Statistical Peak Temperature Prediction and Thermal Yield Improvement for 3D Chip Multiprocessors

DA-CHENG JUAN, Carnegie Mellon University
SIDDHARTH GARG, University of Waterloo
DIANA MARCULESCU, Carnegie Mellon University

Thermal issues have become critical roadblocks for achieving highly reliable three-dimensional (3D) integrated circuits (ICs). The presence of process variations further exacerbates these problems. In this article, we propose techniques for the efficient evaluation and mitigation of the impact of leakage power variations on the temperature profile of 3D Chip Multiprocessors (CMPs). Experimental results demonstrate that, due to the impact of process variations, a 4-tier 3D implementation can be more than 40°C hotter and 23% leakier than its 2D counterpart. To determine the maximum temperature of each fabricated 3D IC, we propose an accurate learning-based model for peak temperature prediction. Based on the learning model, we then propose two post-fabrication techniques to increase the thermal yield of 3D CMPs: (1) tier restacking and (2) thermally-aware die matching. Experimental results show that: (1) the proposed prediction model achieves more than 98% accuracy, and (2) the proposed thermally-aware, post-fabrication optimization techniques significantly improve the thermal yield from only 51% to 99% for 3D CMPs.

Categories and Subject Descriptors: B.8.0 [**Performance and Reliability**]: General

General Terms: Algorithms, Performance, Reliability

Additional Key Words and Phrases: 3D IC, temperature, process variation, chip multiprocessor, leakage power, statistical model, prediction, thermal hotspot

ACM Reference Format:
Da-Cheng Juan, Siddharth Garg, and Diana Marculescu. 2014. Statistical peak temperature prediction and thermal yield improvement for 3D chip multiprocessors. ACM Trans. Des. Autom. Electron. Syst. 19, 4, Article 39 (August 2014), 23 pages.
DOI: http://dx.doi.org/10.1145/2633606

1. 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. In a 3D IC, conventionally fabricated planar dies are restacked 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 [Brooks et al. 2007; Puttaswamy and Loh 2006], the thermal concerns for 3D ICs are exacerbated. In particular, those tiers farther away from the heat sink tend to have elevated temperature profiles, often up to 25°C

This work was supported in part by an Intel Ph.D. Fellowship for D.-C. Juan and NSF grant CNS-1128624.
Authors' addresses: D.-C. Juan (corresponding author), Department of Electrical and Computer Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA; email: dacheng@cmu.edu; S. Garg, Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue W, Waterloo, ON N2L 3G1, Canada; D. Marculescu, Department of Electrical and Computer Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA.

DOI: http://dx.doi.org/10.1145/2633606

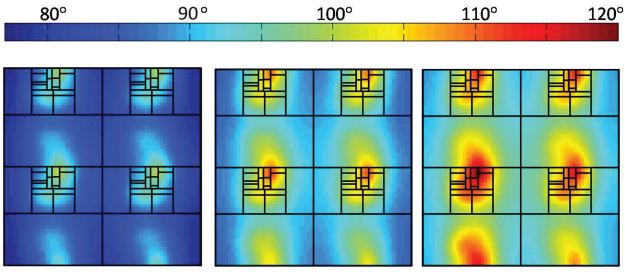


Fig. 1. Temperature maps of top tier in a 3D CMP.

higher than tiers closest to the heat sink [Black et al. 2006]. 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 [Liu et al. 2007; Kursun and Cher 2009]. 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 positive feedback loop that, 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.

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 [Chakraborty and Roy 2010]. Figure 1 depicts three temperature maps of the tier farthest away from the heat sink in a 3D Chip Multiprocessor (CMP) in the absence (Figure 1(a)), or presence (Figures 1(b) and (c)) of process variations. All implementation details are described in Section 4. 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.

In addition, since our experimental results show that leakage power variations result in greater thermal variability for a 3D CMP compared to an equivalent 2D implementation, it is critical to develop techniques to mitigate the impact of leakage variations on the peak temperature of 3D ICs. In this article, we propose two such post-fabrication techniques that exploit the flexibility offered by the 3D stacking process. These techniques are: (1) tier restacking, in which dies with greater leakage power dissipation due to process variations are stacked closer to the heat sink; and (2) optimal die matching.

1.1. Previous Work

Thermal modeling for both conventional planar and 3D ICs has received a lot of attention in the research community. Skadron et al. [2004] proposed Hotspot, an accurate temperature model for planar ICs, and later extended it to account for 3D circuits as well [Huang et al. 2006]. Sridhar et al. [2010] proposed 3D-ICE, a compact transient

thermal model specifically targeting at 3D ICs with multiple inter-tier microchannel liquid cooling. Juan et al. [2012b] used a learning-based autoregressive model to predict, instead of simulating, the operating temperature of a chip multiprocessor (CMP). Vincenzi et al. [2011] proposed a fast thermal model based on neural networks. This work is the first one to explicitly address the impact of leakage power variations on the peak temperature of a 3D IC. From a mitigation perspective, Donald and Martonosi [2006] used dynamic voltage and frequency scaling (DVFS) to avoid thermal emergencies in CMPs. Ebi et al. [2009] presented an agent-based power distribution approach to balance the power consumption of CMPs in a proactive manner. Goplen and Sapatnekar [2005] and Cong and Zhang [2005] proposed novel placement algorithms to reduce the temperatures in 3D systems. Chakraborty and Roy [2010] proposed a method to assign threshold voltages for 3D CMPs. Recently, Zhuo et al. [2010] 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. [2010], Garg and Marculescu [2009], and Ferri et al. [2007]. Recently, Chae and Mukhopadhyay [2012] proposed a post-silicon voltage tuning for 3D ICs. Reda et al. [2009] have extended the process corner simulation procedure to handle the timing and leakage variability for 3D technology, and Ferri et al. [2008] have proposed strategies to model the parametric yield of 3D ICs and optimally pair different die together such that revenues are maximized. However, all these papers consider the impact of variations on the timing or leakage characteristics of 3D ICs, not on temperature or thermal profiles.

1.2. Article Contributions

In this article, we propose techniques for the analysis and mitigation of the impact of process variations on the peak temperature of a 3D IC. We address this issue by 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 severely underestimate the actual maximum temperature observed by a large fraction of 3D systems.

Motivated by the evaluation results, we make three contributions to help 3D CMP designers mitigate the impact of leakage variations on the maximum operating temperature. First, by 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 speedup when compared to actual temperature simulations.

In order to mitigate the impact of leakage variations on the temperature profile of a 3D system, we propose two post-silicon techniques to mitigate the impact of process variations on the thermal envelope of 3D ICs. The first one, namely tier restacking, is applicable to *symmetric* 3D systems in which the tiers of the stack are identical, for example, the symmetric 3D CMPs studied in this article or 3D DRAM stacks, although the latter are not specifically addressed. Based on the proposed temperature modeling methodology, our algorithm is able to select the optimal tier restacking 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 [Loi et al. 2006], the proposed restacking technique increases the *thermal yield*[1] from 51.4% to 81.1%.

In order to further maximize the thermal yield, we perform a post-fabrication, thermally-aware die matching technique that optimally matches leaky and nonleaky dies to form a 3D system. Similar ideas have been proposed before in the context of variability-aware performance optimization for 3D ICs [Garg and Marculescu 2009; Ferri et al. 2007], but not in the context of mitigating the thermal impact of leakage variations. Unlike the proposed tier restacking algorithm that is limited to symmetric designs, this thermally-aware die matching can be applied on both symmetric and nonsymmetric designs. Furthermore, the proposed matching is orthogonal to the restacking algorithm, and therefore can be used in synergy to maximize the thermal yield. The experimental results demonstrate that the proposed die matching further improves the thermal yield to 99% under the peak temperature constraint of 100°C.

