Fast and Accurate Thermal Modeling and Optimization for Monolithic 3D ICs

Sandeep Kumar Samal[†], Shreepad Panth[†], Kambiz Samadi[§],
Mehdi Saedi[§], Yang Du[§], and Sung Kyu Lim[†]
[†]School of ECE, Georgia Institute of Technology, Atlanta, GA
[§]Qualcomm Research, San Diego, CA
sandeep.samal@gatech.edu, limsk@ece.gatech.edu

ABSTRACT

In this paper, we present a comprehensive study of the unique thermal behavior in monolithic 3D ICs. In particular, we study the impact of the thin inter-layer dielectric (ILD) between the device tiers on vertical thermal coupling. In addition, we develop a fast and accurate compact full-chip thermal analysis model based on non-linear regression technique. Our model is extremely fast and highly accurate with an error of less than 5%. This model is incorporated into a thermal-aware 3D-floorplanner that runs without significant runtime overhead. We observe up to 22% reduction in the maximum temperature with insignificant area and performance overhead.

Categories and Subject Descriptors

B.8.2 [**Performance and Reliability**]: Performance Analysis and Design Aids

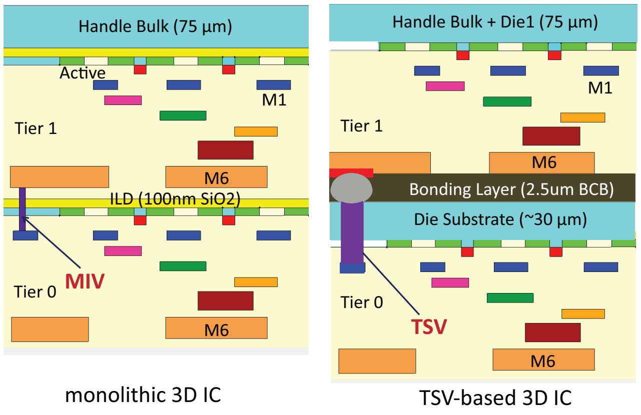
Keywords

Monolithic 3D; Thermal; Modeling; Optimization

1. INTRODUCTION

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

Monolithic 3D ICs can overcome the shortcomings of TSV-based 3D ICs; however, one major concern with 3D ICs in general is the increase in power density which leads to high temperature values. The reduction in footprint area effectively increases the power density by the same factor. Even if we achieve power reduction by going 3D, the increased power density 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 3D ICs. The major bottleneck of considering thermal aspect within 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.



monolithic 3D IC

TSV-based 3D IC

Figure 1: 2-tier 3D IC layer structure (heat sink on top) of Monolithic 3D IC vs TSV-based 3D IC

Attempts have been made to develop accurate temperature evaluation models to be included within the chip design process. The use of compact resistive thermal grid network to estimate the temperature profile of a chip has been studied [2]. They estimate the temperature during the floorplanning process and insert whitespace for dummy vias. The calculation of resistive network solving still consumes some runtime and the insertion of whitespace increases the area further, diminishing the 3D IC benefits. They report 56% reduction in temperature but with a large area increase of 21%. The optimization of silicon area is important in 3D ICs along with the temperature rise and we cannot sacrifice too much area for temperature improvement. The modeling of temperature based on total leakage power dissipation and its use in the tier-planning of similar layout processor chips has also been studied [3]. Another work uses the 3D overlap estimation along with power density calculations for thermal-aware planning [4]. All these methods are either targeted for TSV-based 3D IC design or incur extra runtime and area or use indirect methods of thermal analysis. None of the works address monolithic 3D ICs which interestingly exhibit different thermal behavior due to their layer structure.

In this work (1) We study and explain the thermal characteristics of monolithic 3D ICs for the first time with comparison to TSV-

This work is supported by Qualcomm Research.

DAC'14, June 01 - 05 2014, San Francisco, CA, USA
Copyright 2014 ACM 978-1-4503-2730-5/14/06 ... $15.00.
http://dx.doi.org/10.1145/2593069.2593140.

