Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 14(4): 258-267, Dec. 2016 Regular paper

Machine Learning Based Variation Modeling and Optimization for 3D ICs

Sandeep Kumar Samal[1], Guoqing Chen[2], and Sung Kyu Lim[1*], *Member, KIICE*

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[2]Advanced Micro Devices, Beijing, China

Abstract

Three-dimensional integrated circuits (3D ICs) experience die-to-die variations in addition to the already challenging within-die variations. This adds an additional design complexity and makes variation estimation and full-chip optimization even more challenging. In this paper, we show that the industry standard on-chip variation (AOCV) tables cannot be applied directly to 3D paths that are spanning multiple dies. We develop a new machine learning-based model and methodology for an accurate variation estimation of logic paths in 3D designs. Our model makes use of key parameters extracted from existing GDSII 3D IC design and sign-off simulation database. Thus, it requires no runtime overhead when compared to AOCV analysis while achieving an average accuracy of 90% in variation evaluation. By using our model in a full-chip variation-aware 3D IC physical design flow, we obtain up to 16% improvement in critical path delay under variations, which is verified with detailed Monte Carlo simulations.

Index Terms: 3D ICs, Variation, Machine-learning, Regression

I. INTRODUCTION

With the advancement of very-large-scale integration (VLSI) technology, three-dimensional integrated circuits (3D ICs) have generated great interest in recent years. Researchers have already demonstrated actual 3D IC implementations on silicon [1, 2]. In general, 3D ICs offer many design advantages over their 2D IC counterpart in terms of reduction in footprint area, reduction in wirelength resulting in switching power savings, heterogeneous stacking of different dies, and higher bandwidth by stacking of memory over logic.

However, as with every new technology, 3D ICs come with new issues like high power density leading to thermal and power delivery issues [3, 4], modeling of 3D parasitics which is under research and not yet fully developed, and lack of CAD tools for actual 3D place and route. Another major difference when compared to 2D ICs is the addition of die-to-die variations in TSV-based 3D ICs where the different dies are stacked together and data/clock paths run across multiple dies (Fig. 1). With technology nodes going down to 14 nm and to 7 nm in future, variation is one of the major factors that need to be taken care of very carefully to have a good yield. 3D ICs introduce a new source of variation in the same circuit along with existing systematic and local random variation in 2D ICs which pose new design challenges [5, 6]. The physical design approach needs to incorporate these new variations during optimization to have a good yield in terms of meeting performance and to prevent over design which may lead to higher power and longer design time.

Received 29 November 2016, **Revised** 30 November 2016, **Accepted** 07 December 2016
***Corresponding Author** Sung Kyu Lim (E-mail: limsk@ece.gatech.edu, Tel: +1-404-894-0373)
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

Open Access http://doi.org/10.6109/jicce.2016.14.4.258 print ISSN: 2234-8255 online ISSN: 2234-8883

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright ⓒ The Korea Institute of Information and Communication Engineering

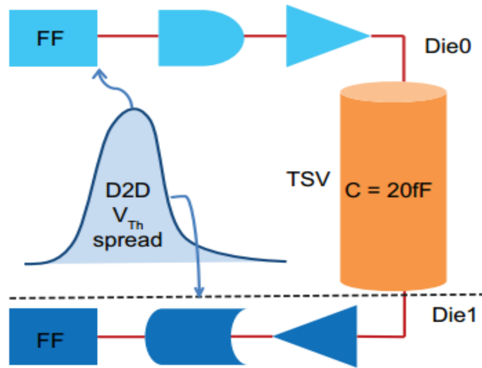

Fig. 1. Addition of die-to-die variations and large TSV (approximately 20 fF load equivalent).

Existing techniques of on-chip variation modeling during physical design are well established for 2D ICs but they need to be developed for 3D IC designs. Use of machine learning-based modeling techniques has recently gained a lot of interest [7]. In this work, we explore non-linear regression as the machine learning techniques to develop a fast accurate variation model for two-tier 3D ICs along with 2D ICs. The major contributions of our work are as follows:

- We study the variation situation in 3D IC compared to 2D IC and the limitation of current 2D IC techniques for advanced on-chip variation analysis in 3D ICs. We also demonstrate the difference in the variation impact on mean delay for both 3D ICs and 2D ICs (Section II).
- We conduct experiments to identify the various factors which impact the delay variation of data paths in real physical layouts. This is the first work to study variation on 3D paths extracted from actual graphic data system (GDS) layouts using commercial RTL-GDSII level flow and not just using a chain of gates.
- We develop a fast and accurate delay variation estimation model for the 3D data paths in digital circuits. Our model is developed with non-linear regression technique and uses input parameters from the design database already available during the place and route steps (Section III).
- We incorporate our developed model into industry quality tools and carry out variation-aware optimization for full-chip 2-tier 3D IC designs. We demonstrate up to 16% improvement in worst critical path delay under variations. To the best of our knowledge, this is the first work to carry out full chip layout-level 3D IC variation-aware design and optimization (Section IV).

We also discuss current CAD limitations involved followed by the conclusion in Section V.