The remainder of this article is organized as follows. Section 2 introduces the related background knowledge required for our work. Section 3 details the proposed methodologies. Section 4 presents the implementation flow. Section 5 demonstrates the experimental results and Section 6 concludes.

2. 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.

2.1. Thermal Model

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

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

where $\boldsymbol{C}$ is a diagonal thermal capacitance matrix, $\boldsymbol{R}$ is a thermal resistance matrix, $\boldsymbol{T}(t) = (T_1 - T_A, \ldots, T_n - T_A)^T$ is the temperature vector, T_A is the ambient temperature, $\boldsymbol{p} = (p_1, \ldots, p_n)^T$ is the power vector, and $U(t)$ is a step function. Bold symbols here represent vectors or matrices instead of scalars.

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

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

From Eq. (1) and Eq. (2), the power vector $\boldsymbol{p}$ 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.

2.2. 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

[1]In this article, *thermal yield* is defined as the proportion of fabricated CMPs whose peak operating temperature is below the given temperature constraint.

the active power is not sensitive to the temperature. Leakage power contains several components, among which subthreshold leakage current and gate leakage current are main contributors [Roadmap 2009]. Recently, due to the introduction of high-k dielectrics, the gate leakage component has become less important. We therefore concentrate only on subthreshold 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 as

$$I_{leak} = K\frac{W}{L_{eff}}\left(\frac{kT}{q}\right)^2 e^{\frac{q(-V_{th})}{nkT}} = K'\frac{W}{L_{eff}}(T)^2 e^{-\frac{\beta \cdot V_{th}}{T}}, \tag{3}$$

where K' is a technology-dependent constant, W is the transistor width, L_{eff} is the effective channel length, T is the temperature, V_{th} is the threshold voltage, and β is a positive constant. According to Eq. (3), I_{leak} will increase when L_{eff} decreases and T increases. Combining Eqs. (1)–(3) determines the interdependency of temperature and leakage power. The increase of either temperature or leakage power will trigger this positive feedback loop [Juan et al. 2012a].

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 Eqs. (3) and (4), it is clear that 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 article mainly focuses on the L_{eff} variation because of the aforementioned exponential dependency.

2.3. Process Variation Model

Process variations significantly affect several important metrics of an IC, such as the maximum clock frequency and leakage power dissipation [Dighe et al. 2011], which in turn affect the temperature due to the interdependency between temperature and leakage power.

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 Eqs. (4) and (5) into Eq. (3) to demonstrate how L_{eff} variations affect leakage currents:

$$I_{leak} = \frac{K''}{(L_{nom} \pm \Delta_{total})}(T)^2 e^{-\frac{\beta_1 - \beta_2 \cdot e^{-\alpha \cdot (L_{nom} \pm \Delta_{total})}}{T}}, \tag{6}$$

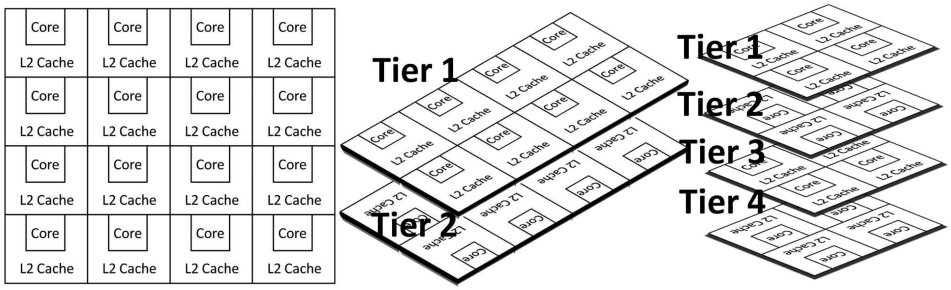
where K'' is a technology-dependent constant, and β_1 and β_2 are both positive constants. From Eq. (6), it is clear that if Δ_{total} increases, I_{leak} may see a significant increase due to the exponentially and inverse-linearly-dependent terms, respectively. Furthermore, Δ_{total} can be decomposed 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

Table I. Processor Parameters

Parameters	Values
Number of cores	16
Frequency	3.0 GHz
Technology	45nm node with Vdd = 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



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

Fig. 2. 2D and 3D CMP implementation.

corresponding weights. Δ_{w2w} and Δ_{rand} can be modeled as Gaussian random variables. In Cheng et al. [2011], the authors proposed an accurate, deterministic model of Δ_{spat} by exploiting across-wafer variation. In this article we use this model assuming Δ_{total} of 5% of L_{nom}; furthermore, based on Sartori et al. [2010], the relative ratio among Δ_{w2w}, Δ_{spat}, and Δ_{rand} is set to 0.7:1:1.

3. METHODOLOGY

In this section, we first introduce symmetric CMPs, the target architecture used in this article. In Section 3.2, 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 are included. In Section 3.3, we present the algorithm to determine the order of tier stacking for 3D CMPs by using the proposed methodology of temperature estimation. Finally, in Section 3.4, we elaborate on the proposed thermally-aware die matching that can be applied on both symmetric and asymmetric designs to further improve the thermal yield.

3.1. Target Architecture

The architecture used throughout this article is a symmetric CMP consisting of 16 out-of-order, Alpha 21264 cores [Kessler 1999]. The corresponding microarchitecture parameters are listed in Table I. The floorplan of a single Alpha 21264 processor taken from Skadron et al. [2004] 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 Figures 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 [Alam et al. 2009]. 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 article, the focus is on thermal evaluation, as opposed to performance that has

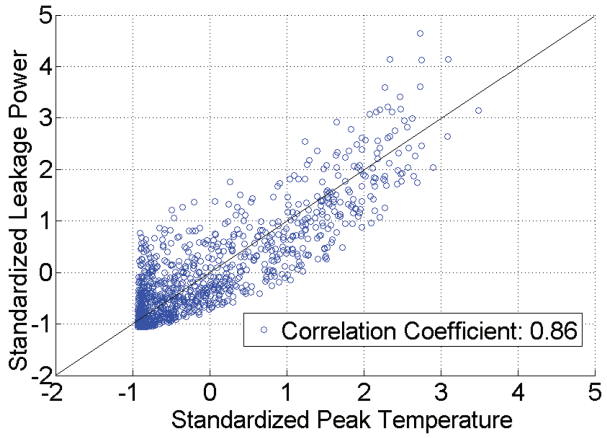


Fig. 3. Scatter plot of leakage power and peak temperature.

been extensively addressed by Ozdemir et al. [2010], Garg and Marculescu [2009], and Ferri et al. [2007]. Therefore, the detailed performance comparison between 2D and 3D CMPs is out of the scope of this article and not addressed here.

3.2. Learning-Based Model for Temperature Prediction

As shown in Figure 1 in Section 1, the maximum temperature of a 3D CMP under the impact of process variations can be significantly higher than the temperature obtained under 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 that account for the impact of leakage variations.

In the presence of leakage power variations, the maximum temperature for each 3D system cannot be 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 [Skadron et al. 2004] or other simulation-based methods to determine the maximum temperature for each fabricated 3D system can be too time consuming. We found that, within the range of the processors' operating temperature (around 70°C to 120°C), 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 Eqs. (3)–(6). Figure 3 illustrates this phenomenon. In Figure 3, the x-axis represents the peak temperature of a 4-tier CMP, whereas the y-axis stands for the leakage power of the tier farthest away from the heat sink. Both values are standardized (scaled to zero mean and unit variance) for better visualization (before standardization, the range of the peak temperature is from 96.92°C to 136.13°C and the range of leakage power is from 13.2 Watts to 38.2 Watts). This observation motivates the development of a learning-based, linear 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. These leakage measurements are routinely performed on bare dies before packaging as part of the popular IDDQ testing methodology [Bushnell and Agrawal 2000], therefore we do not introduce any additional test costs for maximum temperature prediction.

