Adaptive Regression-Based Thermal Modeling and Optimization for Monolithic 3-D ICs

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

***Abstract*—In this paper, we first present a comprehensive study of the unique thermal behavior in monolithic 3-D integrated circuits (ICs) in contrast to through silicon via-based 3-D ICs. In particular, we study the impact of the thin interlayer dielectric between the device tiers on vertical thermal coupling. We then study and compare the impact of different application-based package structures on the thermal behavior of monolithic 3-D ICs. With these unique properties and behavior, we develop a fast and accurate compact full-chip thermal analysis model based on nonlinear regression technique which adapts to the package structure during development and hence considers it during temperature evaluation. Our model is extremely fast and highly accurate with an error of less than 5%. This model is incorporated into a thermal-aware 3-D-floorplanner that runs without significant runtime overhead. We use the floorplanner with our package-aware thermal model and observe up to 22% reduction in the maximum temperature with insignificant area and performance overhead.**

***Index Terms*—3-D floorplanning, 3-D integrated circuit (IC), mobile package, monolithic, monolithic intertier via (MIV), thermal.**

I. Introduction

THE ADVENT of 3-D integrated circuit (IC) technology has opened up the potential of highly improved circuit designs. Through silicon vias (TSVs) enable the vertical integration of separate dies to form a single 3-D chip. However, TSVs consume a lot of area and have large capacitance. This puts a restriction on the number of TSVs and the type of circuits that can be used. Therefore, the greater benefits of 3-D IC are masked by these negative characteristics of TSVs.

Recently developed monolithic 3-D integration technology enables sequential integration of device layers in contrast to bonding of fabricated dies [1]. Monolithic 3-D integration uses nanoscale monolithic inter-tier vias (MIVs) to connect the vertical device layers. MIVs are similar to regular metal-layer vias and their capacitance and area values are negligible compared to those of TSVs that are microscale. This allows the use of many such MIVs for vertical connections which enables significantly higher integration density than that of TSV-based 3-D ICs.

The major contributions of this paper are as follows.

1) For the first time, we study and explain the thermal characteristics of monolithic 3-D ICs with comparison to TSV-based 3-D ICs. We highlight the unique properties of vertical thermal coupling and absence of lateral thermal conduction in the device layers (Section III).
2) We study the new mobile package and cooling solution, its characteristic differences from conventional heat sink cooling solution and its thermal impact on ICs (Section IV).
3) We identify the important factors affecting temperature and develop a very fast and accurate nonlinear regression-based temperature evaluation model which also incorporates package-awareness during design. This is the first such thermal modeling technique for monolithic 3-D ICs for both conventional heat-sink based and modern mobile packages (Section V).
4) We use our model to carry out thermal-aware 3-D floorplanning and show significant reduction in maximum temperature with minimal or no area and performance overhead for designs with both conventional packages with heat sink and mobile packages (Section VI).
5) We study the impact of thickness and conductivity of various materials used in mobile package to help package optimization in 3-D ICs (Section VII).

We discuss the motivation and background in Section II and conclude this paper in Section VIII.

II. Motivation and Background

Monolithic 3-D ICs can overcome the shortcomings of TSV-based 3-D ICs; however, one major concern with 3-D ICs in general is the increase in power density (PD) which leads to high temperature values and thermal issues [2], [3]. Even if we achieve power reduction by going 3-D, the increased PD affects the temperature, especially in the layers away from the heat sink or other equivalent cooling features in modern miniaturized electronics. Therefore, importance of thermal-aware design methodologies become more critical in 3-D ICs. The major bottleneck of considering thermal aspect within

Manuscript received February 25, 2015; revised August 6, 2015; accepted October 23, 2015. Date of publication February 3, 2016; date of current version September 7, 2016. This work was supported by Qualcomm Research. This paper was recommended by Associate Editor L. Benini.

S. K. Samal, S. Panth, and S. K. Lim are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: sandeep.samal@gatech.edu; shreepad.panth@gatech.edu; limsk@ece.gatech.edu).

K. Samadi, M. Saeidi, and Y. Du are with Qualcomm Technologies Incorporation, San Diego, CA 92121 USA (e-mail: ksamadi@qti.qualcomm.com; msaeidi@qti.qualcomm.com; ydu@qti.qualcomm.com).

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.2016.2523983

the physical design process is the huge runtime required for accurate temperature analysis. The inclusion of such detailed analysis within the design process is not practically feasible.

There exists several works which focus on the thermal issues and thermal aware design of TSV-based 3-D ICs [4], [5]. Attempts have been made to develop accurate temperature evaluation models to be included within the chip design process [6]. The use of compact resistive thermal grid network to estimate the temperature profile of a chip has been studied by Cong *et al.* [5]. They use compact resistive model and hybrid model within the floorplanning process to analyze the temperature and insert whitespace for dummy vias. The calculation of resistive network solving still consumes some runtime and the insertion of whitespace increase the area further, diminishing the 3-D IC benefits. They report 56% reduction in temperature but with a large area increase in 21%. The optimization of silicon area is important in 3-D ICs along with the temperature rise and we cannot sacrifice too much area for temperature improvement. Zhou *et al.* [7] proposed a force-directed floorplanner approach to spread high power blocks while simultaneously optimizing wirelength, area, and thermal distribution. The modeling of temperature based on total leakage power dissipation and its use in the tier-planning of similar layout processor chips is demonstrated by Juan *et al.* [8]. The 3-D overlap estimation along with PD calculations for thermal-aware planning has been used in [9]. All these methods are either targeted for TSV-based 3-D IC design or incur extra runtime or use indirect methods of thermal analysis. They also focus only on the conventional package and stack up which has a heat sink at the top.

In order to justify the overall advantages of monolithic 3-D IC over 2-D ICs and over TSV-based 3-D ICs, their thermal-aware design is necessary. Interestingly, monolithic 3-D ICs exhibit different thermal behavior due to their layer structure and are not as thermally bad as TSV-based 3-D ICs even though copper TSVs increase conductivity. These properties allow us to build a very fast temperature model with high degree of accuracy. In addition to that, 3-D ICs also provide huge potential in the design of low power processors for use in mobile applications and monolithic 3-D ICs specifically enable ultrahigh packing density [10]. However, mobile applications have different package structure due to their size and weight constraints. Heat sink is absent in such packages and different materials are used for spreading and dissipation of heat. Therefore, to tap all benefits of monolithic 3-D IC to the full extent, it is very important to also take into account the different types of package structures which will significantly impact the overall thermal quality.

III. New Issues and Unique Thermal Properties of Monolithic 3-D ICs

A. Monolithic 3-D Integration

Monolithic 3-D integration technology enables ultrahigh density vertical integration. The advanced manufacturing technology allows active device layers as thin as 10 nm to be integrated over one another with high alignment precision [1]. To understand the thermal properties of monolithic 3-D ICs, we first need to look into the details of its structure. A typical two-tier monolithic stackup is shown in Fig. 1(a) in a flip-chip configuration. The first set of transistors closer to the handle bulk are processed with standard SOI process and make up Tier 1. A thin interlayer dielectric (ILD) is deposited over the metal layers for the bonding of the next device layer. This device layer along with the metal layers make up the other tier (Tier 0) of the 3-D stackup. The transistors in these layers are processed with low temperature process (<650 °C). Batude *et al.* [11] have demonstrated that performance of devices processed at low temperature can match performance of regular high temperature process devices. For our work and subsequent sections, we consider similar performance of devices in all tiers of monolithic 3-D IC. We also follow the tier numbering convention in 3-D ICs such that Tier0 (bottom tier) is the one closest to the printed circuit board (PCB) and the numbers increase as we go further away.

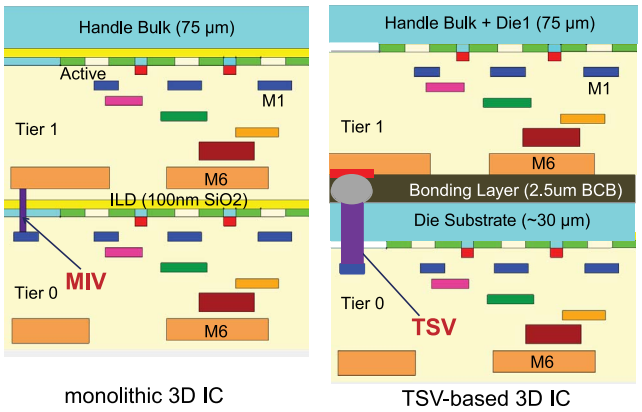

Fig. 1. Two-tier 3-D IC layer structure (heat sink on top) of (a) monolithic 3-D IC versus (b) TSV-based 3-D IC.

B. Material and Structural Differences

The differences in fabrication process of monolithic 3-D and TSV-based 3-D result in significant differences in their thermal behavior. Fig. 1 highlights differences in the materials used in the two technologies and their conductivity and thickness influences the thermal behavior. Table I lists their details for a typical 45 nm technology process. The relative contribution of each material per tier is also shown in the table.