II. MOTIVATION

A. State-of-the-Art for 2D IC

The on-chip variation-aware design technique for 2D ICs is well established in the industry. The basic idea in variation-aware optimization is not reducing the variation (σ) itself but shifting the entire critical path delay distribution curve with variation to below the target clock period (Fig. 2(a)). This way, the designer can ensure that the desired delay for a given critical path is satisfied by 99.7% (3σ) of design samples under consideration. In other words, the paths in the design are made much faster (compared to timing target) during design optimization in the deterministic case such that any factor slowing down the cells under process variations will still not result in more delay than the desired target for 99.7% (3σ) of samples. In practice, this optimization is not done by reducing the timing target but by intentionally derating the cell delays during timing optimization to meet the same timing target as the deterministic case. The commonly used cell derating techniques in sub-65nm digital designs are on-chip variation (OCV), advanced on-chip variation (AOCV), and parametric on-chip variation (POCV) [8]. In OCV, a single pessimistic timing derating factor is applied to all the cells making them slower simultaneously. The disadvantage of this approach is the pessimism in the timing analysis for

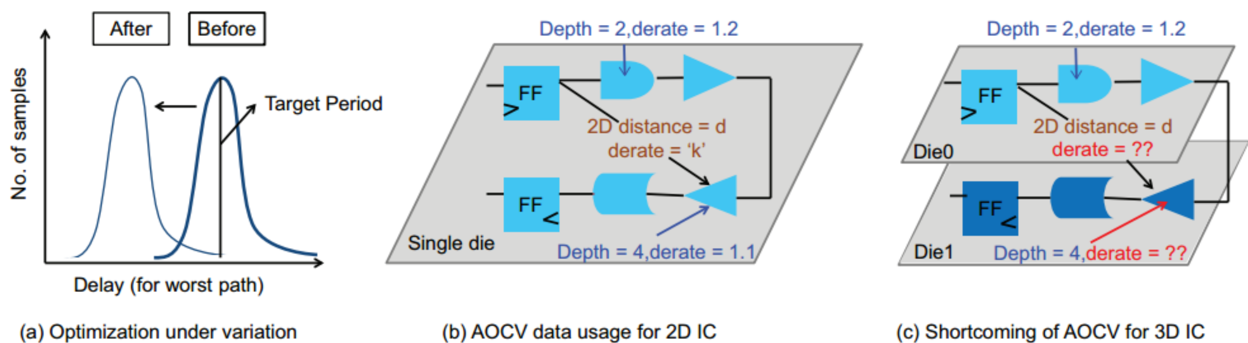

Fig. 2. State-of-the-art advanced on-chip variation (AOCV) optimization technique. (a) Delay optimization to reduce the variation impact, (b) use of AOCV data in 2D ICs, and (c) limitation of AOCV data in 3D ICs.

deep data paths, leading to additional power due to insertion of more timing buffers and larger sized cells along with longer timing closure cycles. It was acceptable for older technologies because the global timing derating factor was not very high. But with increased variation in advanced technology nodes, the common derating number tends to be very high leading to very pessimistic results.

Statistical analysis has shown that the random variation is relatively less for deeper timing paths since not all cells are simultaneously fast or slow [8]. AOCV (advanced OCV) technique is a more accurate and practical approach as it assigns timing derating factors to individual cells based on the depth and the physical span of the timing path (Fig. 2(b)). While the former takes into account the local random variation, the latter models spatial variation within the die. These derating factors are provided by the foundry in form of a look-up table which has timing derating values for all cells for different depths and distances. POCV (parametric OCV) is another method where delay variations are modeled by addition of random variables.

The AOCV and POCV techniques for 2D IC design and optimization under variation have been universally accepted as the most efficient method in terms of computational cost and design quality. In the next sub-section, we discuss the new challenges in 3D ICs and why these techniques are not directly applicable and need to be modified or expanded to be used in 3D IC design.

B. What Is New in 3D ICs?

A typical TSV based 3D IC has different dies stacked together. This introduces a new source of die-to-die variation in same design in addition to already existing within-die random and systematic variations. Garg and Marculescu [6] have studied the impact of this variation mathematically. For die-to-die variation in a lot, the method of corner analysis is generally used in 2D IC designs where each die is independent of another. However, for 3D IC, the same design has multiple dies. The die-to-die variations in the same design itself are not considered during the compilation of AOCV tables by semiconductor foundries. Moreover, the variation differences among dies in the same 3D IC will also vary from sample to sample.

Fig. 2 explains the shortcoming of currently existing AOCV approach for 3D ICs. Consider a data path as shown in Fig. 2(b) consisting of 4 cells from Q-output of a flip-flop to D-input of another. Using AOCV tables provided by foundry for that particular technology node, timing derating values are assigned to the various cells and then the design optimized. However, once the path moves from one die to another, the situation is different (Fig. 2(b)). The timing derating factors cannot be applied to the cells in a similar fashion as 2D IC and new factors need to be considered. The use of current AOCV tables in its current format will not give a practical picture of the variation and will affect the final design quality significantly. This necessitates the requirement of delay variation modeling for 3D paths. OCV approach of assigning a conservative global timing derating to all cells across all tiers is one solution, but going one step back defeats the purpose of moving into AOCV technique itself and such a pessimistic approach will not be fruitful for design quality, power and runtime.