Based on the aforementioned observation, we propose a learning-based regression model that 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 CMP under the steady-state condition, a_is 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 article, 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. By convention, a_is are annotated according to the distance from the heat sink, that is, a_1 is the tier furthest away from the heat sink, whereas a_4 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 (the number of 3D CMPs whose T_k^{max} are known). In this article, m is empirically set to 500. 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. Furthermore, $\hat{a}_i$s are not the same 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 relearned by reevaluating 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 Eqs. (8) and (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. We point out 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.

To evaluate the accuracy of the proposed learning model, we use tenfold cross-validation [Kohavi et al. 1995] to calculate the prediction error. Cross-validation is a nearly unbiased error estimator widely used in machine learning and statistics fields. Figure 4 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%. To further evaluate the sensitivity of the proposed model to inter-die

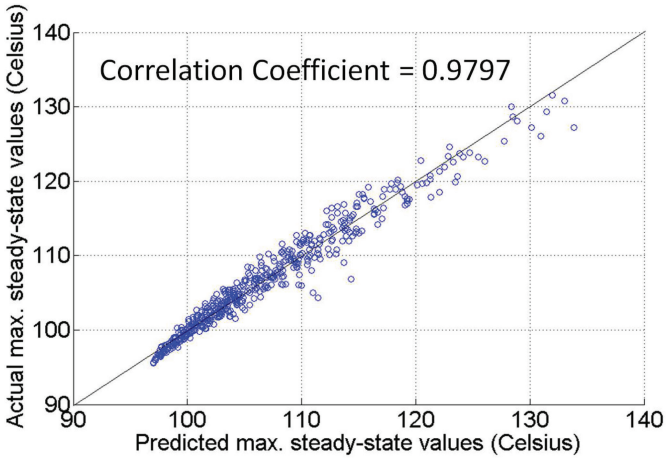


Fig. 4. The prediction accuracy for a 16-core, 4-tier 3D CMP.

Table II. The Values of $\hat{\alpha_i}$

$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$	$\hat{\alpha_4}$
2.052	0.7182	1.2312	0.5643

thickness (the default value of Hotspot is 20μm), we change the values from 10μm to 30μm according to Ivankovic et al. [2011] and Jung et al. [2014], and repeat the whole experiment. The errors are almost identical (less than 0.01% different) to 1.02%, showing the proposed model is robust to the change of inter-die thickness.

We also reimplemented the method proposed by Juan et al. [2012b] and received a similar prediction error (1.05%). 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. Since the active power is not sensitive to the change of temperature as mentioned in Section 2.2, it is implicitly modeled as part of the constant term $\hat{c}$. We would like to stress that this learning model does not aim to replace the original thermal model described in Eqs. (1) and (2) or thermal simulators like Hotspot. The model relies on accurate temperature analysis or simulation to provide inputs with good quality to learn the fitting coefficients. The goal of the proposed model is to allow for fast 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 or optimal die matching for generic 3D ICs that minimizes maximum temperature.

As an example, we list the values of $\hat{\alpha_i}$s for a 16-core 4-tier CMP in Table II. The values of $\hat{\alpha_i}$s would be expected to be monotonically decreasing according to the tier ordering: $\hat{\alpha_1} > \hat{\alpha_2} > \hat{\alpha_3} > \hat{\alpha_4}$, since the impact on the maximum temperature of tiers farther away from the heat sink is expected to be higher. As expected, $\hat{\alpha_1}$ 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, $\hat{\alpha_3}$ is larger than either $\hat{\alpha_2}$ or $\hat{\alpha_4}$. The reason for this nonmonotonic behavior stems from the relative positioning of processing cores: in tier 3, the cores are located directly underneath the cores of the hottest tier (tier 1), as shown in Figure 5. The fourth tier is intentionally not shown so that we can focus on the interplay between tier 1 and tier 3. Therefore, the vertical heat conduction of cores between tier 3 and tier 1 increases the

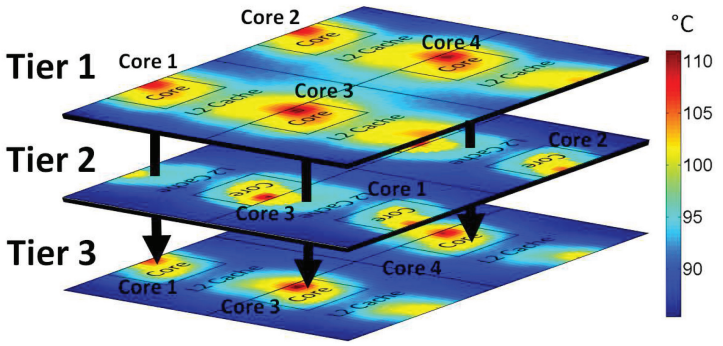


Fig. 5. The correlation between tiers 1 and 3.

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 [Rusu et al. 2007], thereby making them more robust to leakage variations.

3.3. Tier Restacking

As observed from the $\hat{a}_i$s of the learning model in Section 3.2, 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 restack the tiers based on their leakage values so as to keep the tiers with high leakage values closer to the heat sink and thereby achieve a potential reduction in the maximum temperature? As a result, this stacking technique would only be applicable for symmetric 3D systems, that is, where each tier has the same layout. This is certainly the case for the 3D CMP system we study in this article. Also, the stacking technique would be applicable to those systems that contain multiple identical stacked SRAM or DRAM layers, or to the recently introduced Reciprocal Design Symmetry (RDS)-based 3D ICs [Alam et al. 2009].

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 useful 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 second largest leakage on the tier with the second smallest $\hat{a}_i$, and so on. This stacking algorithm would lead to the minimal $\hat{T}^{max}$ and 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 five days, while using our learning model and the proposed stacking algorithm would only take four hours; therefore a 30X speedup is achieved. It is worth mentioning that these four hours are spent in the training phase to learn $\hat{a}_i$s. Once $\hat{a}_i$s are learned, no additional simulation is required and the optimal stacking order can be instantly determined.

The tier restacking algorithm is shown in Algorithm 1. The algorithm takes as inputs the measured leakage power dissipation of each die, P^i_{leak}, and provides as an output $S_i(1 \leq i \leq n)$, the tier to which die i is assigned. For simplicity and without losing

ALGORITHM 1: Tier Restacking

Input: The leakage power of each die, P^i_{leak}
Output: The tier assignment of each die, S
/* Begins algorithm */
Sort $\hat{a}_j$ in a descending order;
Let $\mathcal{R}(j)$ be the rank of the j^{th} coefficient;
Sort P^i_{leak} in ascending order;
Let $Q(i)$ be the rank of i^{th} die in terms of leakage power dissipation;
for *each die* $i \in [1, n]$ **do**
    $S(i) = j$ such that $Q(i) = \mathcal{R}(j)$
end
/* Ends algorithm */

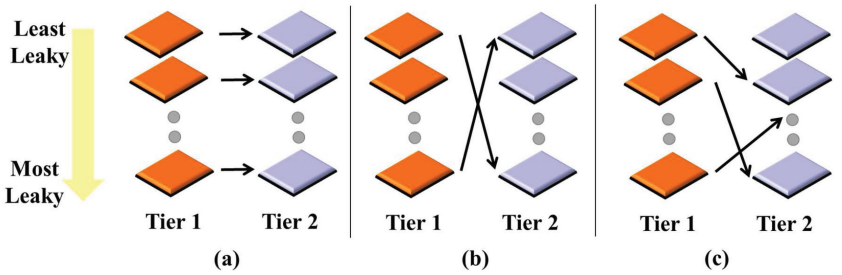
generality, we use a 2-tier CMP as an example to describe how the proposed algorithm works. All notations here are the same as in Section 3.2. Let us assume the leakage power dissipation of CMP 1 is 2W, of CMP 2 is 1W, $\hat{a_1} = 20$, $\hat{a_2} = 10$, and $\hat{c} = 40$. According to the proposed 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, $P^1_{leak} = 1$W and $P^2_{leak} = 2$W. Hence, $\hat{T}^{max}$ equals $\hat{a_1} \times P^1_{leak} + \hat{a_2} \times P^2_{leak} + \hat{c} = 80^\circ$C, which is the minimal value. Any other stacking order will lead to larger $\hat{T}^{max}$. The complexity of the proposed algorithm is $O(n \times log(n))$ because two instances of sorting are required to obtain sorted $\hat{a_i}$s and leakage values.