In TSV-based 3-D ICs, copper TSVs and μ-bumps improve the conductivity. However, the presence of bonding layer (underfill) which is necessary for stress-related issues worsens the overall conductivity significantly [Fig. 1(b)]. Typical materials used for underfill are required to be soft and elastic and in general such materials have poor thermal conductivity. BCB is one of the commonly used materials and it has a thermal conductivity of 0.29 W/m-K. Copper metal on the other hand has a thermal conductivity of 401 W/m-K. The presence of this underfill which is around 2.5 μm thick impedes the heat flow from Tier0 toward the heat sink present above the handle bulk resulting in considerable temperature rise in Tier0. However, heat from Tier0 passes through silicon substrate before reaching the underfill wall. Silicon being a good

TABLE I
DIFFERENT MATERIALS USED, THEIR THERMAL CONDUCTIVITIES, VERTICAL THICKNESSES, AND RELATIVE % IN TOTAL STACK

Layer/Structure	Material	Thermal Cond. (W/m-K)	Vertical Thickness	% of total Tier0	% of total Tier1
Monolithic					
Handle Bulk	Silicon	141	$75\mu m$	-	97.1
ILD (Inter-tier)	SiO_2	1.38	100nm	4.3	0.13
BEOL	SiO_2/Cu	1.38/401	$2.2\mu m$	93.6	2.84
TSV-based					
Handle Bulk+Die 1	Silicon	141	$75\mu m$	-	97.2
Die0 Substrate	Silicon	141	$30\mu m$	86.5	-
Bonding Layer	BCB	0.29	2.5 μm	7.2	-
TSV	Copper	401	$30\mu m$	in Die0 sub	
TSV-bump	Solder	50	2.5 μm	in BCB	
BEOL	SiO_2/Cu	1.38/401	$2.2\mu m$	6.3	2.9

conductor of heat spreads out the thermal profile of Tier0 by allowing many lateral heat flow paths in its 30 μm thickness. Tier1 in TSV-based 3-D ICs does not have any buried oxide between the device layer and the handle bulk. This helps in better conduction of heat from Tier1 to the heat sink.

In contrast to TSV-based 3-D ICs, the bonding layer and bulk substrate are absent in monolithic 3-D ICs while the different tiers are separated by ILD which function as the buried oxide for the SOI process for formation of subsequent device layers. Also the MIVs are tiny compared to the huge TSVs. These particular differences change the heat dissipation phenomenon of monolithic 3-D ICs from that of TSV-based 3-D ICs. The absence of bulk substrate and the extremely thin device layers reduce the lateral conductivity to almost zero which results in heavy tier-to-tier thermal coupling. The heat flows only vertically up until it reaches the handle bulk where there is lateral spreading due to its very large thickness compared to all other layers. The presence of buried oxide also increases the thermal resistance from top tier to handle bulk. All these factors considered together result in similar temperature profiles for all the tiers irrespective of the whitespace locations in the different tiers. A high power block in the tier closer to the heat sink will also result in a hot spot in all other tiers away from the heat sink. There is a difference in the temperature value of the same 2-D location in two tiers due the rise across the 100 nm ILD. Also the maximum temperature of the tier closest to the heat sink is more than that of TSV-based 3-D IC due to the presence of additional oxide layer which is a poor conductor.

C. Vertical Tier-to-Tier Coupling in Monolithic 3-D ICs

Fig. 2(a) shows the layouts of a three-tier monolithic block level 256 bit multiplier. Fig. 2(b) are the temperature maps of the individual tiers with 2-D thermal analysis performed on each tier independently for conventional package with heat sink. We can see the cooler regions (blue) and their spread exactly following the whitespace locations in the corresponding layouts of Fig. 2(a). Since it is 2-D thermal analysis, these whitespace locations have no heat generation and hence the cooler spots. However, when we carry out the 3-D thermal analysis for this three-tier 3-D design considered as a whole, the temperature maps change significantly [Fig. 2(c)]. The hotspots of all the individual tiers overlap with each other in

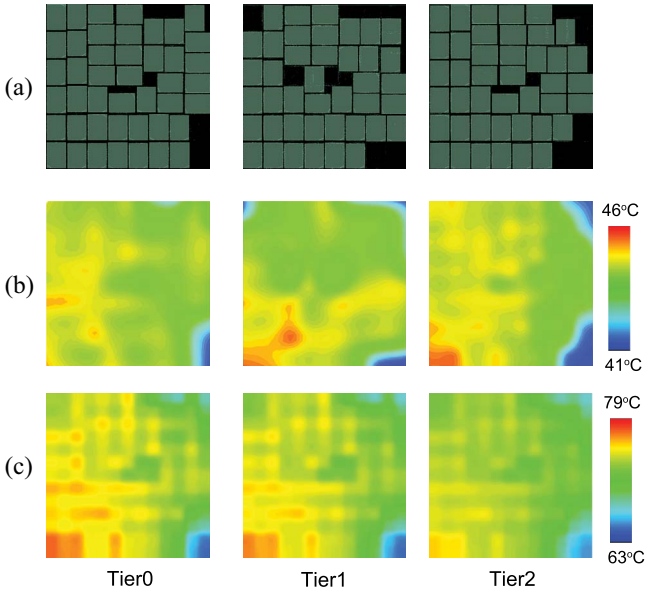

Fig. 2. Tier-to-tier coupling in monolithic 3-D ICs. (a) 3-D-Floorplan for three-tier 256-bit multiplier. (b) Temperature maps with (b) independent 2-D thermal analysis of each tier and (c) stacked 3-D thermal analysis.

3-D and affect the temperature of all the three tiers with bottom tier suffering the most due to additional poor conducting ILDs on the way to the heat sink. The hotspots (red) in Tier2 affect the 3-D design most because it obstructs direct vertical heat flow from the tiers lying below along with addition of its own heat. As we know, there is negligible lateral conduction until the heat reaches the handle bulk. Therefore, the temperature maps are similar in trend of variation across the entire area. The temperature values also increase compared to the individual 2-D analysis of each tier due to almost threefold increase in PD. Only the common whitespace regions (bottom right corner) remain cooler in all the tiers.

D. Temperature Map Comparisons

Fig. 3 shows the temperature map of a same two-tier 3-D layout in monolithic technology and TSV-based technology. We compare and contrast these temperature maps for the two technologies along the lines of the discussions presented earlier and highlight the unique properties in monolithic 3-D ICs.

The layout is originally designed for a two-tier TSV-based 256 bit mutiplier. The TSV locations are shown in yellow in Tier0 layout and their landing pads shown in Tier1. Since our primary objective here is to understand the thermal behavior of the technology, for fair comparison from the thermal point of view, the same 3-D layout with same PD and performance is analyzed for a monolithic structure with TSVs replaced by MIVs at the same 3-D via locations. In practice, MIVs are much smaller and their design will consume much lesser area.

For the TSV-based 3-D IC temperature map, we can clearly see that the presence of TSVs help in improving the conduction significantly in Tier0. There are cooler spots among very hot ones wherever TSVs are present. The temperature of other regions is quite high due to the heat flow obstruction by the

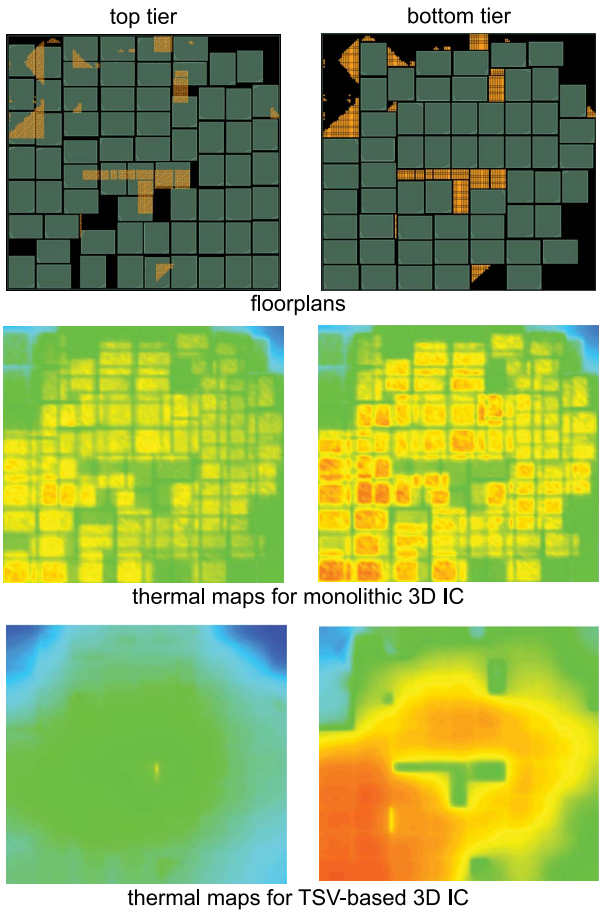

Fig. 3. Temperature maps of same two-tier 3-D floorplan (originally designed for TSV-based 3-D IC) in monolithic 3-D IC technology and TSV-based 3-D technology. The temperature range is [61 °C, 71 °C].

bonding layer. Tier1 is much cooler compared to Tier0 as it is closer to the heat sink. The other important thing to observe is the lateral spreading of temperature across the two tiers which smears the temperature profile of each tier. This is because of the bulk silicon substrate which allows multiple lateral heat flow paths.

