac. Design Autom. Conf.*, Sydney, NSW, Australia, Jan. 2012, pp. 335–340.

[39] S. Garg and D. Marculescu, "3D-GCP: An analytical model for the impact of process variations on the critical path delay distribution of 3D ICs," in *Proc. Int. Symp. Qual. Electron. Design*, San Jose, CA, USA, Mar. 2009, pp. 147–155.

[40] H. Hong, J. Lim, and S. Kang, "Process variation-aware floorplanning for 3D many-core processors," in *Proc. IEEE Elect. Design Adv. Packag. Syst. Symp.*, Taipei, Taiwan, Dec. 2012, pp. 193–196.

[41] W.-L. Hung, G. M. Link, Y. Xie, N. Vijaykrishnan, and M. J. Irwin, "Interconnect and thermal-aware floorplanning for 3D microprocessors," in *Proc. Int. Symp. Qual. Electron. Design*, San Jose, CA, USA, Mar. 2006, pp. 98–104.

Shreepad Panth (S'11–M'15) received the B.S. degree from Anna University, Chennai, India, in 2009, and the M.S. and Ph.D. degrees from the Georgia Institute of Technology, Atlanta, GA, USA, in 2011 and 2015, respectively.

He is currently an SoC Design Engineer with Intel Corporation, Santa Clara, CA, USA. His current research interest includes physical design for current and next generation 3-D ICs. He has authored over 20 publications.

Dr. Panth was a recipient of the Best Paper at ATS'12 and IITC'14, and nominations for Best Paper at ISPD'14 and DAC'14.

Sandeep Kumar Samal (S'12) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology Kharagpur, Kharagpur, India, in 2012, and the M.S. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2013, where he is currently pursuing the Ph.D. degree with the School of Electrical and Computer Engineering.

His current research interests include low power and reliable digital design, and modeling and analysis using TSV-based and monolithic 3-D technology.

Kambiz Samadi (S'04–M'12) received the M.Sc. and Ph.D. degrees from the University of California at San Diego, San Diego, CA, USA, in 2007 and 2010, respectively.

He joined Qualcomm Research, San Diego, CA, USA, in 2011, where he is currently a Staff Research Engineer. He has over 25 publications in refereed journals and conferences. His current research interests include on-chip interconnection modeling and optimization for system-level design, 3-D-IC modeling and optimization, 3-D-IC EDA solutions, and 3-D-IC architecture-level design space explorations.

Yang Du (M'96) received the Ph.D. degree from Columbia University, New York, NY, USA, in 1994.

He is currently a Director of Engineering with Qualcomm Research, San Diego, CA, USA, where he leads a team in advanced nano-technology and semiconductor research. He has authored over 50 patents and numerous conference/journal papers. His current research interests include emerging semiconductor devices, predictive modeling, novel very large scale integration (VLSI) circuits and architecture, 3-D-IC technology and design, emerging 3-D-VLSI circuit, architecture and system integration, design automation, and thermal aware design methodologies.

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

He joined the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2001, where he is currently the Dan Fielder Professor of Electrical and Computer Engineering. His research on 3-D IC reliability is featured as Research Highlight in the Communication of the ACM in 2014. He is the author of the book entitled *Practical Problems in VLSI Physical Design Automation* (Springer, 2008). His current research interests include architecture, circuit design, and physical design automation for 3-D ICs.