Statistical Thermal Evaluation and Mitigation Techniques for 3D Chip-Multiprocessors In the Presence of Process Variations

Da-Cheng Juan
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, U.S.A.
dacheng@cmu.edu

Siddharth Garg
Electrical and Computer Engineering
University of Waterloo
Ontario, Canada
siddharth.garg@uwaterloo.ca

Diana Marculescu
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, U.S.A.
dianam@cmu.edu

***Abstract*—Thermal issues have become critical roadblocks for achieving highly reliable three-dimensional (3D) integrated circuits. This paper performs both the evaluation and mitigation of the impact of leakage power variations on the temperature profile of 3D Chip-Multiprocessors (CMPs). Furthermore, this paper provides a learning-based model to predict the maximum temperature, based on which a simple, yet effective tier-stacking algorithm to mitigate the impact of variations on the temperature profile of 3D CMPs is proposed. Results show that (1) the proposed prediction model achieves more than 98% accuracy, (2) a 4-tier 3D implementation can be more than 40°C hotter than its 2D counterpart and (3) the proposed tier-stacking algorithm significantly improves the thermal yield from 44.4% to 81.1% for a 3D CMP.**

Keywords-thermal; leakage; process variation; 3D; stack; yield; chip-multiprocessor; statistical learning; regression

I. INTRODUCTION

With increased technology scaling, wire length has become a critical factor that limits the performance of integrated circuits. Recently, three-dimensional (3D) integrated circuits (ICs) have been proposed as one of the most promising methodologies to overcome this barrier [1][2][3]. In a 3D IC, conventionally fabricated planar dies are stacked on top of each other and connected using through-silicon vias (TSVs), resulting in lower communication latency. However, due to higher power density and lower thermal conductivity of inter-tier dielectrics [4][5], the thermal concerns for 3D ICs are exacerbated. In particular, the tiers further away from the heat sink tend to have elevated temperature profiles, often up to 25˚C higher than tiers closest to the heat sink [3]. Such a high operating temperature may reduce the mean time to failure and speed up the aging of a 3D IC.

Another concern introduced by technology scaling is the increased contribution of leakage power dissipation to total power consumption. In addition, higher temperatures result in increased leakage power dissipation, because leakage power has exponential dependency on temperature [6]. To make matters worse, the increased leakage power leads to higher total power consumption, which in turn generates more heat and further increases the temperature. This interdependency between temperature and leakage power forms a feedback loop, which in the worst case, may lead to thermal runaway. This phenomenon necessitates the accurate modeling of the interplay between leakage and temperature for 3D ICs, in order to ensure that the operating temperature of the 3D system lies within the maximum temperature constraint.

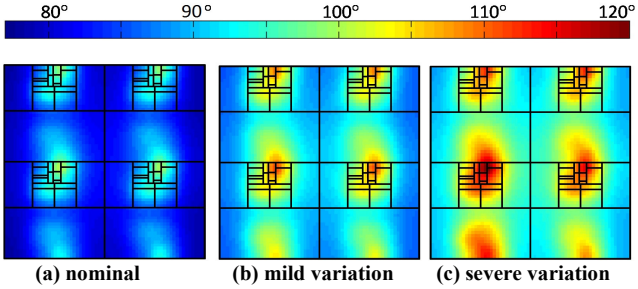


Figure 1: Temperature maps of top tier in a 3D CMP.

In this context, it is critical to also accurately model the impact of process-induced leakage variability on the temperature profile of a 3D IC. Due to the exponential dependency of leakage power dissipation on process parameters, more than one order of magnitude difference can exist in the leakage power profile from one die to another [7]. Figure 1 depicts three temperature maps of the tier farthest away from the heat sink in a 3D Chip-Multiprocessor (CMP) in the absence (a) or presence (b)(c) of process variations. The hottest point and the average temperature in Figure 1(c) are approximately 19.7˚C higher than the corresponding ones in Figure 1(a). This shows that leakage variations may significantly increase the temperature of a 3D system. To be able to gauge the impact of process variations at the system level in 3D ICs, a comprehensive framework is needed for variation-aware thermal modeling.

A. *PREVIOUS WORK*

Thermal modeling for both conventional planar and 3D ICs has received a lot of attention in the research community. Skadron et al. [8] proposed Hotspot – an accurate temperature model for planar ICs and later extended it to account for 3D circuits as well. From a mitigation perspective, Donald et al. [9] used dynamic voltage and frequency scaling (DVFS) to avoid thermal emergencies in chip-multiprocessors (CMPs). Ebi et al. [10] presented an agent-based power distribution approach to balance the power consumption of CMPs in a pro-active manner. Goplen et al. [11] and Cong et al. [12] proposed novel placement algorithms to reduce the temperatures in 3D systems. Chakraborty et al. [7] proposed a method to assign threshold voltages for 3D CMPs. Very recently, Zhuo et al. [13] presented a workload-aware framework that accounts for local variations in both the process and temperature.

The impact of process variations on the performance of 3D ICs has been recently addressed by Ozdemir et al. [14], Garg et al. [15] and Ferri et al. [16]. However, all these papers consider the impact of variations on the timing characteristics of 3D ICs, and not on leakage power or temperature.