3.4. Thermally-Aware Die Matching

For die-to-die bonded 3D ICs, there is an additional degree of flexibility in how to combine or match the dies from each tier into 3D systems [Garg and Marculescu 2009; Alam et al. 2009]. Assuming we are given: (1) k fabricated dies in each of the n tiers of the 3D IC, and (2) the measured leakage power dissipation of each die, we need to determine how to assemble k n-tier CMPs so as to maximize thermal yield. As we mentioned in Section 1, the *thermal yield* is defined as the percentage of assembed 3D chips with a peak temperature below the desired peak temperature.

One possible die matching strategy is to match the leakiest dies with the least leaky dies to form a 3D stack. An alternative strategy is to match the least leaky dies with the least leaky dies. Each matching strategy can lead to a different distribution of peak temperature for the resulting 3D ICs and thus different thermal yields. The goal is to pick the matching strategy that maximizes thermal yield. Note that, although die matching has been proposed before in the context of performance maximization of 3D ICs, assuming that the dies from each tier can lie in one of many frequency bins, the optimization technique proposed for this scenario cannot be used for thermally-aware optimal die matching. This is because, unlike chip frequency that can only take discrete values, the leakage power dissipation of each die is a continuous random variable.

We now describe in detail our proposed thermally-aware optimal die matching strategy, which is predicated on our observation that the maximum temperature can be accurately modeled as a weighted sum of the leakage power dissipation of each tier in the 3D IC. We begin by noting that die matching introduces a correlation between the leakage power random variables of the tiers in the 3D stack. This is illustrated in Figure 6, in which the dies from each tier are first sorted in descending order of their leakage power dissipation. The matching strategy in Figure 6(a) results in the leakage power random variables of the two tiers being positively correlated (we refer to this strategy as P-matching), while the strategy in Figure 6(b) results in negatively correlated leakage power dissipation (we refer to this strategy as N-matching).



(a) positively correlated (b) negatively correlated (c) uncorrelated

Fig. 6. Different die matching schemes.

Let P^i_{leak} represent the leakage power dissipation random variable tier i and, without loss of generality, assume that the leakage power random variables have been shifted and scaled to be zero mean and unit variance. Let m_i be the correlation coefficient between the leakage power dissipation of tier i and tier $i+1$, which is mathematically defined as

$$m_i := \mathbb{E}\left[P^i_{leak} \times P^{i+1}_{leak}\right], \tag{11}$$

where $m_i := \mathbb{E}[P^i_{leak} \times P^{i+1}_{leak}]$ represents the correlation coefficient between the leakage power dissipation of tier i and tier $i+1$ and is a direct consequence of the matching between the tiers.

We now show that the matching strategy in Figure 6(a), namely P-matching, maximizes m_i while that in Figure 6(b), that is, N-matching, minimizes m_i.

LEMMA 1. *P-matching results in the largest value of m_i compared to any other matching strategy.*

PROOF. We prove this claim by contradiction. Let $P^{i,j}_{leak}$ represent the leakage power of the die in the i^{th} tier ($1 \le i \le n$) in the j^{th} ($1 \le j \le k$) assembled 3D system. Now assume that the P-matching strategy is not used. Therefore, there must exist integers a and b ($1 \le a < b \le k$) such that $P^{i,a}_{leak} > P^{i,b}_{leak}$ but $P^{i+1,a}_{leak} < P^{i+1,b}_{leak}$. Let $\hat{m}_i$ be the correlation coefficient for this matching strategy.

$$\hat{m}_i = \left(\sum_{j \in [1,k]-\{a,b\}} P^{i,j}_{leak} P^{i+1,j}_{leak} \right) + P^{i,a}_{leak} P^{i+1,a}_{leak} + P^{i,b}_{leak} P^{i+1,b}_{leak}.$$

However,

$$P^{i,a}_{leak} P^{i+1,a}_{leak} + P^{i,b}_{leak} P^{i+1,b}_{leak} < P^{i,a}_{leak} P^{i+1,b}_{leak} + P^{i,b}_{leak} P^{i+1,a}_{leak}.$$

In other words, the correlation coefficient can be increased by swapping the tier $i+1$ dies in 3D ICs (a and b). Therefore, any matching strategy that is not a P-matching cannot maximize the correlation coefficient m_i. □

The converse argument can be used to prove that N-matching results in the smallest value of m_i.

Since the maximum temperature of a 3D stack is a weighted linear combination of the leakage power dissipation of each tier, the average value of the peak temperature[2]

[2]Recall that the peak temperature is a random variable due to process variations.

does not depend on the matching solution. From Eq. (8), and noting that the leakage power random variables are normalized to have zero mean, we observe that

$$\mathbb{E}(T^{max}) = c + \sum_{i=1}^{n} a_i \times \mathbb{E}\left(P^i_{leak}\right) = c. \tag{12}$$

We have verified this observation empirically in the experimental results section (Section 5.4).

On the other hand, as we will show, the die matching solution has a significant impact on the standard deviation of the maximum temperature. Thus, to maximize the thermal yield for any temperature constraint greater than the mean temperature, we seek to determine the optimal matching $M^* = \{m^*_1, \ldots, m^*_{n-1}\}$ that minimizes the variance of maximum temperature, in other words, minimizes $\mathbb{E}[(T^{max} - c)^2]$. Here, we provide a corresponding lemma by using 2-tier CMPs as an example.

LEMMA 2. *For $n = 2$, $M^* = \{m^*_1\} = \{-1\}$. In other words, for a 2-tier 3D stack, the optimal die matching solution is N-matching.*

PROOF. Based on Eq. (12), we can rewrite $\mathbb{E}[(T^{max} - c)^2]$ as

$$\mathbb{E}[(T^{max} - c)^2] = \sum_{i=1}^{2} a_i^2 + 2a_1a_2m_1. \tag{13}$$

Since the a_i coefficients are always positive as their estimates shown in Table II, and since N-matching results in the smallest value of m_1, we can see that N-matching minimizes $\mathbb{E}[(T^{max} - c)^2]$. □

Based on this observation, we propose a greedy algorithm that iteratively determines the best choice of M^*, starting from m_1 to m_{n-1}. When a new tier is added, the algorithm determines the variance in peak temperature that would result from an N-matching versus a P-matching to the current topmost tier and picks the solution that results in lower variance. The proposed greedy algorithm is formally described in Algorithm 2. Please note that we do not claim optimality for $n > 2$, in other words, our solution is optimal only for $n = 2$. For $n > 2$, our proposed algorithm is a greedy algorithm that makes a locally optimal decision as each tier is integrated, but there is no guarantee of global optimality. Our empirical results demonstrate that, even so, the improvements in yield are significant.