There have been a few works on variation mitigation techniques for 3D IC. Tier adaptive body biasing has been suggested as a post silicon tuning method to reduce clock and data path variability in 3D ICs [9]. Impact of distribution of critical paths and process variation in 3D IC on clock frequency has been studied in [6]. However, use of detailed analytical models is computationally expensive especially for full-chip designs with tens of thousands of data paths. There are few other works which have studied modeling of variation for speed up and for including spatial correlation but these works are limited to 2D ICs only [10].

Fig. 3(a) shows the normalized variation curves for both 2D IC and 3D IC for a 15-inverter ring-oscillator after 1,000 Monte Carlo simulations with 5% σ/μ of input threshold voltage variation. In the 3D IC ring oscillator experiment, 7

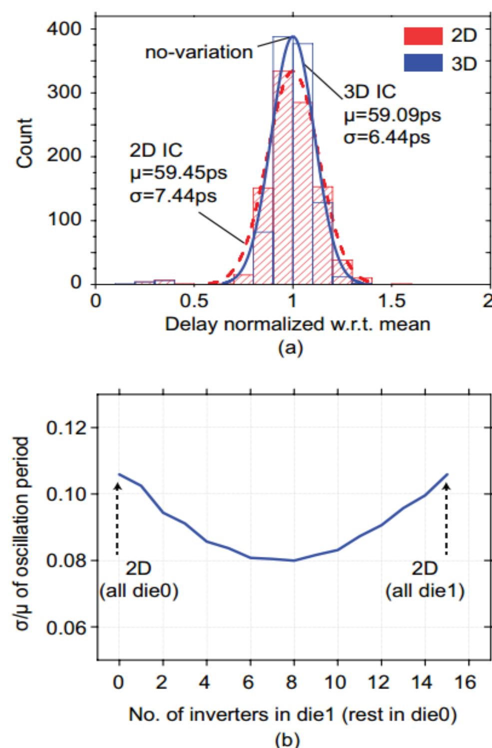

Fig. 3. Variation in 15-stage ring oscillator for 1,000 Monte Carlo simulations under 5% VT_h variation. (a) Stage-delay distribution in 2D IC and 3D IC with 7 INVs on die0 and 8 on die1 connected with 2 TSVs. (b) Relative stage-delay variation (= σ/μ) with different number of cells in die0 and die1.

inverters are in die0 and the other 8 in die1 and both within-die and die-to-die distributions are considered. The x-axis has been normalized to compare the relative variation (σ/μ). The 3D IC ring oscillator experiences lesser overall variation due to additional die-to-die variation which tend to average out over multiple dies along the path.

We also study the impact of having different number of cells distributed across two dies in a 3D IC ring oscillator (Fig. 3(b)). The corner cases include all cells in die0 and all cells in die1 which are equivalent to 2D ICs. We observe that equal number of cells distributed in each tier reduces the overall impact of variation by a good extent. The curve is also symmetric along the center (equal cells per die) which means that the count of cells in each die impacts the relative variation. Theoretically, from a statistical point of view, the addition of independent die-to-die variations are expected to have similar effect as local random variations (also independent) and hence average out overall variation.

C. Key Summary

The key points from these background studies and experiments conducted so far which motivates our overall study and establishes the foundations of variation-aware 3D IC design are:

- 3D IC variation is different from 2D IC and needs additional considerations.
- Industry approach of using AOCV (or POCV) libraries is excellent for 2D ICs but is not a practical technique, if applied directly on 3D IC design. The impact of a path spanning multiple dies needs to be considered.
- Independent die-to-die variations impact the delay variation distribution along with number of cells in each die. The delay variation impact of a fixed cell count in each die is almost the same irrespective of which die those cells are on.

III. FULL-CHIP VARIATION MODELING

A. Machine Learning in VLSI

The use of machine learning techniques has recently generated great interest with such techniques applied in design research problems [7, 11, 12]. The basic idea of machine learning is to use the actual implementation of some process or phenomenon to train or guide a model and then use it to predict the same metric for new input data. The data used for training the model initially is called the training set. To test the goodness of the developed model, it should be tested with a completely different set called the testing set. The fitness value of training set is also important to determine whether the actual set of inputs chosen is highly correlated to the output to be predicted. The key motivation of using machine learning is to replace computationally expensive simulations with predictive models which, though not 100% accurate, are orders of magnitude faster. The primary way to enhance the speed of evaluation further is to identify the key inputs which affect the output the most and reduce the total number of such inputs. Earlier works have demonstrated models of accuracy varying from 70% to 98%. In particular, for 3D IC applications, a learning-based maximum temperature model is developed and used for estimating leakage power variation impact on temperature profile of 3D Chip-Multiprocessors in [11]. Chan et al. [12] proposed a machine-learning based methodology to estimate 3D IC benefits from 2D IC implementations.

B. Modeling Methodology

In our work, we used multivariate adaptive regression splines (MARS) technique to train our model using the training set. The working of MARS is discussed in detail in [13]. We use the software tool available through [14]. MARS is a weighted combination of linear functions with knots and hinges to model non-linearity and handle higher dimensional inputs. The general approach to use MARS for modeling is to have k observations with n inputs and one target output in the training set. The set of candidate basis functions (hinge functions) based on the n inputs are determined with knots specified at the observed values. The overall process includes a forward pass to try different basis functions to reduce the training error and a backward pass to fix the overfit. For each step in forward pass, MARS adds the basis function which reduces residual error to maximum extent. The coefficient of the basis function is determined by least squares regression lines. Backward pass and generalized cross validation (GCV) is used to avoid overfit, i.e., unnecessarily high sensitivity to inputs by having too many terms in the model. This may lead to erroneous predictions for new input sets. GCV values close to zero along with low residual error values indicate a good model.