For monolithic 3-D IC design on the other hand, the temperature profiles of the two tiers are identical and the block layouts can be demarcated in the temperature map itself. This is a result of absence of lateral conduction at the source of power dissipation. The vertical tier-to-tier coupling can be observed by the block outlines from both tiers appearing overlapped in the temperature maps. Tier0 map is hotter than Tier1 due to the heat block by the ILD. Tier1 of TSV-based 3-D IC is cooler than Tier1 of monolithic because of the absence of oxide which is a poor thermal conductor. Tier0 in TSV-based 3-D is much hotter than Tier0 in monolithic 3-D due to bonding layer which is a poorer conductor than SiO_2. Wei *et al.* [12] also compared TSV-based 3-D IC with monolithic 3-D ICs but did not consider the underfill layer. The mass production of TSV-based 3-D ICs without any underfill is highly unlikely due to stress-related issues. Therefore, we need to consider them during thermal behavior study of TSV-based 3-D ICs and then compare with monolithic 3-D ICs. This very poor

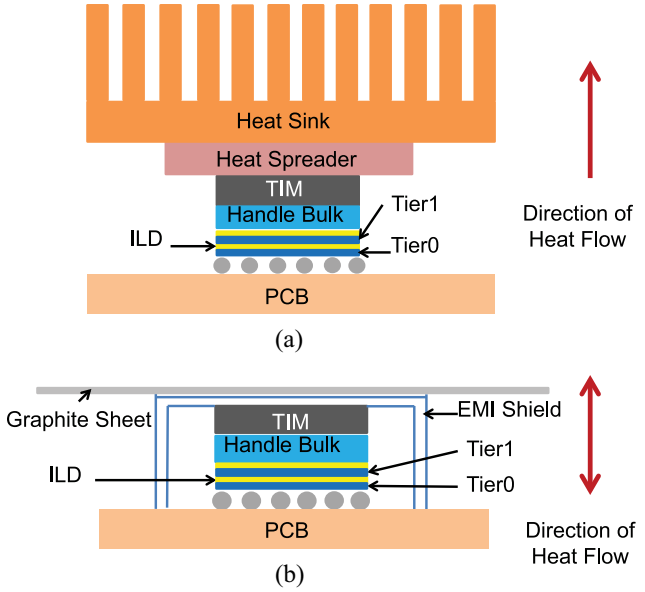

Fig. 4. 3-D IC Packaging structure for cooling. (a) Conventional cooling (with heat sink). (b) Modern mobile cooling (no heat sink).

conducting bonding layer in TSV-based 3-D ICs significantly worsen the temperature of tiers away from heat sink. If we do not consider this layer, then TSV-based 3-D ICs will be better than monolithic 3-D ICs thermally.

The key points from the above thermal study of monolithic 3-D ICs and comparison with TSV-based 3-D ICs are as follows.

1) Monolithic 3-D ICs have almost zero lateral conduction at the source of power due to very thin layers and show no lateral spreading in the device layers.
2) There is heavy vertical tier-to-tier coupling in monolithic 3-D and all tiers have similar temperature profile with differing absolute values due to rise across ILDs.
3) In monolithic 3-D ICs, handle bulk is the first layer in the path of heat flow where noticeable lateral conduction occurs. Therefore, the individual neighbors in a floorplan have an indirect effect unlike TSV-based 3-D ICs where they directly affect each other by conduction through silicon substrate.
4) MIVs do not play an important role in heat conduction such as TSVs due to small size and thickness.

IV. New Mobile Package Structure and Properties

Miniaturization is one of the key characteristics of modern very-large-scale integration (VLSI). With low power devices such as smart phones, smart watches, and sensor nodes, there is a need for compactness and light weight materials and high integration density. The power dissipation in such applications is much lower than that of high-performance servers and desktop computers. Large and heavy heat sinks with cooling fans can be avoided for such systems. Therefore, industry uses a different kind of packaging structure for the ICs used in mobile applications. Fig. 4(b) shows the structure and materials used for packaging and cooling of mobile processors [13].

TABLE II
PROPERTIES OF THE DIFFERENT LAYERS IN MOBILE PACKAGE STRUCTURE

Layer	Vertical thickness (μm)	Therm. cond. (W/mK)	
		Vertical	Lateral
PCB	800-1500	1.5-4.5	25-60
Handle Bulk	50-200	141	141
Therm. Int. Material (TIM)	500-1200	0.5-5	0.5-5
EMI Shield (Steel/Al)	100-250	20/120	20-120
Graphite Sheet	25	2.9-4.5	300-500

Since monolithic 3-D ICs enable very high integration density, they are a very good candidate for use in mobile processors to increase functions in the same form factor. Such mobile systems use the new mobile package structure and there is a need for good thermal planning and budgeting for the use of monolithic 3-D ICs. This is the key motivation to study mobile package structure in detail, analyze the properties and impact of the new materials used, and develop thermal model which can incorporate the package characteristics during fast accurate temperature evaluation. Furthermore, knowing the impact of various materials used in the package will enable designers and packaging engineers to carry out package optimization after thermal optimization during physical design. In this section, we first discuss the package structure used in mobile phones. We then discuss the differences in the cooling phenomenon for such packages in contrast to conventional packages with heat sink. We also discuss the thermal behavior with various number of tiers in 3-D IC design.

A. Structure and Materials

Fig. 4 shows the package structure for conventional cooling with heat sink [Fig. 4(a)] and mobile applications without any heat sink [Fig. 4(b)]. The major differences in the mobile package is the absence of a copper heat sink with multiple fins and copper heat spreader.

The absence of heat sink with multiple fins and cooling fan in mobile package structure reduces the dominance of the upward path in heat conduction. The red arrows in Fig. 4 show the primary direction of heat flow in the respective structures. Because of a very large heat sink and low thermal resistive path, almost all of the heat flows toward the heat sink in conventional packages. However, for mobile packages, heat flows in both directions and therefore, the importance of all other layers increases. The different layers in the mobile package structure and their thickness and conductivity values are shown in Table II. Note that the values are shown in ranges as the properties of some of the layers can be different in different systems based on the actual composition and requirement.

Along with the PCB inside the mobile phone toward the display side, the back body also helps in heat dissipation and the very thin graphite sheet helps in spreading the heat to the entire back cover instead of having concentrated hot spots. The electromagnetic insulator (EMI) which is usually a steel or aluminum sheet also helps in spreading of the heat by providing a low heat resistive path along with its primary function of shielding. Another important factor is that the lateral conductivity of PCB and graphite sheet are much better than their vertical conductivities. Therefore, they play a significant role

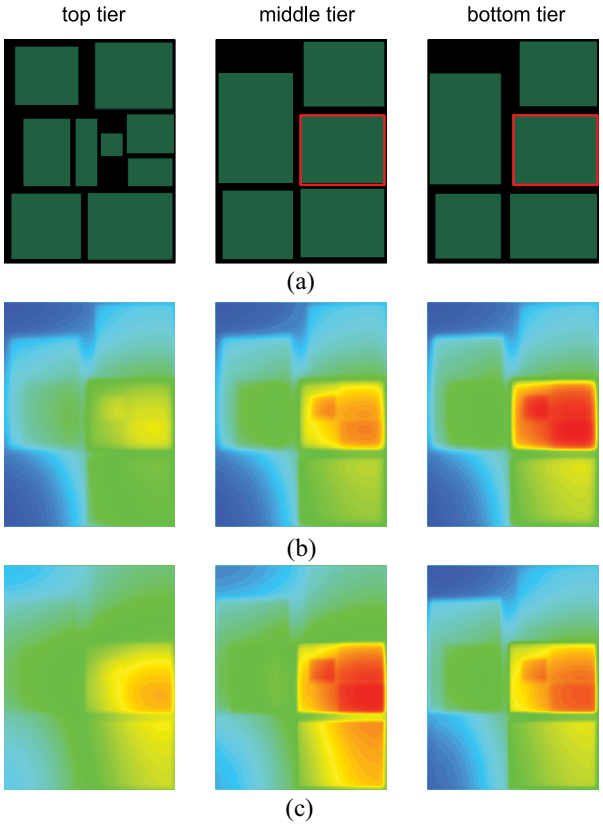

Fig. 5. Temperature hotspot and distribution comparison for conventional cooling (with heat sink) and modern mobile cooling (no heat sink) for openSPARC T2 core. Temperature scales are normalized with blue as minimum and red as maximum temperature for respective package structure with lower PD for mobile systems. Red outline blocks in (a) are the maximum PD blocks. (a) Floorplans. (b) Temperature map with conventional heatsink package. (c) Temperature map with mobile package (no heat sink).

in good lateral spreading and hence increasing the surface area of contact with the external environment. The same design with same power maps will have hotter spots with mobile package than a package with heat sink. However, mobile processors have significantly reduced average PD (< 2 W/cm^2) compared to high performance servers (20–30 W/cm^2) and hence the maximum temperature is well within control even with the absence of heat sink and fan in a different package. In our thermal analysis related to mobile packages, we assume that the PCB, the graphite sheet, and the free regions of EMI are all connected to ambient environment.

B. Comparison With Conventional Package Structure

The absence of heat sink significantly changes the thermal behavior of mobile packages in contrast to that of conventional packages with heat sink. This difference becomes more prominent in multitier 3-D ICs where there are multiple layers of heat source. Fig. 5(a) shows the block level layout of three-tier OpenSPARC T2 core [14] with the highest PD execution unit blocks highlighted in red outline. The floorplan is targeted toward minimum wirelength. Fig. 5(b) and (c) are the temperature maps with the conventional and mobile package structures, respectively. It is important to note that the power dissipation of the system under the two packaging structures

TABLE III
MAXIMUM TEMPERATURE RISE VALUES (ABOVE ROOM) ACROSS DIFFERENT TIERS IN DESIGNS WITH DIFFERENT PACKAGES

Design	Conventional Package			Mobile Package		
	tier0	tier1	tier2	tier0	tier1	tier2
2D IC	24.37	-	-	16.54	-	-
2-tier 3D IC	47.10	44.30	-	32.27	31.60	-
3-tier 3D IC	65.82	63.10	59.95	43.18	44.15	42.4

is different with mobile package having much lower PD. The temperature ranges are normalized with blue color for minimum and deep red for maximum temperature in the respective packages.

As discussed in previous section, almost all of the heat flows toward the heat sink in conventional packages. This makes the tier away from the heat sink most critical in terms of thermal reliability. This is evident from the red hot spots in the bottom tier in Fig. 5(b) which is farthest from heat sink. On the other hand, bi-directional heat-flow in mobile packages has two important consequences. First is that the middle tier is most critical in terms of thermal reliability unlike conventional packages [Fig. 5(c)]. Second, due to bi-directional flow, the relative temperature difference between tiers is lesser compared to conventional package structure. This difference can be seen in the temperature maps of Fig. 5, where the relative difference in maximum temperature across tiers is lesser for mobile packages and more for conventional packages.