ALGORITHM 2: Die Matching

Input: The thermal sensitivity of tier i, a_i
Output: The correlation between tier i and tier $i + 1$, m_i
/* Begins algorithm */
$m_1 = -1$;
for *each die* $i \in [2, n-1]$ **do**
  $v = \sum_{j=1}^{i} a_j \times a_{i+1} \times \prod_{k=j}^{i} m_k$;
  if $v \geq 0$ **then**
    Use N-matching
  else
    Use P-matching
  end
end
/* Ends algorithm */

Table III. Hotspot Setup

Parameters	Values
Chip size for a 1-tier CMP	3.2cm × 3.2cm
Chip size for a 2-tier CMP	1.6cm × 3.2cm
Chip size for a 4-tier CMP	1.6cm × 1.6cm
Heat sink thermal resistance	0.24
Heat spreader size	Same as the chip size
Sampling rate	500K clock cycles
Resolution for thermal analysis	64 × 64 per die

4. IMPLEMENTATION

In this section, we describe the experimental setup in detail, followed by providing the overall implementation flow. First, we use SimpleScalar [Burger and Austin 1997], Wattch [Brooks et al. 2000], and Hotspot [Huang et al. 2006] for the performance, power, and thermal simulators, respectively. We modified the leakage power model in Wattch based on Butts and Sohi [2000] and Rusu et al. [2007] and as described in Section 2.2, for more accurate leakage values. The average power consumption of the whole CMP is approximately 139 Watts. As for the Hotspot configuration, Table III lists the detailed parameter settings; the parameters not mentioned here are assumed to be the default values. Note that the through silicon vias (TSVs) in Hotspot are assumed to be deployed homogeneously, leading to a uniform inter-die thermal conductivity [Huang et al. 2006]. The spatial resolution of Hotspot is set to 64 × 64 per die and the same resolution is also used for the leakage power calculation. Since the proposed learning model is independent from any simulator or any configuration, the parameters can be changed to meet the required accuracy and the proposed framework will still be applicable.

According to Skadron et al. [2004], we separate SPECcpu2000 benchmarks into two categories, namely intermediate and intensive thermal demands, and then randomly select eight benchmarks from each category to form a representative multiprogram 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 multiprogrammed workload mixes. With the preceding settings, we perform a full-system simulation for 500 million instructions and then collect the power profiles for the temperature simulation.

Figure 7 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 2.3 to our in-house variation map generator based on the models in Cheng et al. [2011] and Sartori et al. [2010], for L_{eff} within each die. Third, we characterize leakage power by simulating a 14-stage ring oscillator under all possible values of L_{eff} (from 35nm to 55nm) and operating temperature (from 20°C to 150°C) via HSPICE, with the 45nm high-performance predictive technology model [Zhao and Cao 2007]. 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 3.2 to determine the coefficients.

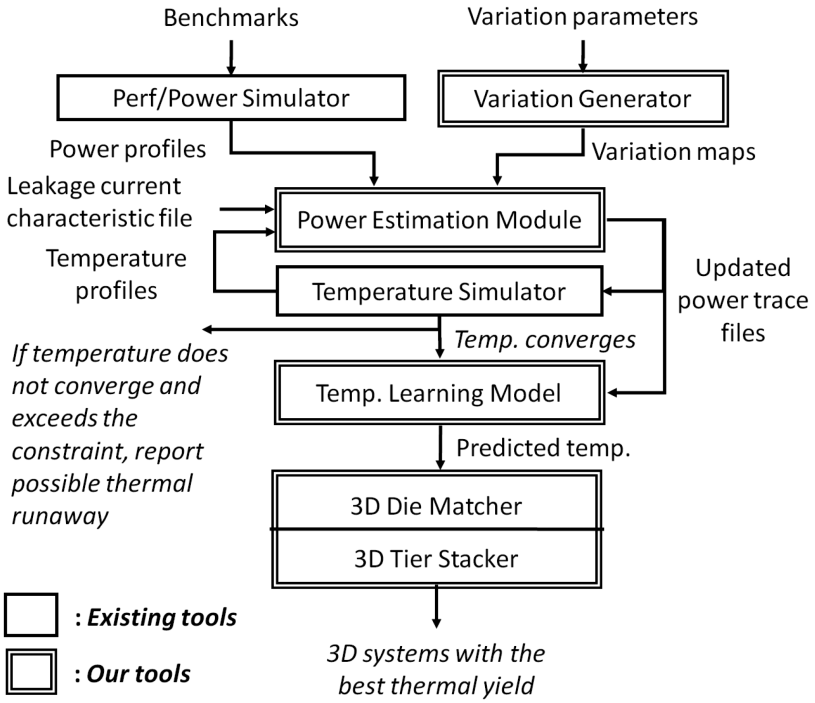


Fig. 7. Overall flow.

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 and die matcher described, respectively, in Sections 3.3 and 3.4 output the best stacking order and die matching solutions that maximize the thermal yield.

5. 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 as well as leakage power for a 1-tier (2D), 2-tier, and 4-tier CMP, and (3) the distributions of maximum operating temperatures and the corresponding thermal yield improvements after tier restacking, die matching, and both. All experiments are implemented with the settings described in Section 4.

5.1. Transient Thermal Behavior

Figure 8 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 a 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 8, we can make three interesting observations: (1) During the time interval between the 150^{th} and 250^{th} million cycle, a "sawtooth" behavior periodically

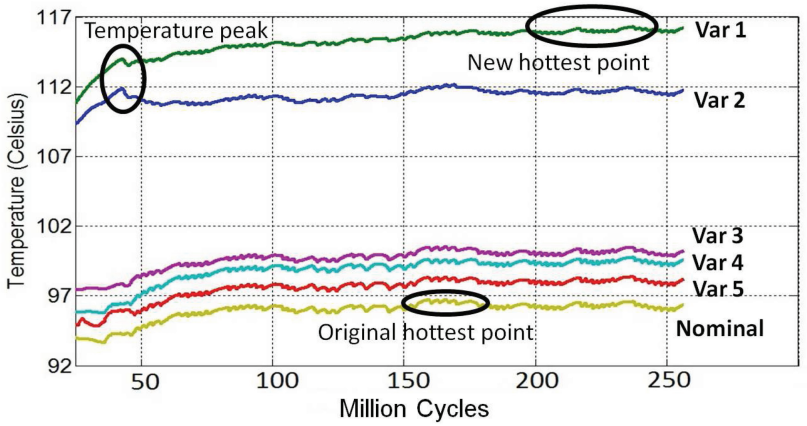


Fig. 8. The distributions of maximum transient temperatures.

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. Also, this result matches the observation in Skadron et al. [2004]. (2) Leakage variations may alter the time point when the maximum temperature occurs. In the Nominal case, the maximum temperature occurs between the 150^{th} to 200^{th} million clock cycle. However, in Var 1, the time point of maximum temperature is shifted to around the 220^{th} million cycle. (3) In both Var 1 and Var 2, an unexpected high thermal peak occurs at around the 45^{th} million cycle, which is completely different from the thermal behavior of Nominal. We investigated this phenomenon and found that the thermal peak originated from the *bzip2* application running in tier 3. The benchmark *bzip2* has a higher thermal envelope than other benchmarks around the 45^{th} million cycle. At the same time, the processing core in tier 1, right above the core executing *bzip2*, has very high 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 suppressed and therefore not reflected explicitly on the overall temperature curve. This is the reason why the thermal peak does not occur either after the 50^{th} million cycle or in Var3, Var 4, and Var 5. In sum, all aforesaid three important observations show that the leakage variations may dramatically change the nature of the temperature profiles.

It is worth mentioning that, in Ozdemir et al. [2010], the difference of the maximum temperature between a 2D CMP and a 2-tier 3D CMP is around 10°C, whereas 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 Ozdemir et al. [2010] used the steady-state values instead. Compared to the steady-state power values, the cycle-accurate ones can reflect real operating conditions more accurately and thereby more precisely capture the peak temperature.