C. Design of Experiments

In order to have a good model for variation estimation, we need to have a good and extensive training set from actual data obtained through detailed Monte Carlo simulations. To achieve this, we used three different benchmark circuits and implemented them as TSV-based 2-tier 3D IC designs using 28 nm technology. These designs have both 2D and 3D data paths and therefore provide us with a large data set for training and testing. The design details of the three different benchmarks used for model development are shown in Table 1. The other three

benchmarks are used later for full chip optimization. 3D IC partitioning and TSV planning was carried out with the algorithm used in [15] and the place and route for each individual die was done using Cadence Encounter.

One of the key features of our work is the use of actual layouts to extract critical path information in the form of spice netlists and carry out detailed Monte Carlo simulations to obtain the training set. These spice netlist are extracted using Primetime from full layout design with wire parasitic information. Therefore, the actual physical parasitic of the data paths has also been considered in the form of R and C values in the detailed Monte Carlo simulations. The TSVs are 3.5 μm in diameter with resistance of 40 mΩ and capacitance of 20 fF [16]. As input variation for Monte Carlo analysis, we modeled within-die and die-to-die threshold voltage variation as independent Gaussian distribution functions with a relative standard deviation of 0.05. For a given technology node in a particular foundry process, the relative standard deviation is the same across all chips fabricated using that particular technology. Therefore, for design optimization purpose, the relative standard deviation for a given technology node in a given foundry can be a fixed value. In fact, AOCV tables are provided by foundry along with the PDK and used directly by designers during variation-aware optimization of 2D ICs. Since, we had to obtain the observations for many thousands of paths (Table 1), we limited our Monte Carlo simulations to 1,000 runs per path. The key objective is to demonstrate the modeling approach.

Our primary focus here is on obtaining a good modeling technique and methodology for variation modeling in two tier 3D ICs. Though, we only include threshold voltage variation for simulations, it is important to note that our modeling technique using MARS will capture all relevant information from the training set. Therefore, any other form of variation can be included during development of training data set and its impact will be incorporated while building the basis functions in MARS. Moreover, multi-tier 3D IC variation models can also be developed with relevant training samples covering multi-tier 3D ICs. While it is true that there needs to be more observations in training set, it is important to consider that similar to AOCV characterization, this is a one-time process during initial model development for a technology, and therefore does not incur any runtime overhead during design optimization.

Table 1. Benchmark circuits used

	Cells	Data paths	TSVs
jpeg	400.2 K	37.68 K	1668
aes	187.2 K	21.54 K	844
cf-fft	275.9 K	75.56 K	180
cf-rca	135.3 K	18.43 K	1016
des	31.5 K	1.99 K	258
itc99_b19	77.4 K	11.10 K	114

Top 3 are used for model development and bottom 3 for demonstration of full-chip variation-aware 3D IC design.

D. Input Selection for the Model

The major research contribution of any machine learning based modeling approach is to select an appropriate modeling methodology and more importantly to identify a minimal yet sufficient input set good enough to model the desired output. From the detailed Monte Carlo simulations on the thousands of paths, we generate detailed tables of information of each path along with the output standard deviation (σ) of path delay. The detailed information includes the actual deterministic path delay (mean delay), the distribution of different cell sizes (X1–X32) and the distribution of cells in terms of different complexity. To reduce the number of inputs, we categorize the complexity of cells in terms of their logic function. For example inverters and buffers fall in the simplest category, NAND, NOR and similar gates fall in the second category while flip-flops and latches fall in the most complex category. Each timing path starts from a flip-flop and ends in another. Therefore, only the size of the flip-flops are important and their count is redundant (=2). We also include the physical extent of the cells in terms of half perimeter bounding box of the timing path. This is a rough representation of the parasitic data of the entire path. It will also prove useful to capture spatial variation per die, if included in analysis. It is important to note that all of the above information is already available in the design database during the place and route process.

To identify the relevant inputs, we divided the entire data set into training set (60%), validation set (10%) during training and an independent testing set (30%). MARS modeling reports the importance of all inputs in predicting the target output. This importance metric is basically the degradation of GCV when that particular input is dropped from modeling. The higher the degradation in GCV, the more important the input is for a good model. We modeled with the different number of input variables as discussed above. In addition, we modeled relative standard deviation (σ/μ) as well as absolute standard deviation (σ) with mean path delay (μ) as an additional input to model. The validation set is used to avoid over-fitting of data during training which makes the model too sensitive to minute changes in input variables. We then tested the developed model using the testing data for accuracy evaluation. From the different models and their accuracy values, and the importance of input variable, we found out that not all the input variables are necessary to build a good and accurate

model. As discussed earlier, GCV is the key metric used for evaluating the success of training a model. The closer it is to zero (with respect to actual output range), the more accurate the model is. While the GCV values were quite low for all the iterations using different number of inputs, it was observed that too many input variables over-fitted the training samples, giving bad testing results. Also too many inputs is not good for runtime in general during full chip variation estimation. The final GCV values and average testing error % for the different models are shown in Table 2. More details on the final models and accuracy are explained below.