C. Thermal Behavior With Different Number of Tiers

Since heat flow is bi-directional in mobile packages, the middle tiers are more critical for multitier 3-D ICs and the extreme tiers are influenced similarly. This implies that two-tier 3-D ICs are almost similar to 2-D ICs in terms of thermal floorplanning for same mobile package properties unless the power map is heavily unbalanced to have excessive power dissipation on one tier only. A high PD block can be placed in either of the two tiers in two-tier 3-D IC design to have the same overall temperature profile because heat flows in both directions almost equally. This is not the case for conventional packages with heat sink because the tier away from the heat sink is always more critical thermally and it is desirable to have the high PD blocks closer to the heat sink. The maximum temperature of different tiers for a 2-D, two-tier 3-D, and three-tier 3-D design with same total power is shown in Table III. The maximum temperature for the two tiers in two-tier 3-D IC is almost same for both tiers with mobile package but for three-tier case, the middle tier has higher maximum temperature than both the extreme tiers. The increase is uniform for conventional package with the tier away from the heat sink having worst temperature in all 3-D designs.

V. FAST THERMAL ANALYSIS MODEL

A. Model Development

Steady-state finite element thermal analysis will lead to large matrix calculations of an equivalent thermal resistive network with multiple power sources. We use nonlinear regression to accurately model the steady-state temperature of monolithic 3-D ICs after generating a large number of representative samples. The method of approximating a quantity dependent on certain number of predictor inputs using such techniques has been used in earlier studies [15]. While regression helps in determining direct correlation between target and inputs, nonlinearity helps is reducing the total number of required inputs without affecting prediction accuracy. We set temperature as our target quantity and model it after successfully determining the different parameters of the monolithic 3-D IC on which it depends. Our developed temperature model evaluates the steady-state thermal behavior of monolithic 3-D ICs of given dimensions, number of tiers, and power distribution.

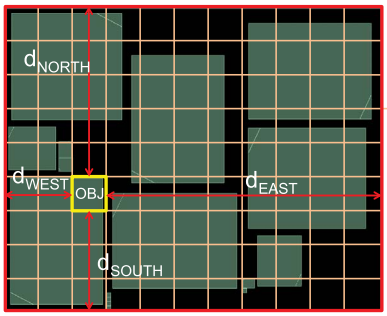

Fig. 6. Final model structure with an objective tile. The red rectangles show the objective tile and rest of the chip. Their power values along with 2-D distances from boundary are used as inputs for temperature calculation.

1) Initial Experiments: We divided the entire chip into a tile based structure for each tier (Fig. 6). Each of the tiles is randomly assigned a power value such that the PD lies between 0 and 100 W/cm^2. Full chip thermal FEA with 20 μm $\times$ 20 μm mesh is carried out on these test cases. To address different types of applications, we consider two kinds of packaging structure independently. Fig. 4(a) is the conventional cooling method which uses heat spreader and heat sink. Almost 100% of the heat dissipates through the heat sink. Fig. 4(b) is the packaging structure used in modern smart phones due to size limitations [13]. The thermal resistance in both directions is of the same order and therefore there is bidirectional heat dissipation. We use ANSYS Fluent and SPICE simulations to carry out the thermal analysis [16].

Based on our study of the thermal properties discussed earlier, we conducted various experiments involving different power distributions, different granularity of tile division, multiple neighboring levels considered separately versus considered as single entity lumped together, and temperature dependence on 2-D and 3-D location of the objective tile. The neighboring tiles of an objective tile were found to have a unified effect on the temperature of the objective. The reason is that they affect the objective indirectly through the handle bulk and not directly because immediate lateral conduction is almost absent. We also observed that the location of a particular tile in the layout affects its temperature value. Some of the experiments conducted are explained in the following.

The primary goal is to divide the entire chip into tiles of 100 μm $\times$ 100 μm and then obtain a model with minimum number of inputs to calculate the temperature of each

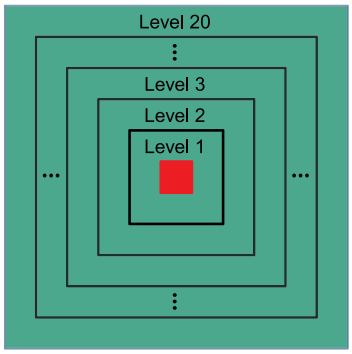

Fig. 7. Experimental setup with 20 neighboring levels and objective at the center.

TABLE IV
EXPERIMENTAL RESULTS WITH DIFFERENT NUMBER OF NEIGHBORS CONSIDERED DURING MARS MODELING

No. of levels considered	GCV	Avg Error (%)	RMSE	Most important variable
20	0.108	1.31	0.46	Power_ Level20
19	3.855	8.04	2.66	Power_ Level19
18	6.550	11.26	3.90	Power_ Level18
17	7.475	13.66	4.73	Power_ Level17

tile accurately without carrying out full FEA simulations. The lesser the number of input variables required, faster is the full-chip temperature analysis. To correctly determine the number neighboring levels to be covered, we carried out experiments in starting with 20 levels of neighboring tile levels and dropping the farthest neighbor one at a time to see if it affects the results (Fig. 7). Since each tile is 100 μm $\times$ 100 μm and there are 20 levels of neighbor rings, the chip size is 4.1 mm $\times$ 4.1 mm. The effectiveness of a model is measured in terms of the generalized cross validation (GCV) values of the model development and average error. A GCV value close to zero implies perfect modeling. The results of this experiment show that the amount of error increases as we remove the farthest neighbors (Table IV). This shows that the farthest neighbors do have a considerable effect on the temperature of the objective even though they are far away laterally. This is because the total power of the larger rings of tiles are much more than the objective and as we have already pointed out, all of this heat primarily goes vertically to the handle bulk layer therefore indirectly affecting the objective tile temperature. Therefore, we cannot ignore any power dissipation irrespective of the lateral distance from the objective tile.

Using the same raw data as mentioned earlier, we carry out analysis from a different viewpoint. The entire region (20 levels) is divided into different number of equal regions, namely 20 partitions (default), 10 partitions, 5 partitions, 4 partitions, 2 partitions, and finally a single partitions, where all 20 neighboring tile rings are treated as one. We then use these different partitions' power as variables to develop the model and compare the resulting model in terms of GCV and average absolute error (Table V). The results show that it is not necessary to have fine grained neighbors in the model. All the neighbors near or far have similar effect. Once again, this is explained by the indirect effect of neighbors through the handle bulk which

TABLE V
EXPERIMENTAL RESULTS OF MODELING WITH THE ENTIRE CHIP AREA CONSIDERED COMPLETELY BUT WITH DIFFERENT NUMBER OF LEVEL PARTITIONING

No. of partitions	GCV	Avg Error	Most important variable
20	0.108	1.31	Power_ Level20
10	0.105	1.53	Power_ Level10
5	0.199	1.77	Power_ Level5
4	0.20	1.95	Power_ Level4
2	0.626	2.14	Power_ Level2
1	0.727	2.32	Power_ Level1

is 75 μm thick and is silicon. The most important variable is always the last partition which has maximum magnitude of power.

2) Modeling Technique: From our experiments, we determine the following important parameters which influence the chip temperature: 1) power of objective tile; 2) total power of rest of the tiles in the same tier; 3) lumped sum of power of all tiles exactly above the objective; 4) lumped sum of power of rest of tiles of the above tiers (excluding the ones directly above); 5) lumped sum of power of all tiles exactly below the objective; 6) lumped sum of power of rest of tiles of the tiers below (excluding the ones directly below); 7) distance of the tile from each of the four 2-D boundaries (four variables); and 8) distance from vertical boundaries (3-D location). We can sum up the contributions of all power values other than that of the objective and immediate vertical neighbors because of the fact that all lateral influence is indirect due to lateral conduction at the handle bulk only which is above all the device layers [Fig. 1(a)]. The exponential increase in leakage with temperature can be taken care of by separating the power inputs into its components, namely dynamic and leakage powers and updating the leakage powers with temperature increase till a specified tolerance level is met.

Fig. 6 shows the division of chip and the 2-D-related variables. The target variable of the model is the rise in temperature above room temperature. Modeling is carried out with the help of multivariate adaptive regression splines (MARS) which is a nonlinear regression technique [17]. We minimize the number of inputs to keep the final temperature evaluation runtime less but with very good accuracy. The chip dimensions are implicitly taken care of by the distance variables and are excluded in the inputs. The tier number of the objective is also included to include the 3-D distance from the package boundaries. The individual tile size is fixed at 100 μm $\times$ 100 μm. Further granularity does not improve modeling results much but adds to the evaluation time for the whole chip which will affect the overall runtime of the thermal-aware floorplanner discussed later. We develop our thermal analysis models for each of the packaging structures separately for both two-tier and three-tier 3-D cases.