5.2. Maximum Temperature Distribution

We perform 1,000 Monte Carlo simulations with the settings described in Section 4. Figure 9 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

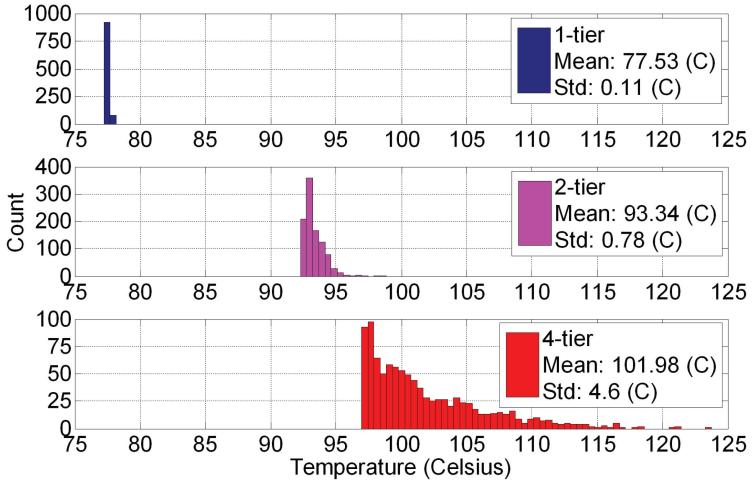


Fig. 9. Maximum transient temperature distribution.

y-axis stands for the count of parts in that temperature bin. The nominal maximum temperatures, namely 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 9, the mean of the temperature distribution increases dramatically from 2D to 3D CMP implementations. As 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 7× higher than that of a planar CMP; for a 4-tier CMP, the standard deviation dramatically increases to approximately 40× 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.

We also provide the corresponding leakage power values in Figure 10. Note that the values are normalized to the nominal leakage power of a 2D CMP, that is, the value without considering process variations. As references, we also clearly mark the respective nominal leakage of a 2-tier, and a 4-tier CMP. For all three cases (a 1-tier (2D), a 2-tier, and a 4-tier CMP), the mean of the leakage power distribution increases with respect to their respective nominal values by 32%, 36%, and 35%. In all three cases, the standard deviations have similar values and range from 0.32 to 0.34.

From Figure 10, we observe three important phenomena: (1) when the number of tiers increases, the mean leakage power dissipation also increases but the standard deviation remains almost the same, resulting in a tighter distribution around the mean value. (2) For 2D CMPs, the worst-case leakage power dissipation is about 2.3× larger than the nominal value, while for 2-tier 3D CMPs, it is 2.35× larger than its respective nominal and reaches 2.43× for 4-tier 3D CMPs, indicating that the worst-case leakage power dissipation increases as the number of tiers in the 3D stack increases. (3) For 2D CMPs, process variations can occasionally lead to lower than nominal total leakage power dissipation. From the experimental results, we noticed approximately 140 of 1,000 CMPs that consumed less leakage power compared to the nominal value, while for 2-tier and 4-tier 3D cases, that proportion is slightly less (121 and 109 out of 1,000 dies, respectively). All aforementioned phenomena indicate that process variations deteriorate leakage power issues more seriously in 3D systems than in 2D ones.

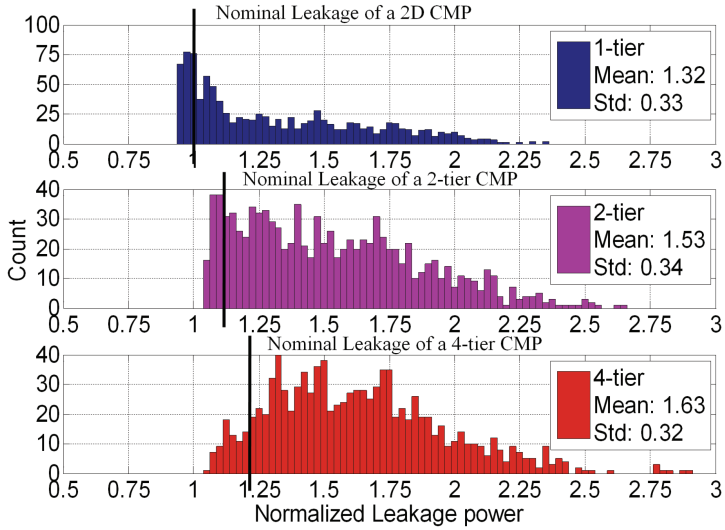


Fig. 10. Average leakage power distribution.

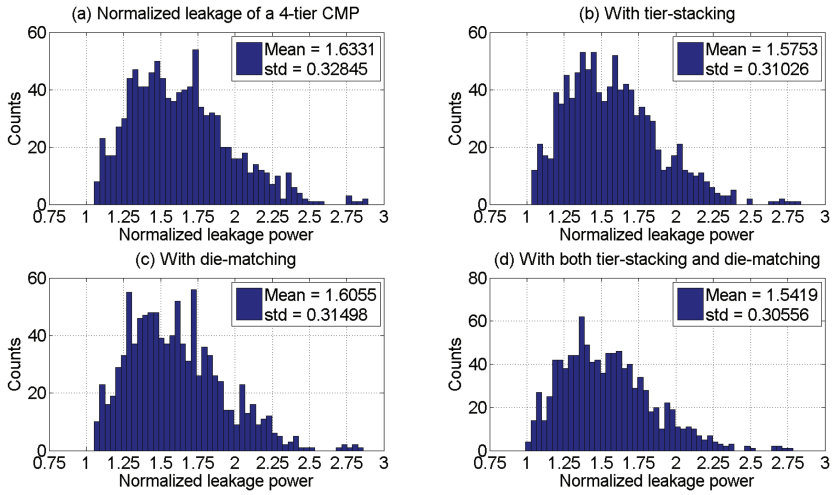
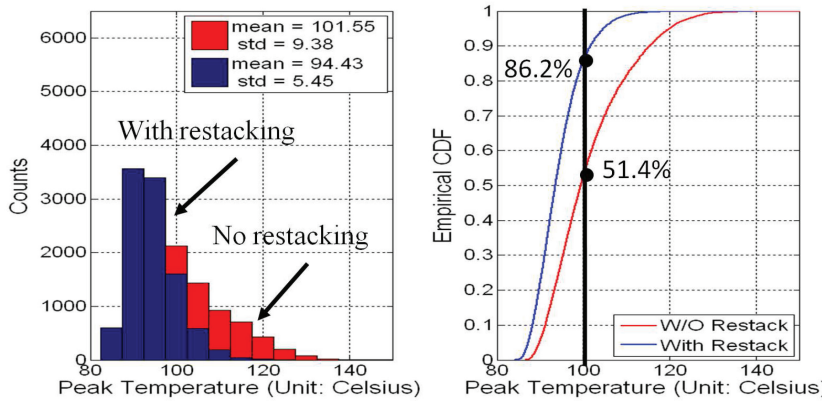


Fig. 11. Average leakage power distributions with tier stacking and die matching.

Similar to the illustration of Figure 10, Figure 11 further shows the leakage power distribution of a 4-tier CMP: (a) without tier stacking and die matching, (b) with tier stacking only, (c) with die matching only, and (d) with both tier stacking and die matching. From Figure 11, two phenomena can be observed. First, although the goal of the proposed tier stacking and die matching is to reduce the peak temperature, they also help reduce the thermally-induced leakage power. All mean leakage values in Figures 11(b), 11(c), and 11(d) are smaller than Figure 11(a). Second, the trend of leakage power distribution remains approximately the same, with a 5.5% smaller mean value (comparing (a) with (d)).