For a given path, the number of cells present in each die impacts the overall variation significantly. Therefore, it is important to include these two parameters as inputs to the model. It is also known that smaller cells in a given library (X1 and X2) suffer more from variation compared to larger cells. The larger cells have multiple transistors in parallel and therefore tend to average out the overall effect of variation. The other input found to be important was the half perimeter bounding box of the cell i.e. the physical extent of the timing path. The deterministic path delay was one of the most important factor to model the absolute σ values. We found that this approach of modeling absolute σ with mean delay as input was better than modeling relative variation (σ/μ). With our experiments on different set of inputs for training and then testing the model, we found that these inputs were sufficient to model the absolute standard deviation of the path delay distribution.

To summarize, the 8 inputs used for variation modeling in 2-tier 3D IC are

- Deterministic path delay
- 3D half perimeter bounding box (including constant TSV height of 30 μm).
- Total number of cells in each die (two variables).
- Number of minimum sized (X1) cells in each die (two variables).
- Number of size X2 cells in each die (two variables).
- For 2D IC variation modeling the number of input variables reduces to 5 as the different cell counts are limited to just one die.

E. Model Development and Runtime

We used MARS to generate our model from the training data set obtained with extensive Monte Carlo simulations. For all of the models including 2D IC and3D IC models, we used around 60% samples for training and validation and around 30% samples for the testing set. The final model consists of a set of max functions called basis functions, involving the input variables, certain offset values and some constants. These basis functions (hinges) are then added together with different weights and interaction levels (forming knots) to give us the final mathematical model which is very fast.

Table 2. Modeling results

	Max σ (nm)	**GCV**	**Avg. error %**
	Training	Training	Testing
3D IC	0.173	3.34×10^{-23}	10.30
2D IC	0.192	3.33×10^{-23}	7.66

Generalized cross validation (GCV) for model development using the training set and the average error % using the testing set. The maximum σ values are provided for comparison to GCV.

Since our objective is to integrate this model to the full chip optimization tool, we use Tool Command Language (Tcl) and other helping scripts to implement this mathematical formulation within the design process. There is no additional computational expense in estimating the variation using our model as the input data to our model comes from various paths and is already available in the design database (since, cell-types, count, path-length, etc., are all readily available during place and route). While an extensive Monte Carlo analysis (1,000 simulations) of a single data path takes almost 5 minutes (=300 seconds) even after using 10 parallel CPUs, our fast model estimates the variation of all the data paths (many thousands) in a given design in a fraction of a second.

If we compare the time required to develop the regression-based model itself, it is very similar to that required by foundry in building the AOCV tables where they have to carry out extensive spice simulations and test chip measurements. Therefore, our model is unique to 3D IC variation modeling and simultaneously good in terms of both development time and execution time compared to state-of-the-art 2D IC variation analysis and accurate Monte Carlo analysis respectively. For 3D ICs, there is no established variation modeling methodology since AOCV tables cannot be used directly to separate dies in same design (Section II-B).

F. Modeling Results

Fig. 4(a) shows the fitting of the model training, while Fig. 4(b) shows the prediction accuracy of the model when used for the testing set. For an ideal model, the curve would be a 45° straight line (x = y), with data points lying on that straight line. We see that the training data is very close to the ideal case. But the actual quality of a model is determined by the estimation results on the testing data and not the training data itself. A good match of testing data prediction establishes the goodness of the model. The testing data prediction is spread along a band with the center at the x=y line with an average accuracy of around 90%. Therefore, the accuracy in prediction using our developed

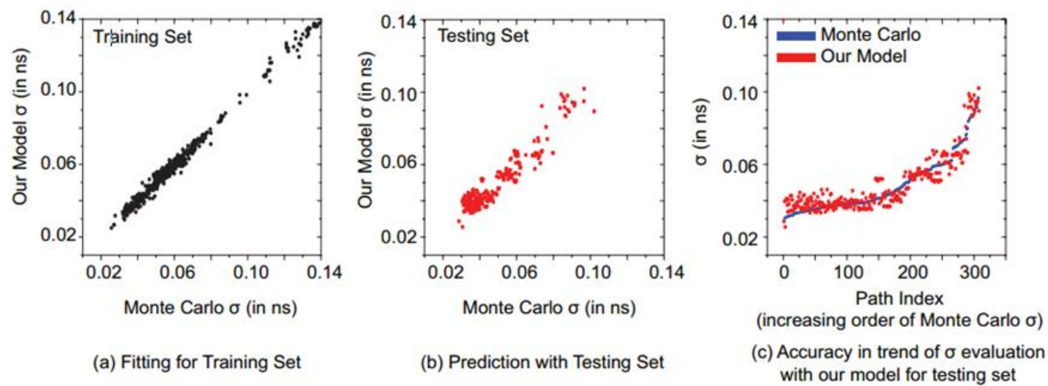

Fig. 4. 3D IC variation modeling results using our modeling technique and chosen input variables.

technique is very high. Fig. 4(c) shows the trend in prediction of σ for our model compared to the actual vales of standard deviation (σ). The samples have been arranged in increasing order of actual σ based on detailed Monte Carlo simulations. It is clear that the output of our model closely follows the Monte Carlo results which are obtained only after extensive spice simulation. The relative trend in path delay prediction is well maintained with our model.