B. *PAPER CONTRIBUTIONS*

To the best of our knowledge, previous research has not addressed the impact of leakage variations on the temperature profile of 3D ICs from neither an analysis nor a mitigation perspective. In this work, we address this issue for the first time, using a 2-tier and a 4-tier 3D implementation of a 16-core CMP as a case study. The experimental results confirm that the temperature profile of 3D CMPs shows much larger susceptibility to leakage variations when compared to an equivalent 2D implementation. In particular, using nominal leakage values to determine the maximum temperature for a 3D CMP can

This research was supported in part by NSF Grant CCR0702451

978-3-9810801-7-9/DATE11/©2011 EDAA

severely underestimate the actual maximum temperature observed by a large fraction of 3D systems.

Motivated by the evaluation results, we make two contributions to help 3D CMP designers mitigate the impact of leakage variations on the maximum operating temperature. First, using a learning-based regression model, we show that the maximum steady-state temperature of a 3D CMP can be accurately predicted as a linear function of total leakage power dissipation of each tier in a 3D system. The proposed learning model can be used to make quick and accurate post-silicon predictions of the maximum temperature for each fabricated 3D system and, as we will show, to enable thermal hotspot mitigation strategies. Our model shows less than 2% error and more than 30X speed-up when compared to actual temperature simulations.

Second, in order to mitigate the impact of leakage variations on the temperature profile of a 3D system, we propose a simple yet effective tier-stacking algorithm that can be used for any symmetrically-partitioned 3D system, such as the 3D CMPs studied in this paper. Based on the proposed temperature modeling methodology, our algorithm is able to select the optimal tier-stacking order to minimize overall temperature. The experimental results show that the proposed method significantly reduces the standard deviation of the maximum temperature distribution by 54%. In addition, for a 100˚C temperature constraint [31], the proposed stacking technique increases the *thermal yield* from 44.4% to 81.1%.

The remainder of this paper is organized as follows. Section II introduces the related background knowledge required for our work. Section III details the proposed temperature prediction and tier stacking algorithm. Section IV presents the proposed implementation flow. Section V demonstrates the experimental results and Section VI concludes this paper.

II. BACKGROUND

In this section, we introduce the background knowledge relevant to our proposed methodology, including thermal modeling and leakage current characterization. We also discuss how these models are modified to account for the impact of process variation on 3D systems.

A. *THERMAL MODEL*

This section presents the thermal model used throughout the paper. Conventionally, heat flow is approximated as a heat current flowing through thermal resistance [17], resulting in temperature differences. This phenomenon can be modeled as the electrical current in an RC network, and the temperature differences can be expressed as:

$$\boldsymbol{C}\frac{d\overrightarrow{\boldsymbol{T}}(t)}{dt} = \boldsymbol{R^{-1}}\overrightarrow{\boldsymbol{T}}(t) - \overrightarrow{p}U(t) \tag{1}$$

where $\boldsymbol{C}$ is a diagonal thermal capacitance matrix, $\boldsymbol{R}$ is a thermal resistance matrix, $\overrightarrow{\boldsymbol{T}}(t) = (T_1 - T_A, ..., T_n - T_A)^T$ is the temperature vector and T_A is the ambient temperature, $\overrightarrow{p} = (p_1, ..., p_n)^T$ is the power vector, and $U(t)$ is a step function.

For the purpose of steady-state thermal analysis the temperature does not vary with time. Therefore, Eq.(1) can be modified as follows:

$$\overrightarrow{p} = \boldsymbol{R^{-1}}\overrightarrow{\boldsymbol{T}}(t) \implies \overrightarrow{\boldsymbol{T}}(t) = \boldsymbol{R}\overrightarrow{p} \tag{2}$$

From Eq.(1) and Eq.(2), the power vector $\overrightarrow{p}$ is proportional to both transient-state and steady-state temperatures. In other words, the increase of power consumption directly affects temperature if other factors remain fixed.

B. *LEAKAGE CURRENT MODEL*

The power consumption of a circuit consists of active power and leakage power. This section focuses on modeling the feedback loop between leakage and temperature, since the active power is not sensitive to the temperature. Leakage power contains several components, among which sub-threshold leakage current and gate leakage current are main contributors [18]. Recently, due to the introduction of high-k dielectrics, the gate leakage component has become less important. We therefore concentrate only on sub-threshold leakage power dissipation. Eq.(3) describes the leakage current model for a single transistor. For simplicity and without losing accuracy, terms not sensitive to temperature or effective channel length are merged together:

$$\begin{aligned} I_{leak} &= K\frac{W}{L_{eff}}(\frac{kT}{q})^2 e^{\frac{q(-V_{th})}{nkT}} \\ &= K'\frac{W}{L_{eff}}(T)^2 e^{-\frac{c\cdot V_{th}}{T}} \end{aligned} \tag{3}$$

where K' is a technology dependent constant, W is transistor width, L_{eff} is effective channel length, T is temperature, V_{th} is threshold voltage, and c is a positive constant. According to Eq.(3), I_{leak} will increase when L_{eff} decreases and T increases. Combining Eq.(1)(2)(3) determines the interdependency of temperature and leakage power. The increase of either temperature or leakage power will trigger this positive feedback loop.

It is worth mentioning that V_{th} also exponentially depends upon L_{eff} due to drain induced barrier lowering (DIBL). Eq.(4) models the relationship between V_{th} and L_{eff}:

$$V_{th} = V_{th0} - V_{dd}\cdot e^{-\alpha_{DIBL}\cdot L_{eff}} \tag{4}$$

where V_{th0} is the threshold voltage for long channel transistors, α_{DIBL} is the DIBL coefficient, and V_{dd} is the supply voltage. From Eq.(3)(4), it is clear that the decrease of L_{eff} determines both exponential and linear scaling factors for leakage current. As a result, the L_{eff} variation usually increases leakage power exponentially. Although there are also other factors affecting leakage current, such as gate oxide thickness and doping density variations, this paper mainly focuses on the L_{eff} variation because of the aforementioned exponential dependency.