5.3. Tier Restacking Improvements

To extensively evaluate the proposed tier restacking, we further perform 40,000 Monte Carlo simulations with the settings described in Section 4 for collecting the leakage



(a) peak temperature distribution (b) cumulative distribution function (CDF)

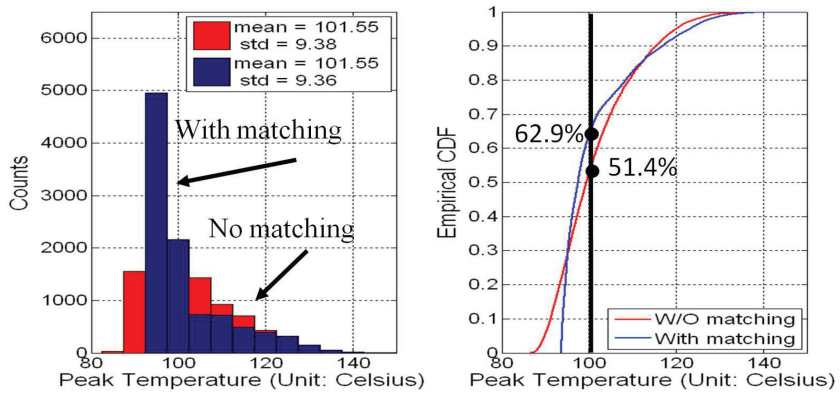
Fig. 12. Improvement after tier restacking.

measurements and the corresponding peak temperature from 10,000 4-tier CMPs. Here, we target 4-tier CMPs because the thermal yield issue is not as severe for 2-tier CMPs. Figure 12 demonstrates the results of the proposed tier stacking algorithm for a 4-tier CMP. For better visualization, in Figure 12(a), we overlap the distribution after tier stacking (highlighted as blue) with the results before stacking (highlighted as red). All dies here are randomly matched instead of using the proposed thermally-aware die matching. It is clear that the variance of the maximum temperature distribution is much smaller. The standard deviation is reduced by 42%, from 9.38°C to 5.45°C; the mean is also reduced by 7%, from 101.55°C to 94.43°C. These statistical evaluations provide useful references for designers to determine to what degree they need to guard band the temperature constraints.

From the cumulative distribution function (CDF) in Figure 12(b), the red line depicts the peak temperature from randomly-stacked 3D CMPs, whereas the blue line depicts the one from the proposed restacking algorithm. If the temperature constraint is set to 100°C [Loi et al. 2006], the thermal yield for the 3D system after restacking is 86.2% compared to the original yield 51.4%. For symmetric designs, this improvement in thermal yield clearly demonstrates the strength of our proposed tier restacking technique.

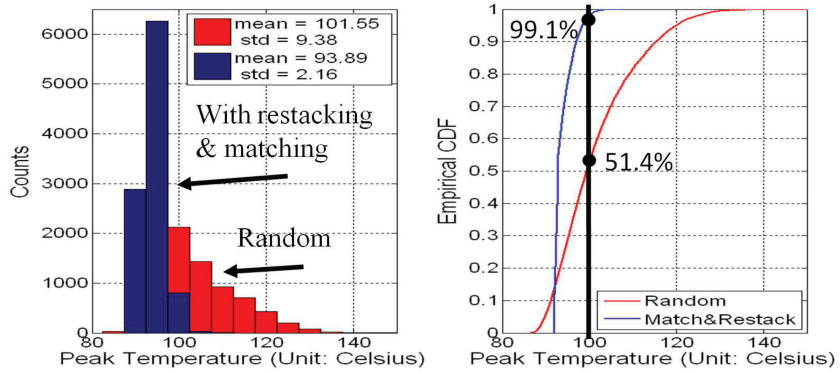
5.4. Die Matching Improvements

Figure 13(a) shows the peak temperature distribution of a 4-tier CMP from 40,000 dies. Similar to Figure 13, we overlap the original peak temperature distribution (highlighted as red) with the one after applying the die matching described in Section 3.4 (highlighted as blue). Please note that all dies here are randomly stacked instead of using the proposed tier restacking. Although the mean temperatures of both distributions are the same (101.55°C) as we point out in Section 3.4, the distribution with die matching, namely the blue one, is more skewed toward to the left and centered around 95°C. This trend can be seen more clearly in Figure 13(b). Figure 13(b) provides the cumulative density function (CDF) of these two distributions. The red line depicts the peak temperature from randomly matched 3D CMPs, whereas the blue line depicts the one from the proposed matching algorithm. As can be seen, the blue line has an abrupt surge around 95°C that leads to a higher thermal yield since a larger proportion of dies can stay within the peak temperature constraint. Under the thermal constraint of 100°C, the thermal yield improves to 62.9% compared to the original thermal yield



(a) peak temperature distribution (b) cumulative distribution function (CDF)

Fig. 13. Improvement after die matching.



(a) peak temperature distribution (b) cumulative distribution function (CDF)

Fig. 14. Improvement after matching and restacking.

of 51.4%. Therefore, a total of 11.5% thermal yield improvement is achieved. We want to again stress that this die matching technique can be applied to both symmetric and asymmetric designs and is thus more general than the proposed tier restacking.

5.5. Overall Thermal Yield Improvements

Since the proposed die matching and tier restacking are orthogonal to each other and thus can be used in synergy, we evaluate and demonstrate their combined effects in this section. Figure 14(a) provides the peak temperature distributions of randomly matched and randomly stacked 3D CMPs (depicted as red) and of the 3D CMPs after applying the proposed matching and restacking (depicted as blue). Compared to the original (red) distribution, the proposed framework significantly reduces both the standard deviation and the mean by 77% (from 9.38°C to 2.16°C) and 8.1% (from 101.55°C to 93.89°C), respectively. Therefore, the new peak temperature distribution is not only shifted to the left but also centered at the mean, as Figure 14(a) suggests.

Figure 14(b) provides the CDFs of these two distributions. It is clear that the proposed framework can drive the CDF to 100% very quickly, which in turn dramatically improves the thermal yield. Under the peak temperature constraint of 100°C, the

thermal yield increases to 99.1% compared to the original yield of 51.4%. Also, we want to point out that, once the leakage measurements are obtained at test time, these two techniques can be used in synergy to improve the post-fabrication thermal yield without introducing any design overhead or extra manufacturing costs. All these aforementioned results confirm the effectiveness of the proposed methods.

6. CONCLUSION

In this article, we propose a methodology to perform statistical thermal evaluation for 3D ICs showing that such systems are much more susceptible to process variations than their 2D counterparts. 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 peak temperature and improves the thermal yield. To further maximize the thermal yield, a thermally-aware die matching is proposed and applied in synergy with the proposed tier restacking. This holistic framework significantly reduces the standard deviation and the mean by 77% and 8.1%, respectively, for the maximum operating temperature distribution of a 3D CMP. Under the peak temperature constraint of 100°C, the proposed framework improves the thermal yield from 51% to 99%.

REFERENCES

S. Alam, R. Jones, S. Pozder, and A. Jain. 2009. Die/wafer stacking with reciprocal design symmetry (RDS) for mask reuse in three-dimensional (3D) integration technology. In *Proceedings of the 10th International Symposium on Quality of Electronic Design (ISQED'09)*. 569–575.

B. Black, M. Annavaram, N. Brekelbaum, J. Devale, L. Jiang, G. Loh, D. McCaule, P. Morrow, D. Nelson, D. Pantuso, et al. 2006. Die stacking (3D) microarchitecture. In *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'06)*. 469–479.

D. Brooks, R. Dick, R. Joseph, and L. Shang. 2007. Power, thermal, and reliability modeling in nanometer-scale microprocessors. *IEEE Micro* 27, 3, 49–62.

D. Brooks, V. Tiwari, and M. Martonosi. 2000. Wattch: A framework for architectural-level power analysis and optimizations. *ACM SIGARCH Comput. Archit. News*. 28, 83–94.