The development of model is a major part of the design process but it is necessary to study its use and effectiveness in variation-aware 3D IC physical design. Moreover, it is important that the developed model and variation optimization approach work well for any new design case and not the designs used for initial training and development. In the following section, we demonstrate the impact of our model in successfully improving timing under statistical variation in full chip 3D IC design of benchmarks which are completely new and independent of training benchmarks.

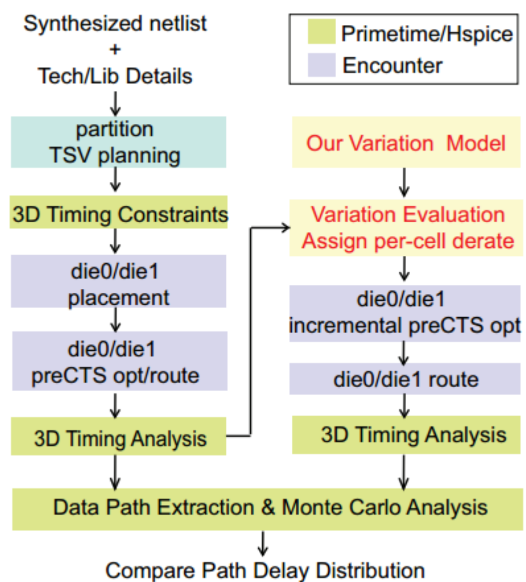

Fig. 5. Variation-aware 3D IC design optimization flow using our variation model.

IV. VARIATION-AWARE OPTIMIZATION

A. Design Methodology

We use commercial place and route tools to carry out full-chip 3D IC variation-aware design. The detailed design flow is shown in Fig. 5. We start from a synthesized netlist and library and carry out partitioning and TSV planning using technique used in [15]. This is followed by a detailed 3D timing analysis including 3D parasitics to determine the timing constraints and boundary conditions for each die under the given overall timing constraints for the design. After detailed placement, optimization and routing for both dies, we carry out 3D timing analysis using Primetime. Primetime is used for more accurate 3D timing analysis including the TSVs parasitic information and the 3D connections.

We assume ideal clock and ignore Clock Tree Synthesis in our current work as clock path variation is another important study in itself. Based on the deterministic timing results after static timing analysis, we use our model to evaluate σ of all the timing paths. We then evaluate the ratio $(\mu + 3\sigma)/t$ where μ is the delay of the path, σ is the estimated standard deviation using our model and t is the target clock period of the design under consideration. If this ratio is greater than 1, it violates the timing requirement under statistical variations. While calculating this ratio, we use the actual timing target of design and not delay of respective paths. This helps us to avoid over optimization of paths which do not violate the timing constraint under variations. Then we assign timing derating factors equal to $1 + 3\sigma/\mu$ for each cell in the violating path. If a cell is a part of multiple paths, the worst derating calculated for that cell is used. Our

Table 3. Full-chip variation-aware optimization for 2-tier 3D ICs

	Cells (×1000)					Wirelength (m)	Worst path delay (ns) $(\mu+3\sigma)_{worst}$	Power-delay-product (pJ)		
	X1	X2	X4-X8	X16-X32	Total			Dynamic	Leakage	Total
cf_rca										
Default	47.9	73.0	13.6	0	134.6	0.888	0.9445	28.86	1.82	30.68
Optimized	38.9	73.0	23.5	0.062	135.5	0.897	0.8286 (-12%)	28.99	1.82	30.81
des										
Default	12.5	13.2	4.4	0	30.1	0.165	0.8341	10.51	0.34	10.85
Optimized	5.9	12.9	11.3	0.132	30.2	0.166	0.7886 (-6%)	10.52	0.35	10.87
Itc99_b19										
Default	20.0	52.4	4.9	0.001	77.2	0.165	3.45	15.18	3.56	18.74
Optimized	17.7	52.3	7.3	0.078	77.4	0.166	2.91 (-16%)	15.24	3.59	18.83

model uses input parameters which are already known to the place and route tool (Encounter) and therefore has no runtime overhead for extracting information to predict the σ values. After assigning the derating values to the cells, we carry out incremental optimization on the design followed by detailed routing. Incremental optimization is much faster because it optimizes the newly violated paths only. Then we extract the spice netlist with parasitic information for the worst 100 critical paths of the design and carry out detailed Monte Carlo analysis to compare the efficiency of optimization using our approach.

B. 3D IC Design Optimization Results

Table 3 shows the details of the optimization results for three benchmarks implemented as 2-tier 3D IC. These benchmarks are completely different from the ones used for developing the model and therefore give more credibility to our modeling technique in evaluating variation and optimizing the design under variation. We use power delay product (PDP) for fair comparison of power at the respective critical path delays. In general, designs at nominal operating conditions run fast and have lower relative variation (around 10%). Our optimization approach improves the worst path delay under variation by 12% for the cf_rca benchmark, 6% for the des benchmark and 16% for the itc99_b19 benchmark. The table also shows the change in count of different sizes of cells and the overall PDP change due to additional timing optimization carried out on the data paths.