C. *PROCESS VARIATION MODEL*

Process variations affect several important metrics of an IC, such as power consumption and maximum clock frequency. Variations may also affect leakage current, which in turn increases the temperature due to the interdependency between temperature and leakage power. This section presents the variation model used in this paper.

In general, the variation of L_{eff} in 3D systems can be described as:

$$L_{eff} = L_{nom} \pm \Delta_{total} \tag{5}$$

where L_{nom} is the nominal value of L_{eff}, and Δ_{total} is the total variation of L_{eff}. Let us further merge Eq.(4)(5) into Eq.(3) to demonstrate how L_{eff} variations affect leakage currents:

$$I_{leak} = \frac{K''}{(L_{nom} \pm \Delta_{total})}(T)^2 e^{-\frac{C_1 - C_2\cdot e^{-\alpha\cdot(L_{nom}\pm\Delta_{total})}}{T}} \tag{6}$$

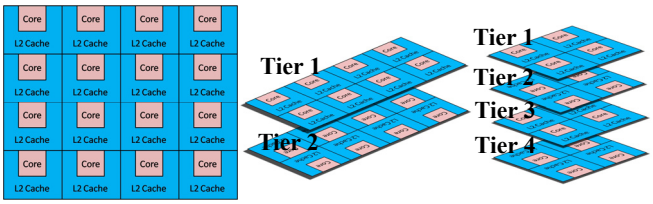
where K'' is a technology dependent constant, C_1, and C_2 are both positive constants. From Eq.(6), it is clear that if Δ_{total} increases, I_{leak} may see a significant increase due to the exponentially and inverse-linearly dependent terms, respectively. Furthermore, Δ_{total} can be decomposed as:

$$\Delta_{total} = w_1\Delta_{w2w} + w_2\Delta_{spat} + w_3\Delta_{rand} \tag{7}$$

where Δ_{w2w} is the wafer-to-wafer (W2W) variation, Δ_{spat} is the die-to-die (D2D) spatial variation, Δ_{rand} is the within-die (WID) random variation, and w_is are the corresponding weights. Δ_{w2w} and Δ_{rand} can be modeled as Gaussian random variables. In [19], Cheng et al. proposed an accurate, deterministic model of Δ_{spat} by exploiting across-wafer variation. In this paper, we use this model assuming Δ_{total} of 5% of L_{nom}; furthermore, based on [20], the relative ratio among Δ_{w2w}, Δ_{spat} and Δ_{rand} is set to 0.7:1:1.

Parameters	Values
Number of cores	16
Frequency	3.0 GHz
Technology	45nm node with V_{dd}=1.0V
On-chip network	4×4 mesh
L1- I/D caches	64KB, 64B blocks, 2-way SA, LRU
L2 caches	1MB, 64B blocks, 16-way SA, LRU
Pipeline	7 stage deeps, 4 instructions wide

Table 1. Processor parameters



(a) 1-tier CMP **(b) 2-tier CMP** **(c) 4-tier CMP**

Figure 2: 2D and 3D CMP implementation.

III. METHODOLOGY

In this section, we first introduce symmetric CMPs, the architecture used in this paper. In Section III.B, we propose a methodology to estimate the maximum steady-state temperature of a 3D CMP. Also, the mathematical formulation as well as the accuracy of the proposed method is included. In Section III.C, we present the algorithm to determine the order of tier stacking for 3D CMPs by using the proposed methodology of temperature estimation.

A. *TARGET ARCHITECTURE*

The architecture used throughout this paper is a symmetric CMP, which consists of 16 out-of-order, Alpha 21264 cores [21]. The corresponding micro-architecture parameters are listed in Table 1. The floorplan of a single Alpha 21264 processor taken from [8] is replicated 16 times in a 4×4 mesh to create a planar 2D CMP. As shown in Figure 2(a), processing cores and caches are placed in a fine-grained, interwoven manner. The floorplans for the corresponding 2-tier and 4-tier CMPs are shown in Figure 2(b) and (c), respectively. Note that the floorplan of every other tier is flipped to ensure that cores are never stacked directly on top of each other [23]. For the 3D CMPs, we assume that tier 1 is the farthest away from the heat sink, while tier 4 is the closest to the heat sink. In addition, we point out that in this paper, the focus is on thermal evaluation as opposed to performance, which has been extensively addressed by [14][15][16]. Therefore, the detailed performance comparison between 2D and 3D CMPs is out of the scope of this paper and not addressed here.

B. *LEARNING-BASED MODEL FOR TEMPERATURE PREDICTION*

As shown in Figure 1 in Section I, the maximum temperature of a 3D CMP under the impact of process variations can be significantly higher than the temperature obtained under using nominal leakage power conditions. In addition to exceeding imposed thermal constraints, the elevated temperature could also lead to dramatically reduced reliability for 3D systems. Therefore, it is crucial to develop thermal modeling and mitigation methodologies which account for the impact of leakage variations.