Table 1: The different materials used in the layers, their thermal conductivities and vertical thicknesses

Layer/Structure	Material	Thermal Conductivity (W/m-K)	Vertical Thickness
Monolithic			
Handle Bulk	Silicon	141	$75\mu m$
ILD (Inter-tier)	SiO_2	1.38	$100nm$
TSV-based			
Handle Bulk	Silicon	141	$75\mu m$
Die0 Substrate	Silicon	141	$30\mu m$
Bonding Layer	BCB	0.29	$2.5\mu m$
TSV	Copper	401	$30\mu m$
TSV-bump	Solder	50	$2.5\mu m$

based 3D ICs.(2) We identify the factors affecting temperature and develop a very fast and accurate non-linear regression based temperature evaluation model for monolithic 3D ICs. This is first work on thermal modeling for monolithic 3D ICs.(3) We use our model to carry out thermal-aware 3D floorplanning and show significant reduction in maximum temperature with minimal or no area and performance overhead.

2. NEW THERMAL ISSUES

2.1 Monolithic 3D Integration

A typical two-tier monolithic stackup is shown in Figure 1 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 inter-layer dielectric (ILD) is deposited over the metal layers for growing the next device layer. This device layer along with the metal layers make up the other tier (Tier 0) of the 3D stackup. The transistors in these layers are processed with low temperature process (<650 ^oC).

2.2 Material and Structural Differences

The differences in fabrication process of monolithic 3D and TSV-based 3D result in significant differences in their thermal behavior. Figure 1 and Table 1 highlight the differences in the materials used in the two technologies.

In TSV-based 3D ICs, copper TSVs and solder 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). Typical materials used for underfill are required to be soft and elastic and usually 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. The presence of this underfill which is around $2.5\mu m$ thick impedes the heat flow from Tier0 towards the heat sink resulting in considerable temperature rise in Tier0. However, heat from Tier0 passes through a thick silicon substrate ($30\mu m$) before reaching the underfill wall and the heat spreads laterally because of many lateral heat flow paths within the substrate. The heat conduction in Tier1 is better as it is closer to the heat sink without any oxide between the device layer and the handle bulk.

Monolithic 3D ICs on the other hand do not have bonding layer and bulk substrate while the different tiers are separated by interlayer dielectrics (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. The absence of bulk substrate and the extremely thin layers reduce the lateral thermal conductivity to almost zero which results in heavy tier-to-tier thermal coupling. The heat flows vertically up only 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 2D location in two tiers due the rise across the 100nm ILD. Also the maximum temperature of the tier closest to the heat sink is more than that of TSV-based 3D IC due to the presence of additional oxide layer which is a poor conductor.

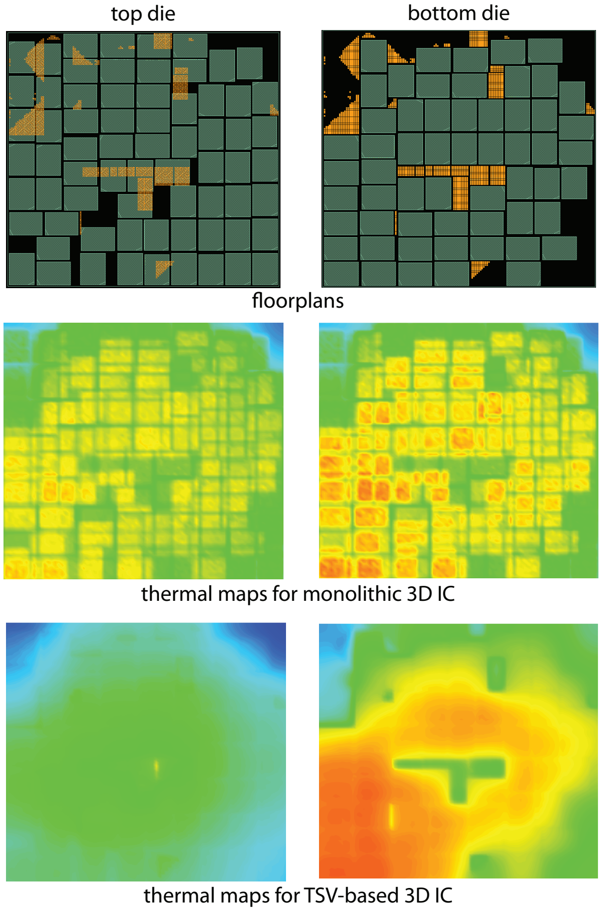


Figure 2: Temperature maps of same 2-tier 3D floorplan (originally designed for TSV-based 3D IC) in Monolithic 3D IC technology and TSV-based 3D technology. The temperature range is [61°C, 71°C].