3) Sample Generation: To develop a good model, we require a large number of samples which cover all the possible variations in the parameters. To correctly capture all the possible 3-D chip sizes and power distributions, we carry out detailed thermal analysis of whole chip testcases which cover chip dimensions from 1 to 5 mm (in steps of 1 mm) with aspect ratio lying between 0.5 and 2. Each chip is divided

into 100 μm $\times$ 100 μm tiles and each such tile forms one sample. The above properties add up to 17 whole chip FEA simulations. These simulations are run only for one time to generate a large number of samples. PD values of the tiles are randomly distributed from 0 to 100 W/cm^2 while ensuring that around 10% of the total chip area is whitespace to correctly simulate practical designs. Around 15% of the samples are used for training of model and the rest used for testing. Since the samples were generated with the respective packages, the training captures the package properties into the final temperature model. For a different package structure or same package with different dimensions or material properties, the training samples generated with the corresponding package will capture the package properties. Therefore, the same modeling approach adapts the model to the package used for generating the training samples. During whole chip thermal simulation to obtain these samples, we treat back end of line (BEOL) material as 100% dielectric (SiO_2) material. This is because these generated samples do not have actual routing and dielectric constitutes maximum portion of BEOL [16].

We observed that the modeling is more accurate when all the samples have a random PD distribution with fixed average rather than with varying average. Therefore, we use samples with PD varying randomly from 0 to 100 W/cm^2 which results in an average PD of all samples close to 50 W/cm^2 (average of a random distribution). However, the PD of the practical case to be modeled will vary from design to design and needs to be taken care of during final evaluation. The trend prediction of our model is always correct irrespective of the actual average PD. However, the values are just shifted up or down and need a constant correction offset depending on the actual PD being greater than or less than 50 W/cm^2. From various practical example cases, we determine this offset as a multiple of the difference of the actual average PD of chip and the samples' PD (=50 W/cm^2 here). To successfully model samples covering different average PD, the number of total samples required increase by orders of magnitude. Since steady-state average temperature is a linear function of average power, our simple offset method avoids the need for generating more samples for modeling. The exact multiplying coefficient depends on the samples' used for modeling but will always remain constant once a model is developed irrespective of the actual chip being evaluated. The temperature evaluated by the model is the rise above room temperature as it is the more appropriate variable to model. To get the absolute temperature, we add it to the room temperature.

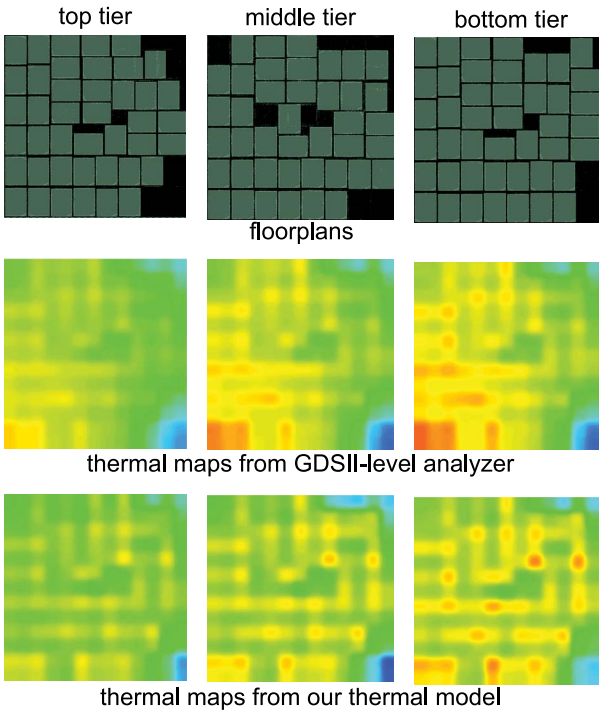

Fig. 8. Model accuracy: FEA simulation versus our temperature model for 256-bit multiplier. The temperature range is [63 °C, 79 °C].

B. Model Accuracy

The testing sample set gives an absolute average error of less than 1%. For practical designs, the average error is less than 5%. Fig. 8 shows the accuracy of model for a testcase, designed for three-tier 3-D. The top row shows the layouts of the individual tiers of three-tier 3-D IC, the middle row is the temperature maps after detailed FEA thermal analysis with the average temperature of each tile plotted, while the last row is the temperature analysis results from our model. We can clearly observe that our model captures the temperature variation trend very well and all the hotspots are accurately detected. This methodology of temperature estimation can be used for any circuit irrespective of whether it is a flat gate level design or a block level design. We just have to distribute the power into the tiles to carry out fast accurate temperature analysis.

The important conclusion is that irrespective of the error, the trend of temperature change within the chip is accurately evaluated with our model. The error primarily comes in the cooler regions of the chip. The reason is that there are immediate whitespaces on all sides (2-D and 3-D) around these cool tiles but we treat them similar to all other tiles. The tile power for such cases is low but the rest of the chip power used as temperature predictor becomes high and thus overestimates the temperature. This error can be easily rectified by adding one more step of checking the very low power tiles with low power 2-D and 3-D neighbors in a given design before feeding the predictors into the model. For these tiles, we can scale down the lumped power value of rest of the chip and then analyze the temperature.

C. Related Compact Modeling Work

There have been many studies on compact thermal modeling for both physical design as well as for model predictive controllers to have real-time thermal management in place of conservative worse-case thermal management for multicore chips. Compact thermal modeling for realistic energy-aware thermal management and control techniques with proper validation for multicore chips has been done in [18]. They develop robust thermal model using graybox approach which uses both statistical content and physical laws for better quality.

TABLE VI
FULL CHIP THERMAL ANALYSIS RUNTIME COMPARISON FOR THREE-TIER 3-D IC (1.3 mm × 1.3 mm FOOTPRINT). (RUNTIME FOR OUR MODEL AVERAGED OVER 10^6 RUNS)

Method	Runtime (in sec)	Normalized Runtime
Our Model	0.00022	**1.0**
GDS-level FEA	1082	**4.9 x 10^6**
HotSpot	5.68	**2.6 x 10^4**

Their adaptive models are used as controllers during operation and cover 2-D multicore designs.

Hotspot tool [6] is one of the most popular thermal analysis tools widely used in thermal studies. They used compact resistive models with different tuning parameters for tradeoff between run-time and accuracy. It has been maintained over the years with many updates to accommodate new features. Their grid model is capable of handling 3-D stacked chips with different power dissipating layers in the compact resistive mesh. Beneventi *et al.* [19] developed a compact thermal model for TSV-based logic+WideIO 3-D stack. Their model can successfully predict the temperature at locations where sensors are absent and can also evaluate the power dissipation based on temperature data. Cong *et al.* [5] also developed a fast but less accurate hybrid resistive model and another accurate but relatively slow resistive model for TSV-based 3-D ICs.

All these works on 3-D IC thermal modeling cover TSV-based 3-D ICs only and involve various forms of matrix manipulation, which though simplified is still computationally expensive. On the other hand, our thermal model is the first to cover monolithic 3-D ICs and is a very fast and simple model using very few input parameters. In the next section, we directly compare the runtime of our tool with Hotspot. Later, in Section VI-C, we also demonstrate the application benefits of our model compared to other thermal optimization tools.

D. Runtime Comparison

Since our model is a compact model with a simple mathematical relation obtained by regression, it is many orders of magnitude faster than full GDS-level analysis and compact resistive network analysis methods. This very important property helps us to use direct temperature estimation during a larger part of the design process. Table VI summarizes the runtime comparison with GDS-level FEA simulation and Hotspot [6]. The runtime is reported after the analysis of a three-tier 3-D design with 1.3 mm × 1.3 mm footprint. Hotspot is run for steady-state thermal analysis with 3-D stacking using a 16×16 grid network such that each grid's size is 81.25 μm × 81.25 μm, that is similar to the tile size used in our model. Our model is 4.9 × 10^6 times faster than FEA simulation and 2.6 × 10^4 faster than hotspot analysis for 3-D stacking.

VI. THERMAL-AWARE FLOORPLANNING

A. Floorplanning Algorithm

We use simulated annealing of sequence pair representation of floorplan to obtain the best floorplan depending on the weighted cost function specified. The nonthermal-aware floorplanner excludes the maximum temperature of chip from the cost function. Since this is a monolithic design, we are not concerned about the number of 3-D connections and hence do not include the number of MIVs in the cost function. It is known that larger area will help in reducing temperature. But area is directly proportional to cost, especially in miniaturized systems. Therefore, we tune our floorplanner to start optimizing temperature only after the specified area constraint is satisfied. Also, there is a tradeoff between maximum temperature reduction and total wirelength to have minimum performance overhead. More wirelength will increase total net switching power in the final design which may increase temperature further. However, if the blocks are not given freedom of movement within the constrained area, the solution space for temperature optimized floorplans within that area becomes smaller and there will not be significant temperature reduction. This freedom of movement of blocks implies wirelength overhead in the overall floorplan. Therefore, we use a step by step process to obtain the temperature optimized floorplan.

We first run the nonthermal floorplanner without any temperature cost and obtain the wirelength number. In the next step, given a certain slack on this wirelength, we include wirelength and maximum temperature in the initial cost function. Once, the wirelength goal is met, we minimize only temperature within that area and wirelength constraint. Any floorplan solution which violates the area and wirelength requirement is rejected. We also run the floorplanner with only 5% area slack to give more room for improvement. The final result obtained can be below this limit. The fact that our developed thermal model is extremely fast with good accuracy enables us to evaluate temperature profile of every sequence pair without any runtime issues and minimize the maximum temperature.

For a design with B blocks, N nets, and T thermal tiles, the complexity changes from $O(B \log B+N)$ to $O(B \log B+N+T)$ by including temperature evaluation [20]. The wirelength calculation for all nets for a given sequence pair is the major time complexity in the floorplanning process. Therefore, the addition of thermal analysis using our model which uses 100 μm × 100 μm tiles has insignificant overhead even with millions of moves during simulated annealing.

After obtaining the temperature optimized floorplan, we place and route the design using Encounter and then analyze the total power and timing in Synopsys PrimeTime to verify that we have no performance overhead. All benchmarks are designed to meet the specified timing requirement with minimal change in worst negative slack. A final full GDS-level thermal FEA is carried out with the specific package structure to check the maximum temperature.