In the presence of leakage power variations, the maximum temperature for each 3D system is not known before fabrication, and hence a post-silicon temperature prediction mechanism is required. This information can help engineers filter the 3D systems that exceed the thermal constraint during the testing process, or be a useful input to post-fabrication thermal management strategies. Using Hotspot [8] or other simulation-based methods to determine the maximum temperature for each fabricated 3D system can be too time-consuming. We found that, under steady-state conditions, the maximum temperature-leakage curve can be approximated very well by a linear dependency function for most workloads, although analytically, the instantaneous leakage power depends exponentially on the operating temperature as shown in Eq. (3)(6). This observation motivates the development of a learning-based model to predict the maximum temperature of a 3D system based on leakage measurements. We aim to exploit per-tier leakage current measurements for predicting the maximum temperature of each fabricated 3D system. Note that these leakage measurements are routinely performed on bare dies before packaging as part of the popular IDDQ testing methodology [22], therefore we do not introduce any additional test costs for maxium temperature prediction.

Based on the aforementioned observation, we propose a learning-based regression model, which expresses the maximum temperature for the 3D system as a linear function of per-tier leakage power values under steady-state conditions:

$$T^{max} = \sum_{i=1}^{n} a_i P_{leak}^{i} + c \tag{8}$$

where T^{max} is the maximum temperature of a 3D CMPs under the steady-state condition, a_i s are fitting coefficients, c is a fitting constant, P_{leak}^{i} is the total leakage power of tier i, and n is the total number of tiers. For the purpose of experimental results presented in this paper, n is set to four. However, the framework is general and can be used for an arbitrary number of tiers. The physical meaning of a_is can be interpreted as the sensitivity of the maximum temperature to the leakage power of the i^{th} tier. In addition, note that by convention, ti's are annotated according to the distance from the heat sink, i.e., $t1$ is the tier furthest away from the heat sink, whereas $t4$ is the tier closest to the sink.

Here, we separate the coeffcient-learning process into two phases: *training phase* and *testing phase*. The goal of the training phase is to learn the fitting coefficients $\hat{a}_i$s and constant $\hat{c}$, where $\hat{a}_i$s and $\hat{c}$ are the estimates of a_is and c. This can be done by minimizing the least square loss function:

$$(\hat{a}_i, \hat{c}) = argmin\left\{ \sum_{k=1}^{m} \left(T_k^{max} - \sum_{i=1}^{n} a_i P_{leak}^{ik} - c \right)^2 \right\} \tag{9}$$

where m is the size of training set, i.e. the number of 3D CMPs whose T_k^{max} are known. In this paper, m is empirically set to 500. Please note that T_k^{max} can be obtained via various methods, including readings from thermal sensors or Hotspot simulation. In Eq.(9), T_k^{max} and P_{leak}^{ik} are fed as inputs, while $\hat{a}_i$s and $\hat{c}$ are the outputs of the training phase. Note that $\hat{a}_i$s are not fixed for all kinds of CMP designs. Depending on different technology nodes, design specs, layouts and other factors, $\hat{a}_i$ s may need to be re-learnt by re-evaluating Eq. (9) with the corresponding T_k^{max} and P_{leak}^{ik}. The accuracy of the proposed model can be further improved if more detailed measurements are available at test time, such as per-core or even per-component leakage power. These measurements can be included in Eq.(8) (9) as extra features to improve the accuracy.

In the testing phase, $\hat{a}_i$ values determined from the training phase are plugged into Eq.(8) to calculate $\hat{T}^{max}$ as an estimate of T^{max}:

$$\hat{T}^{max} = \sum_{i=1}^{n} \hat{a}_i P_{leak}^{i} + \hat{c} \tag{10}$$

By using Eq.(10), $\hat{T}^{max}$ can be calculated if the per-tier leakage measurement P_{leak}^{i} is given. **No time-consuming thermal simulation is required in this phase**. Please note that Eq.(10) is different from Eq.(8) because $\hat{a}_i$ and $\hat{T}^{max}$ of Eq.(10) are estimates but a_i and T^{max} of Eq.(8) are actual values.

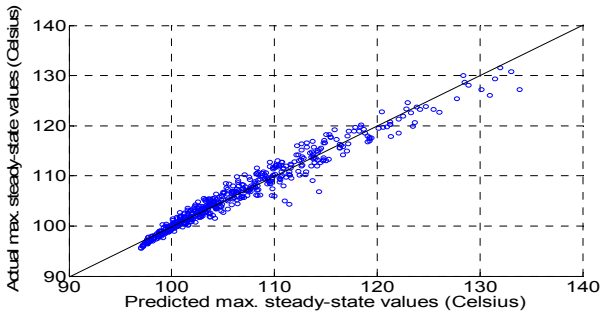


Figure 3: The scatter plot of estimated values vs. actual maximum temperature under the steady-state condition for a 16-core, 4-tier 3D CMP.

To evaluate the accuracy of the proposed learning model, we use 10-fold cross-validation [24] to calculate the prediction error. Cross-Validation is a nearly-unbiased error estimator and is widely-used in machine learning and statistics fields. Figure 3 shows the cross-validated results; the X axis stands for the predicted temperatures by using the learning-based model whereas the Y axis represents the actual simulated results obtained with Hotspot. It is clear that our estimation of maximum temperatures is very accurate. The correlation coefficient is 0.9797; the prediction error rate is 1.02%. Therefore, just relying on per-tier leakage power values obtained at test time, one can determine with high accuracy the maximum temperature for a 3D system, without actually integrating the tiers and performing full system testing. Note that since the active power is not sensitive to the change of temperature as mentioned in Section II.B, it is not included explicitly in the model, but may implicitly be modeled as part of the constant term. We would like to stress that this learning model does not aim to replace the original thermal model described in Eq.(1)(2) or thermal simulators like Hotspot. The model relies on accurate temperature analysis or simulation to provide inputs to learn the fitting coefficients. The goal of the proposed model is to allow for fast, compared to Hotspot, post-fabrication estimation of the maximum temperature of a 3D stack given the leakage power measurements of its constituent tiers. As we will show, this information can be used to determine an optimal tier-stacking order for symmetric 3D CMPs that minimizes maximum temperature.