2.3 Temperature Map Comparisons

Figure 2 shows the temperature map of a same 2-tier 3D layout in monolithic technology and TSV-based technology. The layout was originally designed for a TSV-based design. 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, we use the same layout for fair comparison from the thermal point of view with TSVs replaced by MIVs. In practice, MIVs are much smaller and their design will consume much lesser area. For the TSV-based 3D IC temperature map, we can see clearly that the presence of TSVs help in improv-

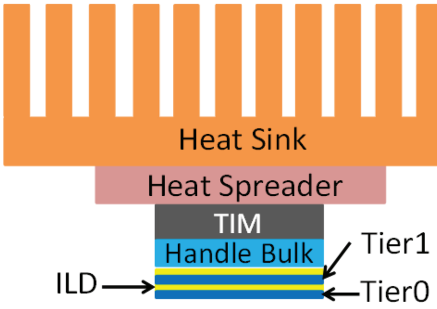


Figure 3: 3D IC Packaging structure for cooling with heat sink

ing the conduction significantly in Tier0 with cooler regions at TSV locations. The temperature of other regions is quite high due to the heat flow obstruction by the 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 heat across the two tiers which smears the temperature profile in each tier. This is because of the bulk silicon allowing multiple lateral heat flow paths.

For monolithic 3D IC design, 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. Tier0 in monolithic 3D is much cooler than Tier0 in TSV-based 3D as there is no bonding layer obstruction while Tier1 of monolithic 3D IC is hotter than Tier1 of TSV-based 3D IC because of the oxide. Wei *et al.* also compared TSV based 3D IC with monolithic 3D ICs but did not consider the underfill layer [5]. The mass production of TSV-based 3D ICs without any underfill is highly unlikely due to stress related issues. Therefore, we need to consider them during the thermal study of TSV-based 3D ICs and then compare with monolithic 3D ICs. This very poor conducting bonding layer in TSV-based 3D ICs significantly worsen the temperature of tiers away from heat sink. If we do not consider this layer, then TSV-based 3D ICs will be better than monolithic 3D ICs thermally.

The key points from the above study which help us to successfully develop a thermal model are (1) Monolithic 3D 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 3D and all tiers have similar temperature profile with differing absolute values due to rise across ILDs. (3) In monolithic 3D 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 3D ICs where they directly affect each other. (4) MIVs do not play an important role in heat conduction like TSVs due to small size and thickness.

3. FAST THERMAL ANALYSIS MODEL

3.1 Model Development

We use non-linear regression to accurately model the temperature of monolithic 3D 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 [6]. We set temperature as our target quantity and model it after successfully determining the different parameters of the monolithic 3D IC on which it depends.

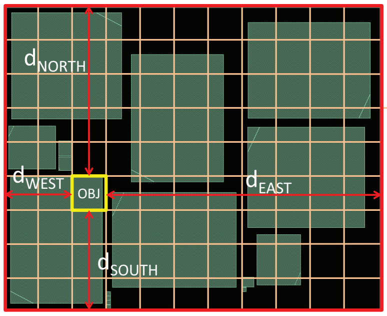


Figure 4: Final model structure with an objective tile (yellow) and rest of the tiles. Their power values along with 2D distances from boundary and 3D tier number are used as inputs for temperature calculation

3.1.1 Initial Experiments

We divided the entire chip into a tile based structure for each tier (Fig 4). Each of the tiles was randomly assigned a power value such that the power density lies between 0-100W/cm^2. Full chip thermal FEA with 20μm X 20μm mesh was carried out on these test cases with ANSYS Fluent and supporting scripts. We consider conventional cooling method which uses heat spreader and heat sink (Fig 3). Almost 100% of the heat dissipates through the heat sink.