B. Floorplanning Results for Conventional Package

We report two benchmark circuits for floorplaning comparison. The FFT benchmark is obtained at RTL level from Opencores and has 49 blocks of different sizes with 1400 interblock nets. The industry circuit benchmark was obtained

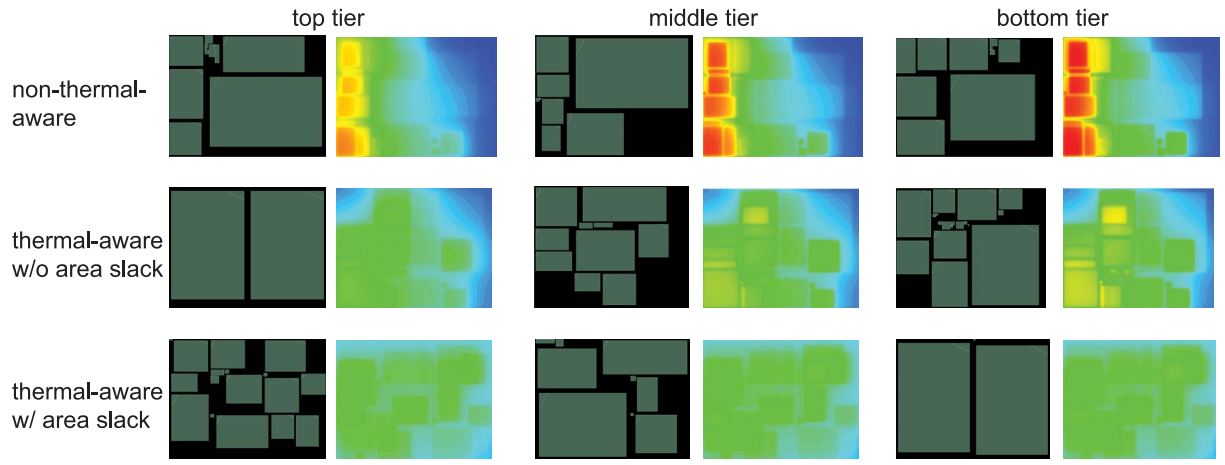

Fig. 9. Three-tier floorplanning layouts (ind_ckt benchmark with conventional package structure) with corresponding absolute temperature maps. The thermal-aware floorplans avoid stacking of high PD blocks and keep them closer to heat sink and result in 22% temperature reduction in lesser total area. The temperature range is [47 °C, 68 °C].

TABLE VII
THERMAL-AWARE FLOORPLANNING WITH TEMPERATURE MODEL DEVELOPED FOR CONVENTIONAL PACKAGE STRUCTURE (I.E., WITH HEAT SINK)

		Footprint ($\mu m x \mu m$)	Si Area (mm^2)	Inter-block WL (m)	Max Temp above room(^oC)	Average Temp above room(^oC)	Temp Gradient(^oC)	Floorplan Runtime(sec)
cf_fft_256_8								
2D		1181 x 1147	1.36	0.56	22.12	13.57	10.39	-
2-tier	Non-thermal	745 x 939	1.40 (**1.00**)	0.34	33.38 (**1.00**)	26.26	10.19	1452 (**1.00**)
	Thermal (w/o area slack)	762 x 920	1.40 (**1.00**)	0.45	31.62 (**0.94**)	25.88	8.37	1723 (**1.18**)
	Thermal (w/ area slack)	867 x 849	1.47 (**1.05**)	0.45	27.36 (**0.82**)	24.37	5.56	1780 (**1.23**)
3-tier	Non-thermal	580 x 824	1.43 (**1.00**)	0.34	48.05 (**1.00**)	39.14	13.00	1486 (**1.00**)
	Thermal (w/o area slack)	577 x 829	1.43 (**1.00**)	0.37	38.47 (**0.92**)	38.47	9.26	1769 (**1.19**)
	Thermal (w/ area slack)	891 x 560	1.50 (**1.05**)	0.35	42.84 (**0.89**)	36.69	11.31	1808 (**1.22**)
ind_ckt								
2D		3939 x 3525	13.89	10.18	15.02	10.71	6.82	-
2-tier	Non-thermal	3680 x 1994	14.68 (**1.00**)	6.43	26.57 (**1.00**)	19.76	11.19	3228 (**1.00**)
	Thermal (w/o area slack)	3603 x 1994	14.37 (**0.98**)	6.86	25.20 (**0.95**)	19.74	9.99	5552 (**1.72**)
	Thermal (w/ area slack)	3050 x 2491	15.19 (**1.03**)	7.33	23.89 (**0.90**)	18.93	9.44	5677 (**1.76**)
3-tier	Non-thermal	2591 x 1960	15.24 (**1.00**)	5.54	40.89 (**1.00**)	28.66	20.30	3600 (**1.00**)
	Thermal (w/o area slack)	2452 x 2070	15.22 (**1.00**)	5.91	35.73 (**0.87**)	28.72	13.80	6471 (**1.80**)
	Thermal (w/ area slack)	2454 x 2037	15.00 (**0.98**)	6.29	32.03 (**0.78**)	28.41	8.00	6074 (**1.69**)

at block level only with interblock nets and block powers. It has 32 blocks with 9203 nets. As we are not provided with the verilog netlist of the industry circuit and the intra-block information, we do not place and route the design and only report the HPWL. The block power numbers result in a large temperature gradient in the nonthermal aware design and the inclusion of temperature cost evaluated using our thermal model improves the temperature profile significantly. The interblock nets' switching power is obtained by timing and power analysis using PrimeTime and is considered in the final GDS-level thermal analysis. The purpose is to ensure that even with slight power increase due to increased wirelength, the thermal aware floorplan results in reduced temperature. Since interblock wirelength is very less compared to total wirelength, there is negligible increase in interconnect power due to increase in interblock wirelength.

The results of the different cases implemented during floorplanning with conventional package are summarized in Table VII. The implementations for two-tier and three-tier 3-D designs are shown for conventional package structure.

TABLE VIII
COMPARISON WITH 3DFP [9] (FFT BENCHMARK)

	Footprint ($\mu m x \mu m$)	Si Area (mm^2)	Inter-block WL (m)	Max Temp above room(^oC)
2-tier				
3DFP [9]	1005 x 735	1.48	0.60	27.63
Our FP	867 x 849	1.47	0.45	27.36
3-tier				
3DFP [9]	972 x 518	1.51	0.46	45.81
Our FP	891 x 560	1.50	0.35	42.84

2-D design metrics are also given for reference. Since the runtime is dependent on number of blocks, number of nets (for wirelength calculation), and size of the chip (for temperature estimation), we observe different runtimes for the different designs but the increase in runtime due to thermal analysis is well within tolerable limits.

We can observe that there is significant reduction in maximum temperature given the fact that there is minimum area overhead therefore satisfying the purpose of the

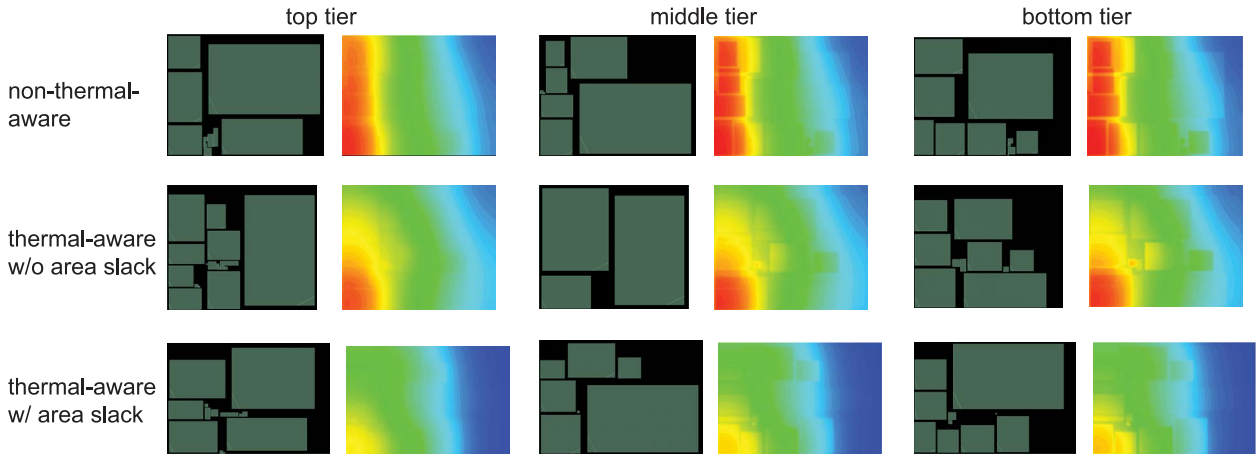

Fig. 10. Three-tier floorplanning layouts (ind_ckt benchmark with mobile package structure) with corresponding absolute temperature maps. The thermal-aware floorplans avoid stacking of high PD blocks and keep low PD blocks in middle tier. The temperature range is [42 °C, 67 °C].