As an example, we list the values of $\hat{a}_i$s for a 16-core 4-tier CMP in Table 2. The values of $\hat{a}_i$s would be expected to be monotonically decreasing according to the tier ordering, i.e., $\hat{a}_1 > \hat{a}_2 > \hat{a}_3 > \hat{a}_4$, since the impact of tiers further away from the heat sink on the maximum temperature is expected to be higher. As expected, $\hat{a}_1$ has the largest value since it represents the weight for the leakage power of the top tier in the 3D CMP. Interestingly, the other three coefficients are not monotonically decreasing; instead, $\hat{a}_3$ is larger than either $\hat{a}_2$ or $\hat{a}_4$. The reason for this non-monotonic behavior stems in the relative positioning of processing cores: in tier 3, the cores are located directly underneath the cores of the hottest tier, i.e., tier 1. Therefore, the vertical heat conduction of cores between tier 3 and tier 1 increases the impact of tier 3 leakage on the maximum temperature. On the other hand, the relative contribution of tier 2 is reduced since L2 caches are placed directly underneath the cores of tier 1. L2 caches tend to run cooler than cores since they have lower dynamic power dissipation. In addition, the channel lengths of L2 caches are increased slightly to reduce the leakage power dissipation [25], thereby making them more robust to leakage variations.

$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$
2.052	0.7182	1.2312	0.5643

Table 2: The values of $\hat{a}_i$

C. *TIER STACKING*

As observed from the $\hat{a}_i$'s of the learning model in Section III.B, the leakage value of each tier has a different impact on the maximum temperature – for example, the coefficient of tier 1 is almost four times greater than the corresponding coefficient of tier 4. This observation raises an intriguing possibility: is it possible to re-stack the tiers based on their leakage values, so as to keep the tiers with high leakage values closer to the heat sink, and therefore achieve a potential reduction in the maximum temperature? As a result, this stacking technique would only be applicable for *symmetric* 3D systems, i.e., each tier has the same layout. This is certainly the case for the 3D CMP system we study in this paper. Also, the stacking technique would be applicable to the systems that contain multiple identical stacked SRAM or DRAM layers, or to the recently introduced Reciprocal Design Symmetry (RDS) based 3D ICs [23].

The central idea of this algorithm is to let $\hat{a}_i$s guide how to determine stacking orders. Recall that $\hat{a}_i$s represent the sensitivity of the maximum temperature to the leakage power of the i^{th} tier, thereby providing a powerful clue of how to assign CMPs of different leakages to the most suitable tiers. In other words, the CMP with the largest leakage power should be placed on the tier with smallest $\hat{a}_i$, the CMP with the 2nd largest leakage on the tier with the 2nd smallest $\hat{a}_i$, and so on. This stacking algorithm would lead to the minimal $\hat{T}^{max}$. Please note that although the proposed learning model is used to predict the maximum temperature of the entire system under steady-state conditions, it is also a very good reference for capturing the trend for the maximum operating temperature. This stacking algorithm is exclusively enabled by our learning model due to the use of $\hat{a}_i$s; using other thermal models such as Hotspot simulation to exhaustively search for the best permutation of stacking orders would be dramatically slow. According to our data, searching for the best stacking order for 1,000 4-tier CMPs by using Hotspot simulation would take more than 5 days, while using our learning model and the proposed stacking algorithm would only take 4 hours; therefore a 30X speed-up is achieved. It is worth mentioning that these 4 hours are spent in the *training phase* to learn $\hat{a}_i$s. Once $\hat{a}_i$s are learnt, no additional simulation is required and the optimal stacking order can be determined instantly.

For simplicity and without losing generality, we use a 2-tier CMP as an example to describe how the proposed algorithm works. All notations here are the same as in Section III.B. Let us assume the leakage power dissipation of CMP 1 is 2W, of CMP 2 is 1W, $\hat{a}_1 = 20$, $\hat{a}_2 = 10$, and $\hat{c} = 40$. According to the proposed technique, CMP 1 will be placed on tier 2 since $\hat{a}_2$ is the smallest, and CMP 2 will be placed on tier 1. That is, P^1_{leak} = 1W and P^2_{leak} = 2W. Therefore, $\hat{T}^{max}$ equals $\hat{a}_1 \times P^1_{leak} + \hat{a}_2 \times P^2_{leak} + \hat{c}$ = 80°C, which is the minimal value. Any other stacking order will lead to larger $\hat{T}^{max}$. The complexity of

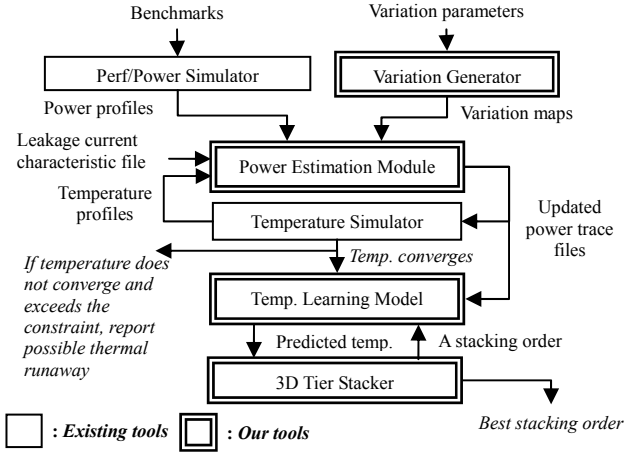


Figure 4: Overall Flow.

the proposed algorithm is $O(nlogn)$ because two instances of sorting are required to obtain sorted $\hat{a}_i$s and leakage values.