The primary goal was to obtain a model to calculate the temperature of each tile without carrying out full FEA simulations. To correctly determine the number neighboring levels to be covered, we carried out experiments by starting with 20 levels of neighboring tile levels and dropping the farthest neighbor one at a time to see if it affects the results . Since each tile is 100μm X 100μm and there are 20 levels, the chip size is 4.1mmx4.1mm. The model effectiveness is measured in terms of the generalized cross validation (GCV). The results of this experiment show that the amount of error increases as we remove the farthest neighbors (Table 2). Therefore, the farthest neighbors 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 very large and as pointed out earlier, all of this heat primarily goes vertically to the handle bulk layer therefore indirectly affecting the objective tile temperature.

Using the same raw data as above, we carried out analysis from a different viewpoint. The entire region (20 levels) was divided into different number of equal regions viz. 20 levels, 10 levels, 5 levels, 4 levels, 2 levels and finally a single level with all 20 tile rings treated as one. We then used these partitions as variables to develop the model and compared the resulting model in terms of GCV and average absolute error (Table 3). The results show that it is not necessary to have fine grained neighbors in the model. This is because the effect of neighbors is indirect through the handle bulk which is 75μm thick and is silicon. The most important variable is always the last level which has maximum magnitude of power. We also observed that the location of a particular tile in the layout affects its temperature value.

3.1.2 Modeling Technique

From the 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)

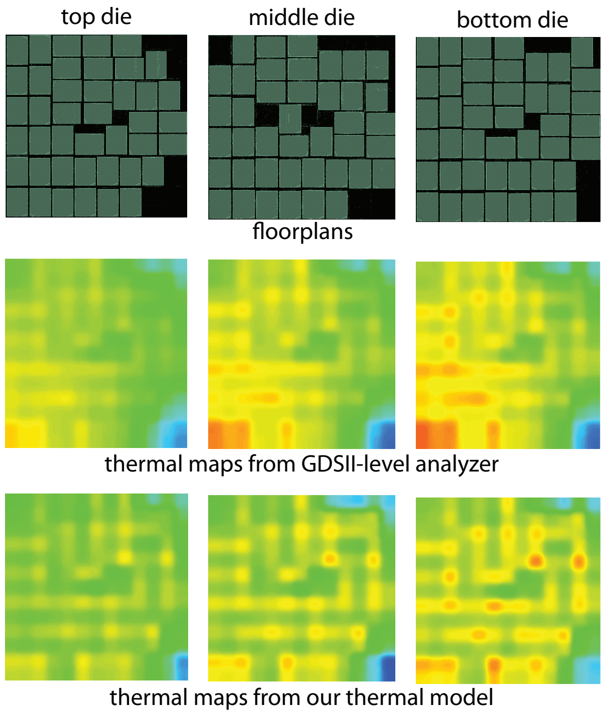


Figure 5: Model Accuracy: FEA simulation vs Our Temperature Model for 256-bit Multiplier. The temperature range is [63°C, 79°C].

Table 2: 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

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 2D boundaries (4 variables) and (8) Distance from vertical boundaries (3D location). The exponential increase in leakage with temperature can be taken care of by separating the power inputs into its components viz. dynamic and leakage powers and updating the leakage powers with temperature increase till a specified tolerance level is met.

Figure 4 shows the division of chip and the 2D dimension 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 non-linear regression technique [7]. 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 account for the 3D distance from heat sink. The individual tile size is fixed at 100μm X 100μm. 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 separate thermal analysis models for both 2-tier and 3-tier 3D cases with conventional packaging structures.

Table 3: Experimental results of modeling with the entire chip area considered completely but with different number of partitions

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

Table 4: Full chip Thermal Analysis Runtime Comparison for 3-tier 3D IC (1.3 mm x 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	59.5	**2.7 x 10^{5}**

3.1.3 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 3D chip sizes and power distributions, we carry out detailed thermal analysis of whole chip testcases which cover chip dimensions from 1mm to 5mm (in steps of 1mm) with aspect ratio lying between 0.5 and 2. Each chip is divided into 100μm X 100μm 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. Power density values of the tiles are randomly distributed from 0-100 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.