TABLE IX
THERMAL-AWARE FLOORPLANNING WITH TEMPERATURE MODEL DEVELOPED FOR MODERN MOBILE PACKAGE (NO HEAT SINK)

		Footprint ($\mu m x \mu m$)	Si Area (mm^2)	Inter-block WL (m)	Max Temp above room(^oC)	Average Temp above room(^oC)	Temp Gradient(^oC)	Floorplan Runtime(sec)
cf_fft_256_8								
2D		1181 x 1147	1.36	0.56	11.96	10.91	1.88	-
3-tier	Non-thermal	580 x 824	1.43 (**1.00**)	0.3	32.78 (**1.00**)	29.13	5.10	1486 (**1.00**)
	Thermal (w/o area slack)	552 x 853	1.41 (**0.99**)	0.32	31.42 (**0.96**)	28.77	4.21	1985 (**1.34**)
	Thermal (w/ area slack)	667 x 739	1.48 (**1.04**)	0.34	28.98 (**0.88**)	28.23	1.43	1889 (**1.27**)
ind_ckt								
2D		3939 x 3525	13.89	10.18	23.24	16.43	12.00	-
3-tier	Non-thermal	2591 x 1960	15.24 (**1.00**)	5.54	39.81 (**1.00**)	27.07	23.06	3600 (**1.00**)
	Thermal (w/o area slack)	2420 x 2097	15.22 (**1.00**)	5.97	38.79 (**0.97**)	27.15	21.06	6564 (**1.82**)
	Thermal (w/ area slack)	2701 x 1949	15.79 (**1.04**)	6.27	35.18 (**0.88**)	23.07	20.23	6962 (**1.93**)

thermal-aware floorplanning (Fig. 9). Our thermal-aware floorplanner tries to reduce the gradient of the temperature variation as the average PD of the chip will remain the same because of the same chip area with the same total power dissipation. It can be clearly observed that the floorplanning process avoids stacking of high PD blocks and also forces such blocks to tiers which are closer to the heat sink. The larger and low PD blocks are placed in the critical tiers. The three-tier designs show more degree of improvement because of more options to move the blocks around. All of this becomes feasible only because of the fast and accurate monolithic 3-D IC temperature estimation model.

C. Comparison With State-of-the-Art

Cong *et al.* [5] show 56% temperature reduction but they report a 21% area increase which is significantly high overhead. They use a fast but less accurate hybrid resistive model and another accurate but relatively slow resistive model selectively within their floorplanning.

PD and total 3-D overlap in the cost function to incorporate thermal awareness during design has been used for thermal aware 3-D floorplanning [9]. Their tool called 3DFP is available for public use. Since our thermal model directly gives an accurate measure of temperature, it is more effective in the design process. We use 3DFP for our benchmarks and compare the results. Since the number of moves during annealing and other annealing parameters differ in the two floorplanners, we compare the quality of the floorplan results and not the total runtime.

Table VIII shows the comparison results of 3DFP and our thermal-aware floorplanner for the FFT benchmark. With the help of direct temperature measurement during annealing using our fast and accurate model, we successfully obtain better floorplans in all respects, namely area, wirelength, and temperature.

D. Thermal Floorplanning for Modern Mobile Package

Table IX shows the results for thermal aware floorplanning for three-tier 3-D ICs with mobile package structure for the two benchmarks. The power densities have been scaled for the designs to satisfy the peak temperature limitations for mobile package structures. Fig. 10 shows the floorplanning results along with their temperature maps for industrial circuit benchmark with mobile packaging structure. For such type of packaging, the middle tier is most critical as heat dissipates in both directions. Our thermal model correctly maps the mobile package system along this line and the larger, low PD blocks get placed in the middle tier (Tier1) without any area overhead. For the thermal-aware floorplanning with no area overhead, Tier1 ends up with only three large blocks with low PD reducing maximum temperature.

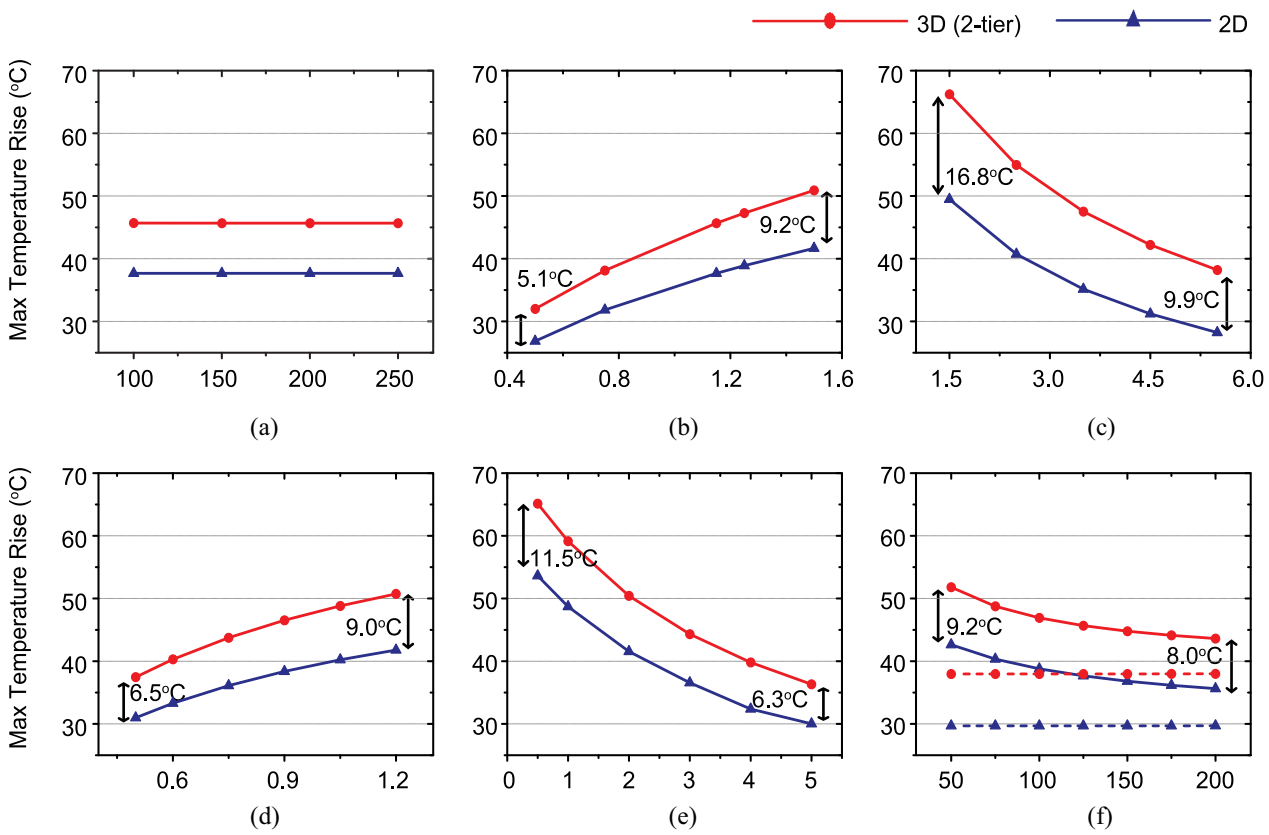

Fig. 11. Impact of change of various material thicknesses and conductivity on maximum temperature of ind_ckt benchmark with mobile package structure for 2-D IC and 3-D IC (two-tier). Dotted lines in (f) is the average temperature variation. (a) EMI Thickness (in μm). (b) PCB thickness (in mm). (c) PCB conductivity (in W/mK). (d) TIM thickness (in mm). (e) TIM conductivity (in W/mK). (f) Bulk Thickness (in μm).

VII. Mobile Package Optimization: Impact of the Materials

Earlier in Section IV, we discussed the different layers which play an important role in thermal behavior of mobile packages due to two-way heat flow. In this section, we study the impact of thickness and conductivity of some of these materials in mobile package on the maximum temperature. In particular, we study the difference in impact on 2-D ICs and 3-D ICs and highlight the fact that a good change in package properties is more beneficial for 3-D ICs than 2-D ICs. The different material properties varied are the ones specified with range of values in Table II. Fig. 11 plots the maximum temperature of 2-D IC and two-tier 3-D IC with change in various material thicknesses and conductivities. 3-D ICs have multiple layers of power dissipation source, while 2-D ICs have just one device layer dissipating power. Therefore, the impact of changing the thickness and conductivities of package layers is more prominent in 3-D IC than 2-D IC.

Thermal interface material (TIM) is a necessary layer to have smooth continuous contact between the uneven bulk surface and the EMI (or heat spreader for conventional packages). They are poor conductors of heat ($<$ 5 W/mK) but provide a better and continuous thermal interface than silicon-air and air-metal interface. Because the TIM provides a high resistive path to heat flow, changes in the thickness of the layers beyond TIM (EMI and graphite) have negligible impact. The thermal circuit is equivalent to a large resistance (TIM here) in series with a small resistance (EMI) whose value changes with change in thickness but has minor impact on equivalent resistance. This is shown in Fig. 11(a) where change in EMI thickness has no impact on the maximum temperature of 2-D IC as well as 3-D IC.

Fig. 11(b) and (c) shows the impact of change of PCB thickness and vertical conductivity, respectively. Since the PCB is the major heat flow path in the downward direction [Fig. 4(b)], improvement in its thermal resistance reduces maximum temperature significantly. Reduction in thickness or increase in vertical conductivity both contribute to reduced thermal resistance. The change is more prominent in 3-D ICs because the bottom tier now finds a much lower resistive path in the downward direction and the amount of heat transferred toward the upper tier is reduced therefore reducing the degree of thermal coupling in the two tiers.

TIM has a similar impact as PCB as shown in Fig. 11(d) and (e). The vertical conductivity change brings about a larger degree of maximum temperature change because the conductivity value itself is changed from 0.5 to 5 W/cm^2 which is a 10$\times$ increase. Though not shown in the plots, the average temperature also follows the same trend as the maximum from Fig. 11(a)–(e) as most of these layers are toward the end of the equivalent thermal circuit. Handle bulk on the other hand is an intermediate layer in the heat flow path. Silicon being a reasonably good conductor helps more in lateral spreading to reduce the temperature gradient across the

design but has no impact on the overall average since there is a poor conducting TIM layer further up in the path. Fig. 11(f) shows the change in maximum temperature with change in handle bulk thickness. The average temperature is also shown in dotted lines and has no change with change in bulk thickness but the maximum temperature reduces due to change in gradient. Also the difference in impact on 2-D IC and 3-D IC is not very high as observed for the PCB and TIM layers.