IV. IMPLEMENTATION

Parameters	Values
Chip size	3.2cm×3.2cm for a 1-tier CMP. 1.6cm×3.2cm for a 2-tier CMP. 1.6cm×1.6cm for a 4-tier CMP.
Heat sink thermal res.	0.24, based on [30].
Heat spreader size	Same as the Chip size.
Sampling rate	500K clock cycles.

Table 3: Hotspot Setup

In this section, we describe the experimental setup in detail, followed by providing the overall implementation flow. First, we use modified versions of SimpleScalar [25], Wattch [26], and Hotspot [8] for the performance, power, and thermal simulators, respectively. We modified the leakage power model in Wattch based on [27] [28] and as described in Section II.B, for more accurate leakage values. As to the Hotspot configuration, Table 3 lists the detail parameter settings; the parameters not mentioned here are assumed to be the default values. In addition, TSVs setup similar to [30] is used and modeled by the thermal conductivity values of grid cells in Hotspot.

According to [8], we separate SPECcpu2000 benchmarks into two categories: intermediate and intensive thermal demands, and then randomly select eight benchmarks from each category to form a representative multi-program workload for a 16-core CMP. To capture the worst-case scenario, the maximum temperature is assumed to occur when all processing cores are consuming power. Also, we assume that the workload executed in each tier includes programs of both intermediate and intensive thermal demands. This might not always be the case, especially when extreme task assignment strategies are used, but it is representative for realistic multi-programmed workload mixes. With the above settings, we perform a full-system simulation for 500 million instructions, and then collect the power profiles for the temperature simulation.

Figure 4 presents the overall flow of the proposed methodology. First, the benchmarks are fed in as inputs of performance and power simulators, to output both active and leakage power profiles. Second, we enter the variation parameters described in Section II.C to our in-house variation map generator based on the models in [19][20], for L_{eff} within each die. Third, we characterize leakage currents by using HSPICE simulation with the 45nm high performance Predictive Technology Model [29]. Next, the power estimation module collects power profiles, variation maps, temperature profiles and the leakage characteristics as inputs, and then updates the power values based on the current temperature value and process variations. The updated power values are fed into the temperature simulator to estimate the new temperature value. This temperature-power iteration will continue updating until the temperature value converges; the converged temperature and power profiles are analyzed by the learning-based regression model described in Section III.B to determine the coefficients. Note that in a real setting, at test per-tier leakage measurements will be used in the validated learning model to estimate temperature values. Finally, the 3D Tier Stacker will outputs the best stacking order that leads to the lowest maximum temperature by using the algorithm described in Section III.C.

V. EXPERIMENTAL RESULTS

This section presents the experimental results, including (1) the maximum operating temperatures of a 4-tier CMP under leakage variations, (2) the distributions of maximum operating temperatures for a 1-tier (2D), 2-tier, and 4-tier CMP, and (3) the distributions of maximum operating temperatures after tier-restacking. All experiments are implemented with the settings described in Section IV.

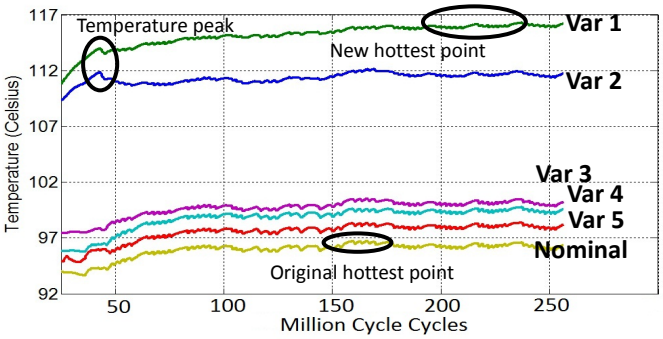


Figure 5: The distributions of max. transient temperatures.

A. *TRANSIENT THERMAL BEHAVIOR*

Figure 5 shows the transient profile of the maximum temperature for a 3D CMP under five different leakage variation maps. Note that the workload remains fixed for all cases. The X axis stands for the simulation time and the time unit is million clock cycles. The Y axis represents the maximum temperature observed across all tiers, at the given time instant on the X axis. "Var 1" and "Var 2" represent the temperatures under two cases of severe variations; "Var 3", "Var 4" and "Var 5" represent the temperatures under mild variations; "Nominal" represents the temperature without any variation. In Var 3, Var 4, and Var 5, the average temperatures are slightly higher than the nominal one, and the trends remain approximately the same.