We observed that the modeling is more accurate when all the samples have a random power density distribution with fixed average rather than with varying average. Therefore, we use samples with power density varying randomly from 0-100W/cm^2 which results in an average power density of all samples close to 50W/cm^2 (average of a random distribution). However, the power density of the practical case to be modeled will vary from design to design. The trend prediction of our model is always correct irrespective of this actual average power density. The values are just shifted up or down and need a constant correction offset depending on the actual power density being greater than or less than 50W/cm^2. From various practical example cases, we determine this offset as a multiple of the difference of the actual average power density (PD) of chip and the samples' power density (=50 W/cm^2 here). The exact multiplying coefficient depends on the samples' power density variation but will always remain constant for a developed model irrespective of the actual chip evaluated. For floorplanning purposes, only the trend of temperature is important and that is always correctly captured with or without offset.

3.2 Model Accuracy

The temperature evaluated by the model is the rise above room temperature. To get the absolute temperature, we add the rise to the

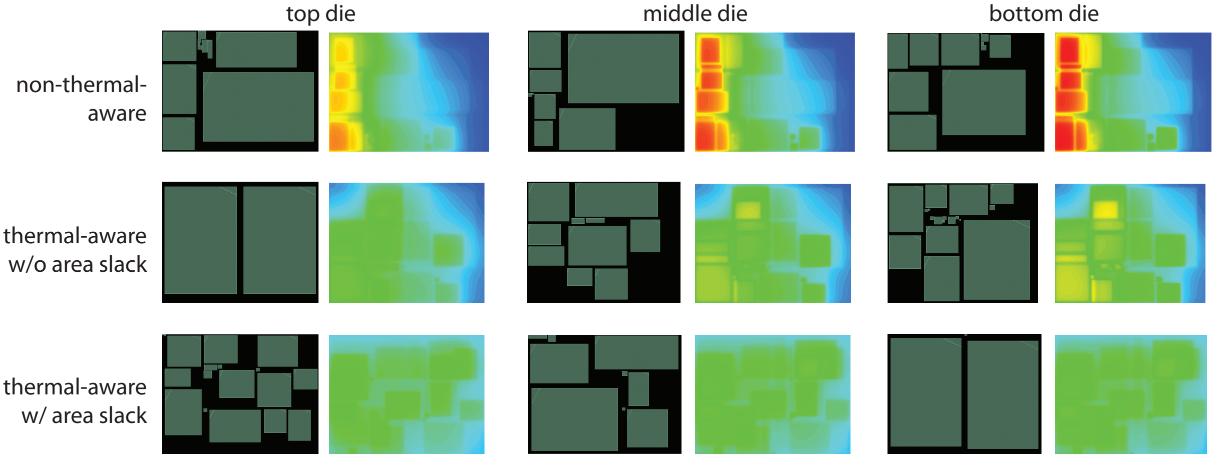


Figure 6: 3-tier Floorplanning Layouts (ind_ckt benchmark) with corresponding absolute temperature maps. The thermal-aware floorplans avoid stacking of high power density blocks and keep them closer to heat sink and result in 22% temperature reduction in lesser total area. The temperature range is [47°C, 68°C].

Table 5: Thermal-Aware Floorplanning With Temperature Model Developed for Conventional Package Structure

		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	44.20 **(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)**

room temperature. The testing sample set gives an absolute average error of less than 1%. For practical designs, the average error is less than 5%. Figure 5 shows the accuracy of model for a testcase, designed for 3-tier 3D. The top row show the layouts of the individual tiers of 3-tier 3D 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 have a temperature analysis.

3.3 Runtime Comparison

Since our model is compact 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 4 summarizes the runtime comparison with GDS-level FEA simulation and Hotspot [8]. Our model is 4.9x10^6 times faster than FEA simulation and 2.7x10^5 faster than 3D Hotspot analysis.

4. THERMAL-AWARE FLOORPLANNING

4.1 Floorplanning Algorithm

We use simulated annealing of sequence pairs to obtain the best floorplan depending on the weighted cost function specified. The non thermal-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 3D 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. Since area is directly proportional to cost, especially in miniaturized systems, 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 which will increase temperature further. However, if the blocks are not given freedom of movement, the solution space for temperature optimized floorplans within the constrained area becomes smaller and there won't be significant temperature reduction. Therefore we use a step by step process to obtain the temperature optimized floorplan.