Fig. 6 shows the effectiveness of our algorithm in improving the timing of most critical paths of the cf_rca design. The paths are arranged in decreasing order of worst delay under variation and the corresponding red dots show the worst delay under variation after full-chip optimization. The worst paths are optimized more due to higher derating factor assignment, therefore, further proving the effectiveness of our model in estimating the variation with sufficient accuracy. These values are compared after Monte Carlo simulations. Fig. 7 shows the distribution of top critical paths in the cf_rca design before and after optimization in layout form. Red paths indicate longer delay values while blue is for low delay values under variation. It is clear from the figure that the $(\mu + 3\sigma)$ of all the paths are reduced after adopting our model for full chip optimization.

Our optimization (similar to AOCV optimization) tries to optimize the timing critical paths to meet stricter timing constraints and does not target reduction in actual variation (σ). Since the tool prioritizes optimization of the most critical paths based on the assigned derating factors, some of the less critical paths may be affected in terms of increase in variation due to use of smaller sized cells. These paths may still satisfy the timing constraint with the low derating factors assigned to them but may end up having larger delay spread than before. This issue is easily fixed by using a few iterations of incremental optimization as the second iteration will give higher optimization priority to these paths.

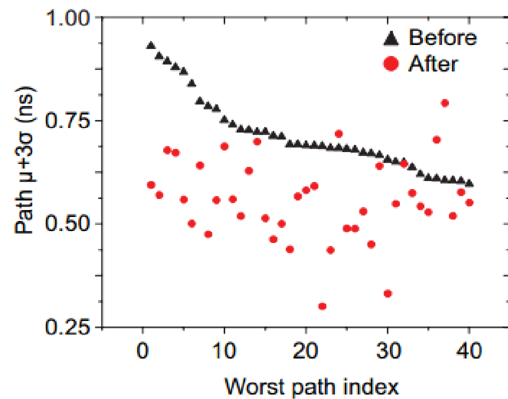

Fig. 6. Worst μ + 3σ values before and after the optimization for cf_rca benchmark (data for 40 worst paths, see Fig. 7).

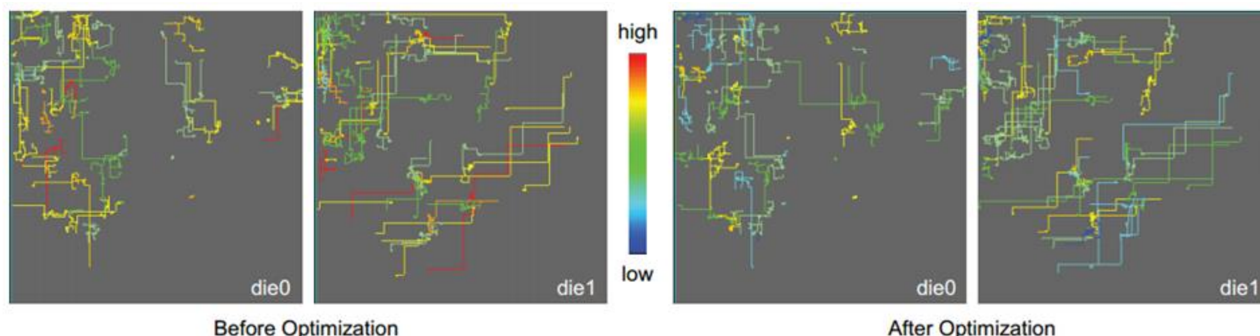

Fig. 7. The ($\mu + 3\sigma$) reduction on critical data paths using our optimization for 3D IC implementation of cf_rca benchmark.

C. Advantages and CAD Limitations

We demonstrated that our model is quite effective in reducing the worst path delay under variations for actual GDSII level 3D IC designs without any power overhead. Tightening clock constraints will help in meeting timing under variation, but the penalty in terms of buffers and cell up-sizing will be huge. AOCV tables, in their current format, are not suitable for 3D ICs. Our model is developed using machine learning based regression techniques, and is a very simple addition of various basis functions involving the different input parameters. Therefore, it is extremely fast and is easily integrated into the design tools themselves. The accuracy, though not 100%, is quite reasonable and assigning path specific derating factors helps us avoid over design and saves additional power dissipation. To be able to get more accuracy, we need to have much larger database for training which, as a matter of fact, is available to the foundries since they use this same database for building AOCV tables.

One important limitation in our work is that we could not modify the internal algorithms of the commercial tools to include our model inside their timing optimization engine and therefore, had to go back and forth outside of the optimization process to evaluate delay variation and then carry out incremental optimization. The use of commercial-grade tools is important to maintain design quality and runtime. Moreover, there are no direct 3D IC place and route tools of commercial quality and our resources were limited in that respect. However, given the fact that 3D IC is gaining more momentum, it is imperative to develop CAD tools for robust and reliable 3D IC designs and we believe that our machine learning technique of developing variation estimation models is a good contribution in this process.