From Figure 5, we can make three interesting observations: (1) During the time interval between the 150th and 250th million cycle, a "sawtooth" behavior periodically occurs in all six profiles. This happens because one of the cores in the top tier is executing *mesa*, a benchmark with a high temperature profile and clear execution phases. This phenomenon shows that the pattern of the maximum temperature distribution in a 3D system may be determined by a single application executed in the top tier. Note that this result matches the observation in [8]. (2) Leakage variations may alter the time point when the maximum temperature occurs. In the Nominal case, the maximum temperature occurs between the 150th to 200th million clock cycle. However, in Var 1, the time point of maximum temperature is shifted to around the 220th million cycle. (3) In both Var 1 and Var 2, an unexpected high thermal peak occurs at around the 45th million cycle, which is completely different from the thermal behavior of Nominal. We investigated this phenomenon, and found that the thermal peak originated from *bzip2* application running in tier 3. The benchmark *bzip2* has a higher thermal envelope than other benchmarks around the 45th million cycle. At the same time, the processing core in tier 1, right above the core executing *bzip2*, has very high L_{eff} variations. These variations make the core in tier 1 more sensitive to temperature changes. Thus, the original thermal profile of this core is altered, due to the strong thermal behaviors of the core underneath. If the process variations in tier 1 are mild, the influences of tier 3 will be supressed, and therefore not reflected explicitly on the overall temperature curve. This is the reason why the thermal peak does not occur either after the 50th million cycle or in Var3, Var 4 and Var 5. In sum, all above three important obervations show that the leakage variations may dramatically change the nature of the temperature profiles.

It is worth mentioning that in [14], the difference of the maximum temperature between a 2D CMP and a 2-tier 3D CMP is around 10°C, wheras our results show that the difference is around 15°C. One of the potential reasons is that we use the cycle-accurate power dissipation to simulate temperature profiles, while the authors of [14] used the steady-state values instead. Compated to the steady-state power values, the cycle-accurate ones can reflect real operating conditions more accurately, and thereby capture the peak temperature more precisely.

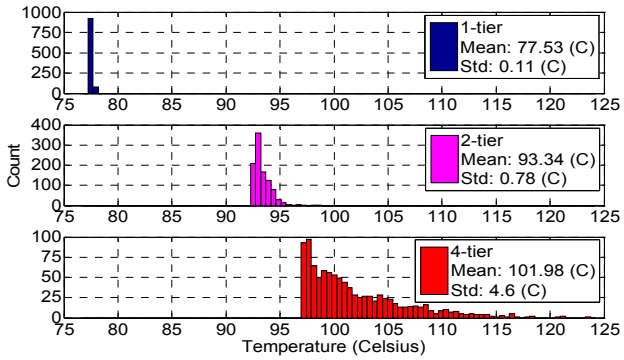


Figure 6: Maximum transient temperature distribution.

B. *MAXIMUM TEMPERATURE DISTRIBUTION*

We perform 1,000 Monte Carlo simulations with the settings described in Section IV. Figure 6 shows the distribution of the maximum temperatures of 1-tier (2D), 2-tier and 4-tier CMPs, respectively. The X axis represents the temperature whereas the Y axis stands for the count of parts in that temperature bin. Please note that the nominal maximum temperatures, i.e., the values without leakage variations of a 1-tier, 2-tier and 4-tier CMPs are 77.42˚C, 92.26˚C and 96.75˚C, respectively. From Figure 6, the mean of the temperature distribution increases dramatically from 2D to 3D CMP implementations. As it can be seen, the distribution for the 2D implementation is very narrow with a standard deviation of only 0.11˚C.

However, in 3D CMPs, the standard deviation of the maximum temperature distribution is significantly larger. For a 2-tier CMP, the standard deviation is approximately seven times higher than that of a planar CMP; for a 4-tier CMP, the standard deviation dramatically increases to approximately 40 times higher than that of a planar CMP. The significant variations in the maximum temperature from one 3D IC to another, necessitates a statistical thermal evaluation of the system along with mitigation techniques.

C. *TIER STACKING IMPROVEMENTS*

Figure 7 demonstrates the results of the proposed tier stacking algorithm for a 4-tier CMP. For better visualization, in Figure 7, we overlap the results after tier-restacking with the results before restacking depicted in Figure 6. It is clear that the variance of the maximum temperature distribution is much smaller. The standard deviation is reduced by 54%, from 4.6˚C to 2.11˚C; the mean is also reduced by 3%, from 101.98˚C to 98.77˚C. These statistical evaluations provide useful references for designers to determine to what degree they need to guard band the temperature constraints. Note that similar results were obtained for a 2-tier CMP, but are not reproduced here due to the space limit.

From the cumulative distribution function (CDF) in Figure 7, if the temperature constraint is set to 100˚C [31], the thermal yield for the 3D system after restacking is 80.1%, compared to the original yield 44.4%. If the temperature constraint is set to 105˚C, the improved yield is 98.0% compared to the original yield of 78.1%. This improvement in thermal yield clearly demonstrates the strength of our proposed techniques.

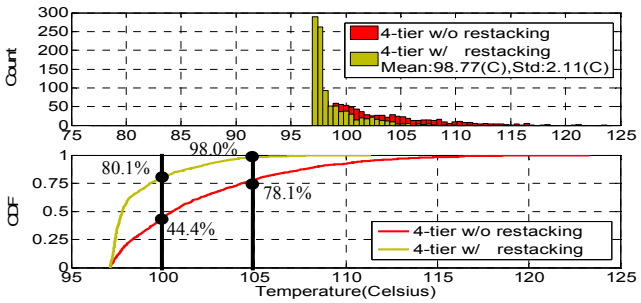


Figure 7: Improvement after tier restacking.