We first run the non-thermal floorplanner without any temperature cost and obtain the wirelength number. In the next step, given

Table 6: Comparison with 3DFP [4] (FFT benchmark)

	Footprint $(\mu mx\mu m)$	Si Area (mm^2)	Inter-block WL (m)	Max Temp above room(^oC)
2-tier				
3DFP [4]	1005 x 735	1.48	0.60	27.63
Our FP	867 x 849	1.47	0.45	27.36
3-tier				
3DFP [4]	972 x 518	1.51	0.46	45.81
Our FP	891 x 560	1.50	0.35	42.84

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 during the annealing process. 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. The wirelength calculation for all nets for a given sequence pair determines the time complexity of the floorplanning process. Therefore, the addition of thermal analysis using our model has insignificant overhead even with millions of moves.

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.

4.2 Floorplanning Results

Due to space limitations, 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 inter-block nets. The industry circuit benchmark was obtained at block level only with inter-block 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 only report the HPWL. The block power numbers result in a large temperature gradient in the non thermal aware design and the inclusion of temperature cost evaluated using our thermal model improves the temperature profile significantly. The inter-block 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 inter block wirelength is very less compared to total wirelength, there is negligible increase in interconnect power due to increase in inter-block wirelength.

The results of the 2-tier and 3-tier 3D floorplans are summarized in Table 5. 2D 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 with minimum area overhead therefore satisfying the purpose of the thermal-aware floorplanning (Fig 6). Our thermal-aware floorplanner tries to reduce the gradient of the temperature variation as the average power density 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 power density blocks and also forces such blocks to tiers which are closer to the heat sink. The larger and low power density blocks are placed in the tiers away from heat sink. The 3-tier designs show more degree of improvement because of more options to move the blocks around.

4.3 Comparison with State-of-the-Art

Power density and total 3D overlap in the cost function to incorporate thermal awareness during design has been used for thermal aware 3D floorplanning [4]. 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 with that of our thermal-aware floorplanner. 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 6 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 terms of area, wirelength and temperature.

5. CONCLUSION

We studied the unique thermal properties of monolithic 3D ICs and compared their thermal behavior with TSV-based 3D 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 improves the temperature profile of the tiers away from the heat sink unlike TSV-based 3D ICs. We utilized these properties to our advantage and developed a methodology to obtain packaging-aware fast and accurate thermal analysis models for monolithic 3D ICs with different number of stacking layers. 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. The speed of our thermal model enables us to use it in the floorplanning process without any runtime issues.

6. REFERENCES

[1] P. Batude, *et al*, “3-D Sequential Integration: A Key Enabling Technology for Heterogeneous Co-Integration of New Function With CMOS,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 4, pp. 714–722, 2012.

[2] J. Cong, J. Wei, and Y. Zhang, “A Thermal-Driven Floorplanning Algorithm for 3D ICs,” in *ICCAD*, 2004, pp. 306–313.

[3] 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 *DATE*, 2011, pp. 1–6.

[4] W.-L. Hung, G.M.Link, Y. Xie, N. Vijaykrishnan, and M.J.Irwin, “Interconnect and Thermal-aware Floorplanning for 3D Microprocessors,” in *ISQED*, 2006, pp. 99–104.

[5] H. Wei, *et al*, “Cooling three-dimensional integrated circuits using power delivery networks,” in *IEDM*, 2012, pp. 14.2.1–14.2.4.

[6] A. B. Kahng, B. Lin, and K. Samadi, “Improved On-Chip Router Analytical Power and Area Modeling,” in *ASP-DAC*, 2010, pp. 241–246.

[7] Salford Systems, http://www.salford-systems.com/products/mars.

[8] W. Huang, M. Stan, K. Skadron, K.Sankaranarayanan, S. Ghosh, and S. Velusamy, “Compact thermal modeling for temperature-aware design,” in *DAC*, 2004, pp. 878–883.