V. CONCLUSION

We studied the new issues and challenges of variation modeling in 3D IC designs and developed a fast variation estimation model using non-linear regression. Variation is a major concern in modern technology nodes and needs to be addressed for all kinds of design techniques including 3D ICs, where the problem is unique. During model development, we discussed the various factors which impact path delay variations in real GDSII level circuits. Our modeling technique is applicable to both 2D IC and 3D IC and is unique to 3D paths spanning multiple dies. We demonstrated the accuracy of our model using completely independent set of sample points for testing. Moreover, we demonstrated the effectiveness of our developed model on completely different benchmarks. By assigning cell specific timing derating factors based on variation estimation with our model, we achieved significant improvement in the critical path delay distribution with minimal power overhead. Our future work is focused on variation modeling for highly sensitive low voltage circuits and the exploration of other machine learning models to further reduce the error in variation estimation.

REFERENCES

[1] D. Fick, R. G. Dreslinski, B. Giridhar, G. Kim, S. Seo, M. Fojtik, et al., "Centip3De: a cluster-based NTC architecture with 64 ARM Cortex-M3 cores in 3D stacked 130 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 104–117, 2013.

[2] D. H. Kim, K. Athikulwongse, M. Healy, M. Hossain, M. Jung, I. Khorosh, et al. "3D-MAPS: 3D massively parallel processor with stacked memory," in *Proceedings of 2012 IEEE International Solid-State Circuits Conference*, San Francisco, CA, pp. 188–190, 2012.

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

[4] P. Falkenstern, Y. Xie, Y. W. Chang, and Y. Wang, "Three-dimensional integrated circuits (3D IC) floorplan and power/ground network co-synthesis," in *Proceedings of the 2010 15th Asia and South Pacific Design Automation Conference (ASP-*

DAC), Taipei, Taiwan, pp. 169–174, 2010.

[5] D. C. Juan, S. Garg, and D. Marculescu, "Impact of manufacturing process variations on performance and thermal characteristics of 3D ICs: Emerging challenges and new solutions," in *Proceedings of 2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, Beijing, China, pp. 541–544, 2013.

[6] S. Garg and D. Marculescu, "3D-GCP: an analytical model for the impact of process variations on the critical path delay distribution of 3D ICs," in *Proceedings of 2009 10th International Symposium on Quality Electronic Design*, San Jose, CA, pp. 147–155, 2009.

[7] A. B. Kahng, B. Lin, and K. Samadi, "Improved on-chip router analytical power and area modeling," in *Proceedings of 2010 15th Asia and South Pacific Design Automation Conference (ASP-DAC)*, Taipei, Taiwan, pp. 241–246, 2010.

[8] S. Walia, *PrimeTime Advanced OCV Technology*. Mountain View, CA: Synopsys Inc., 2009.

[9] K. Chae, X. Zhao, S. K. Lim, and S. Mukhopadhyay, "Tier adaptive body biasing: a post-silicon tuning method to minimize clock skew variations in 3-D ICs," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 3, no. 10, pp. 1720–1730, 2013.

[10] H. D. H. Qian, S. S. Sapatnekar, and K. Bazargan, "Fast and accurate statistical criticality computation under process variations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 28, no. 3, pp. 350–363, 2009.

[11] 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 *Proceedings of 2011 Design, Automation Test in Europe*, Grenoble, France, pp. 1–6, 2011.

[12] W. T. J. Chan, S. Nath, A. B. Kahng, Y. Du, and K. Samadi, "3DIC benefit estimation and implementation guidance from 2DIC implementation," in *Proceedings of 2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, San Francisco, CA, pp. 1–6, 2015.

[13] J. H. Friedman, "Multivariate adaptive regression splines," *Annals of Statistics*, vol. 19, no. 1, pp. 1-67, 1991.

[14] Salford Systems [Internet]. Available: http://www.salford-systems.com/products/mars.

[15] D. H. Kim, K. Athikulwongse, and S. K. Lim, "Study of through-silicon-via impact on the 3-D stacked IC layout," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 5, pp. 862–874, 2013.

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

Sandeep Kumar Samal
is a PhD student in the School of Electrical and Computer Engineering at Georgia Institute of Technology. He received the B.Tech. degree in electronics and electrical communication engineering from 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. His current research interests include low power and reliable digital design, modeling, and analysis using through-silicon-via-based and monolithic 3D IC technology.

Guoqing Chen
received the B.S. (with honors) and M.S. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1998 and 2001, respectively, and Ph.D. in electrical engineering from the University of Rochester, Rochester, NY, USA in 2007. His research interests include low-power circuits and architectures, clock and power distribution networks, on-chip interconnects, and 3D ICs.

Sung Kyu Lim
received the B.S., M.S., and Ph.D. degrees from the Computer Science Department, University of California, Los Angeles (UCLA), in 1994, 1997, and 2000, respectively. He is a full Professor at the School of Electrical and Computer Engineering, Georgia Institute of Technology. His research focus is on the architecture, design, test, and EDA solutions for 3D ICs. His research on 3D IC reliability is featured as Research Highlight in the Communication of the ACM in 2014. Dr. Lim received the National Science Foundation Faculty Early Career Development (CAREER) Award in 2006. His work is nominated for the Best Paper Award at ISPD'06, ICCAD'09, CICC'10, DAC'11, DAC'12, ISLPED'12, and DAC'14. He is an Associate Editor of the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.