D. Burger and T. Austin. 1997. The simplescalar tool set, version 2.0. *ACM SIGARCH Comput. Archit. News* 25, 3, 13–25.

M. Bushnell and V. Agrawal. 2000. *Essentials of Electronic Testing for Digital, Memory, and Mixed-Signal VLSI Circuits*. Vol. 17, Springer.

J. Butts and G. Sohi. 2000. A static power model for architects. In *Proceedings of the 33rd Annual ACM/IEEE International Symposium on Microarchitecture*. 191–201.

K. Chae and S. Mukhopadhyay. 2012. Tier-adaptive-voltage-scaling (TAVS): A methodology for post-silicon tuning of 3D ICS. In *Proceedings of the 17th Asia and South Pacific Design Automation Conference (ASP-DAC'12)*. 277–282.

K. Chakraborty and S. Roy. 2010. Rethinking threshold voltage assignment in 3D multicore designs. In *Proceedings of the 23rd International Conference on VLSI Design (VLSID'10)*. 375–380.

L. Cheng, P. Gupta, C. Spanos, K. Qian, and L. He. 2011. Physically justifiable die-level modeling of spatial variation in view of systematic across wafer variability. *IEEE Trans. Comput.-Aided Des. Integr. Circ. Syst.* 30, 3, 388–401.

J. Cong and Y. Zhang. 2005. Thermal via planning for 3-D ICS. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD'05)*. 745–752.

S. Dighe, S. Vangal, P. Aseron, S. Kumar, T. Jacob, K. Bowman, J. Howard, J. Tschanz, V. Erraguntla, N. Borkar, et al. 2011. Within-die variation-aware dynamic-voltage-frequency-scaling with optimal core allocation and thread hopping for the 80-core teraflops processor. *IEEE J. Solid-State Circ.* 46, 1, 184–193.

J. Donald and M. Martonosi. 2006. Techniques for multicore thermal management: Classification and new exploration. *ACM SIGARCH Comput. Archit. News* 34, 2, 78–88.

T. Ebi, M. Faruque, and J. Henkel. 2009. Tape: Thermal-aware agent-based power econommulti/many-core architectures. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design-Digest of Technical Papers (ICCAD'09)*. 302–309.

C. Ferri, S. Reda, and R. Bahar. 2007. Strategies for improving the parametric yield and profits of 3D ICS. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD'07)*. 220–226.

C. Ferri, S. Reda, and R. Bahar. 2008. Parametric yield management for 3D ICS: Models and strategies for improvement. *ACM J. Emerg. Technol. Comput. Syst.* 4, 4, 19.

S. Garg and D. Marculescu. 2009. System-level process variability analysis and mitigation for 3D mp-socs. In *Proceedings of the Design, Automation, and Test in Europe Conference and Exhibition (DATE'09)*. 604–609.

B. Goplen and S. Sapatnekar. 2005. Thermal via placement in 3D ICS. In *Proceedings of the International Symposium on Physical Design*. ACM Press, New York, 167–174.

W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. Stan. 2006. Hotspot: A compact thermal modeling methodology for early-stage VLSI design. *IEEE Trans. VLSI Syst.* 14, 5, 501–513.

A. Ivankovic, B. Vandevelde, K. Rebibis, A. La Manna, G. Van Der Plas, V. Cherman, E. Beyne, I. De Wolf, and D. Vandepitte. 2011. Thermo-mechanical impact of the underfill-microbump interaction in 3D stacked integrated circuits. In *Proceedings of the 13th IEEE Electronics Packaging Technology Conference (EPTC'11)*. 34–38.

D. Juan, Y. Chuang, D. Marculescu, and Y. Chang. 2012a. Statistical thermal modeling and optimization considering leakage power variations. In *Proceedings of the Design, Automation, and Test in Europe Conference and Exhibition (DATE'12)*. 605–610.

D. Juan, H. Zhou, D. Marculescu, and X. Li. 2012b. A learning-based autoregressive model for fast transient thermal analysis of chip-multiprocessors. In *Proceedings of the 17th Asia and South Pacific Design Automation Conference (ASP-DAC'12)*. 597–602.

M. Jung, J. Mitra, D. Z. Pan, and S. K. Lim. 2014. TSV stress-aware full-chip mechanical reliability analysis and optimization for 3D IC. *Comm. ACM* 57, 1, 107–115.

R. Kessler. 1999. The alpha 21264 microprocessor. *IEEE Micro* 19, 2, 24–36.

R. Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*. Vol. 2. 1137–1145.

E. Kursun and C. Cher. 2009. Temperature variation characterization and thermal management of multicore architectures. *IEEE Micro* 29, 1, 116–126.

Y. Liu, R. Dick, L. Shang, and H. Yang. 2007. Accurate temperature-dependent integrated circuit leakage power estimation is easy. In *Proceedings of the Design, Automation, and Test in Europe Conference*. 1526–1531.

G. Loi, B. Agrawal, N. Srivastava, S. Lin, T. Sherwood, and K. Banerjee. 2006. A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy. In *Proceedings of the 43rd Annual Design Automation Conference*. ACM Press, New York, 991–996.

S. Ozdemir, Y. Pan, A. Das, G. Memik, G. Loh, and A. Choudhary. 2010. Quantifying and coping with parametric variations in 3D-stacked microarchitectures. In *Proceedings of the 47th Design Automation Conference*. ACM Press, New York, 144–149.

K. Puttaswamy and G. Loh. 2006. Thermal analysis of a 3D die-stacked high-performance microprocessor. In *Proceedings of the 16th ACM Great Lakes Symposium on VLSI*. ACM Press, New York, 19–24.

S. Reda, A. Si, and R. Bahar. 2009. Reducing the leakage and timing variability of 2D ICS using 3D ICS. In *Proceedings of the 14th ACM/IEEE International Symposium on Low Power Electronics and Design*. ACM Press, New York, 283–286.

Roadmap. 2009. International technology roadmap for semiconductors. Executive summary. http://www.itrs.net/Links/2009ITRS/2009Chapters_2009Tables/2009_ExecSum.pdf.

S. Rusu, S. Tam, H. Muljono, D. Ayers, J. Chang, B. Cherkauer, J. Stinson, J. Benoit, R. Varada, J. Leung, et al. 2007. A 65-nm dual-core multithreaded Xeon l3 cache. *IEEE J. Solid-State Circ.* 42, 1, 17–25.

J. Sartori, A. Pant, R. Kumar, and P. Gupta. 2010. Variation-aware speed binning of multi-core processors. In *Proceedings of the 11th International Symposium on Quality Electronic Design (ISQED'10)*. 307–314.

K. Skadron, M. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan. 2004. Temperature-aware microarchitecture: Modeling and implementation. *ACM Trans. Archit. Code Optim.* 1, 1, 94–125.

A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza. 2010. 3D-ice: Fast compact transient thermal modeling for 3D ics with inter-tier liquid cooling. In *Proceedings of the International Conference on Computer-Aided Design*. IEEE Press, 463–470.

A. Vincenzi, A. Sridhar, M. Ruggiero, and D. Atienza. 2011. Fast thermal simulation of 2d/3d integrated circuits exploiting neural networks and gpus. In *Proceedings of the 17th IEEE/ACM International Symposium on Low-Power Electronics and Design*. IEEE Press, 151–156.

W. Zhao and Y. Cao. 2007. Predictive technology model for nano-CMOS design exploration. *ACM J. Emerging Technol. Comput. Syst.* 3, 1, 1.

C. Zhuo, D. Sylvester, and D. Blaauw. 2010. Process variation and temperature-aware reliability management. In *Proceedings of the Design, Automation, and Test in Europe Conference.* 580–585.

Received December 2012; revised April 2014; accepted April 2014