From these studies and observations, it is clear that package structure plays an important role in determining the maximum temperature and the same change in package properties exhibit more benefits for 3-D ICs with mobile packages. This can be used to plan a good package structure to start with or as a post physical design technique to further improve the thermal reliability of 3-D ICs after obtaining a thermal aware layout.

VIII. Conclusion

We studied the unique thermal properties of monolithic 3-D ICs and compared their thermal behavior with TSV-based 3-D ICs. It was observed that due to the absence of bulk silicon substrate in monolithic technology, their is no lateral spreading near the device layer. Also the very thin ILD and absence of bonding layer results in heavy vertical thermal coupling and improves the temperature profile of the tiers away from the heat sink compared to TSV-based 3-D ICs. We utilized these properties to our advantage and developed a methodology to obtain package-aware fast and accurate thermal analysis model for monolithic 3-D ICs with different number of stacking layers with the help of nonlinear regression. These models were verified against full chip FEA thermal simulations and found to be highly accurate. We used this model in a thermal aware floorplanner to show significant temperature reduction with minimum or no area overhead for both conventional packages with heat sink and mobile packages. The speed of our thermal model enables us to use it in the floorplanning process without any runtime issues. We also studied the impact of material property changes in mobile package structure to enable thermal package optimization for 3-D ICs.

References

[1] 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.

[2] S. S. Sapatnekar, "Addressing thermal and power delivery bottlenecks in 3D circuits," in *Proc. Asia South Pac. Design Autom. Conf.*, Yokohama, Japan, Jan. 2009, pp. 423–428.

[3] P. Emma *et al.*, "3D stacking of high-performance processors," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit.*, Orlando, FL, USA, Feb. 2014, pp. 500–511.

[4] Y. Chen, E. Kursun, D. Motschman, C. Johnson, and Y. Xie, "Analysis and mitigation of lateral thermal blockage effect of through-silicon-via in 3D IC designs," in *Proc. Int. Symp. Low Power Electron. Design*, Fukuoka, Japan, Aug. 2011, pp. 397–402.

[5] J. Cong, J. Wei, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3D ICs," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design.*, San Jose, CA, USA, Nov. 2004, pp. 306–313.

[6] W. Huang *et al.*, "Compact thermal modeling for temperature-aware design," in *Proc. DAC*, Austin, TX, USA, 2004, pp. 878–883.

[7] P. Zhou *et al.*, "3D-STAF: Scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design.*, San Jose, CA, USA, Nov. 2007, pp. 590–597.

[8] D.-C. Juan, S. Garg, and D. Marculescu, "Statistical thermal evaluation and mitigation techniques for 3D chip-multiprocessors in the presence of process variations," in *Proc. DATE*, Grenoble, France, 2011, pp. 1–6.

[9] 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. 6–104.

[10] Y.-J. Lee and S. K. Lim, "Ultrahigh density logic designs using monolithic 3-D integration," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 12, pp. 1892–1905, Dec. 2013.

[11] P. Batude *et al.*, "3D sequential integration opportunities and technology optimization," in *Proc. IEEE Int. Interconnect Technol. Conf. Adv. Metallization Conf. (IITC/AMC)*, San Jose, CA, USA, May 2014, pp. 373–376.

[12] H. Wei *et al.*, "Cooling three-dimensional integrated circuits using power delivery networks," in *Proc. IEDM*, San Francisco, CA, USA, 2012, pp. 14.2.1–14.2.4.

[13] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Tier-partitioning for power delivery vs cooling tradeoff in 3D VLSI for mobile applications," in *Proc. 52nd Annu. Design Autom. Conf. (DAC)*, San Francisco, CA, USA, 2015, pp. 1–6. [Online]. Available: http://doi.acm.org/10.1145/2744769.2744917

[14] Oracle. (Sep. 30, 2013). *OpenSPARC T2*. [Online]. Available: http://www.oracle.com

[15] A. B. Kahng, B. Lin, and K. Samadi, "Improved on-chip router analytical power and area modeling," in *Proc. ASP-DAC*, Taipei, Taiwan, 2010, pp. 241–246.

[16] K. Athikulwongse, M. Ekpanyapong, and S. K. Lim, "Exploiting die-to-die thermal coupling in 3-D IC placement," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, no. 10, pp. 2145–2155, Oct. 2014.

[17] (Sep. 30, 2013). *Salford Systems*. [Online]. Available: http://www.salford-systems.com/ products/mars

[18] F. Beneventi, A. Bartolini, A. Tilli, and L. Benini, "An effective gray-box identification procedure for multicore thermal modeling," *IEEE Trans. Comput.*, vol. 63, no. 5, pp. 1097–1110, May 2014.

[19] F. Beneventi, A. Bartolini, P. Vivet, D. Dutoit, and L. Benini, "Thermal analysis and model identification techniques for a logic + WIDEIO stacked DRAM test chip," in *Proc. Design Autom. Test Europe Conf. Exhibit. (DATE)*, Dresden, Germany, Mar. 2014, Art. ID 332.

[20] X. Tang, R. Tian, and D. F. Wong, "Fast evaluation of sequence pair in block placement by longest common subsequence computation," in *Proc. Design Autom. Test Europe.*, Paris, France, 2000, pp. 106–111.

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, modeling, and analysis using through-silicon-via-based and monolithic 3-D IC technology.

Shreepad Panth (S'11) received the B.S. degree from Anna University, Chennai, India, in 2009, and the M.S. degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2011, where he is currently pursuing the Ph.D. degree under the supervision of Dr. S. K. Lim.

He has authored over 20 publications in top conferences and journals. His current research interests include physical design methodologies for monolithic 3-D ICs.

Mr. Panth was a recipient of the Best Paper Award at Asian Test Symposium 2012 and the Best Paper Award Nominee at International Symposium on Physical Design 2014 and Design Automation Conference 2014.

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

He joined Qualcomm Research, San Diego, in 2011, where he is currently a Staff Research Engineer, focusing on 3-D IC EDA solutions and 3-D IC architecture-level design space explorations. He has authored 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, and very-large-scale integration design manufacturing interface.

Dr. Samadi was a recipient of the two best paper nominations and a best paper award.

Mehdi Saeidi (M'11) received the B.Sc. degree from the Sharif University of Technology, Tehran, Iran, the M.Sc. degree from Shiraz University, Shiraz, Iran, and the Ph.D. degree from Auburn University, Auburn, AL, USA, in 2005, all in mechanical engineering.

He held various positions as Thermal, Packaging, and Test Engineer at Broadcom Corporation, Irvine CA, USA, and Intel Corporation, Chandler, AZ, USA. He is currently a Senior Staff Engineer/Manager with Qualcomm Technologies Inc., San Diego, CA, USA. His current research interests include electronics thermal design and simulation, SoC design and architecture, and test and packaging technologies. He holds 24 patents (published/pending) and has published several conference/journal papers in the above areas.

Mr. Saeidi was a recipient of the best paper award from the 2014 IEEE International Conference on 3-D System Integration.

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

He is currently the Director of Engineering with Qualcomm Research, San Diego, CA, USA, where he leads a team in advanced nanotechnology and semiconductor research. He has held various engineering positions in Analog Devices, Norwood, MA, USA, AMD, Sunnyvale, CA, USA, Motorola, Chicago, IL, USA, and Qualcomm. He has authored/co-authored over 50 patents/patent publications and numerous conference/journal papers in very-large-scale integration (VLSI) technology, SPICE modeling, IC design, test, and design automation. His current research interests include emerging semiconductor devices, predictive device, circuit modeling, novel VLSI circuits and architecture, next generation 3-D IC technology and design, emerging 3-D VLSI circuit, architecture and system integration, design automation, advanced thermal modeling, and thermal aware design methodologies.

Dr. Du was a recipient of the IEEE Subthreshold Microelectronics Conference. He has been serving on the Technical Program Committee since 2011. He has also been serving on the Advisory Committee of the IEEE S3S Conference since 2013.

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), Los Angeles, CA, USA, in 1994, 1997, and 2000, respectively.

From 2000 to 2001, he was a Post-Doctoral Scholar with UCLA, and a Senior Engineer with Aplus Design Technologies, Inc., Los Angeles. He joined the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2001, where he is currently a Dan Fielder Professor of Electrical and Computer Engineering. He led the Cross-Center Theme on 3-D Integration for the Focus Center Research Program, Semiconductor Research Corporation, from 2010 to 2012. He has authored 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. His research on 3-D IC reliability is featured as a research highlight in the Communication of the ACM in 2014.

Dr. Lim was a recipient of the National Science Foundation Faculty Early Career Development Award in 2006, the ACM Special Interest Group on Design Automation Distinguished Service Award in 2008, and the Best Paper Awards from TECHCON 2011, TECHCON 2012, and ATS 2012. His work was nominated for the Best Paper Award at ISPD 2006, International Conference on Computer-Aided Design 2009, Custom Integrated Circuits Conference 2010, DAC 2011, DAC 2012, International Symposium on Low Power Electronics and Design 2012, and DAC 2014. He was on the Advisory Board of the ACM SIGDA from 2003 to 2008. He was an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS from 2007 to 2009. He has also been an Associate Editor of the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS since 2013. He has served on the Technical Program Committee of several premier conferences in Electronic Design Automation. He was a member of the Design International Technology Working Group of the International Technology Roadmap for Semiconductors.