VI. CONCLUSION

In this paper, we propose a methodology to perform statistical thermal evaluation for 3D ICs. We also propose an accurate learning-based regression model to predict the maximum steady-state temperature that does not rely on expensive simulations, and can be used in an iterative design exploration environment for improving thermal yield. More precisely, based on this model, we propose an effective algorithm to determine the best tier stacking order that minimizes the maximum temperature and maximizes the thermal yield. The proposed algorithm significantly reduces the standard deviation and the mean by 54% and 3%, respectively, for the maximum operating temperature distribution of a 3D CMP.

ACKNOWLEDGMENT

The authors would like to thank Wangyang Zhang and the anonymous reviewers for the valuable suggestions and comments.

REFERENCES

[1] J. W. Joyner et al. “Impact of three-dimensional architectures on interconnects in gigascale integration,” *IEEE TRANS. ON VLSI SYSTEMS,* VOL. 9, pp. 922-928, 2001.
[2] Y. Xie et al, “Design space exploration for 3D architecture,” *ACM JETC*, VOL 2, pp. 65-103, 2006.
[3] B. Black et al., “Die stacking (3D) microarchitecture,” *MICRO,* pp. 469-479, 2006.
[4] D. Brook et al., “Power, thermal and reliability modeling in nanometer-scale microprocessors,” *MICRO*, pp. 49-62, 2007.
[5] K. Puttaswamy et al., “Thermal analysis of a 3D die-stacked high performance microprocessor,” *Proc. of ACM Great Lakes symposium on VLSI*, pp. 19-24, 2006.
[6] Y. Liu et al., “Accurate temperature-dependent integrated circuit leakage power estimation is easy,” *DATE*, pp. 1526-1531, 2007.
[7] K. Chakraborty et al., “Rethinking threshold voltage assignment in 3D multicore designs,” *Proc. of International Conference on VLSI Design*, pp. 375-380, 2010.
[8] K. Skadron et al., “Temperature-aware microarchitecture: Modeling and Implementation,” *TACO*, Volume 1 , Issue 1, pp. 94-125, 2004.
[9] J. Donald, “Techniques for multicore thermal management: classification and new exploration,” *ISCA*, pp. 78-88, 2006.
[10] T. Ebi et al., “TAPE: thermal-aware agent-based power economy for multi/many-core architectures,” *ICCAD*, pp. 302-309, 2009.
[11] B. Goplen et al., “Thermal via placement in 3D IC,“ *ISPD*, pp. 167-174,2005.
[12] J. Cong et al., “Thermal via planning for 3-D ICs,” *ICCAD*, pp. 745-752, 2005.
[13] C. Zhuo et al., “Process variation and temperature-aware reliability management,” *DATE*, pp. 580-585, 2010.
[14] S. Ozdemir et al, “Quantifying and coping with parametric variations in 3D-stacked microarchitectures,” *DAC*, pp. 144-149, 2010.
[15] S. Garg et al., “System-level process variability analysis and mitigation for 3D MPSoCs,” *DATE*, pp. 604-609, 2009.
[16] C. Ferri et al., ”Strategies for improving the parametric yield and profits of 3D ICs,” *ICCAD*, pp. 220-226, 2007.
[17] J. Fourier, *The analytical theory of heat*, 1822.
[18] “International technology roadmap for semiconductors,” 2009.
[19] L. Cheng et al., “Physically justifiable die-level modeling of spatial variation in view of systematic across wafer variability,” *DAC*, pp. 104-109, 2009.
[20] J. Sartori et al., “Variation-aware speed binning of multi-core processors," *ISQED*, 2010.
[21] R. E. Kessler, “The alpha 21264 microprocessor,” *MICRO*, pp. 24-36, 1999.
[22] M. Bushnell et al., “Essentials of electronic testing for digital, memory, and mixed-signal VLSI circuits,” *Kluwer Academic Publishers*, pp. 439-460, 2002.
[23] S.M. Alam et al., “Die/wafer stacking with reciprocal design symmetry (RDS) for mask reuse in three-dimensional (3D) integration technology,” *ISQED*, pp. 569-575, 2009.
[24] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *IJCAI*, pp. 1137-1143, 1995.
[25] D. Burger et al., “The SimpleScalar tool set, version 2.0,” *ACM SIGARCH Computer Architecture News*, pp.13-25, 1997.
[26] D. Brooks et al., “Wattch: a framework for architectural-level power analysis and optimizations”, *ISCA*, pp. 83-94 ,2000.
[27] J. A. Butt et al., “Static power model for architects,” *MICRO*, pp.191-201, 2000.
[28] S. Rusu et al., “A 65-nm dual-core multithreaded Xeon processor with 16-MB L3 Cache,” *IEEE JOURNAL OF SOLID-STATE CIRCUITS*, pp. 17-25, 2007.
[29] W. Zhao et al., "New generation of predictive technology model for sub-45nm early design exploration," *IEEE Transactions on Electron Devices*, vol. 53, no. 11, pp. 2816-2823, 2006.
[30] D. Sekar et al., “3D-IC technology with integrated microchannel cooling,” *Interconnect Technology Conference*, pp. 13-15, 2008.
[31] G. L. Loi et al., “A thermally-aware performance analysis of Vertically Integrated (3-D) Processor-Memory Hierarchy,” *DAC*, pp. 991-996, 2006.