Integrated Thermal and Power Delivery Network Co-Simulation Framework for Single-Die and Multi-Die Assemblies

Yang Zhang, *Student Member, IEEE*, and Muhannad S. Bakir, *Senior Member, IEEE*

***Abstract*— This paper presents a thermal and power delivery network (PDN) co-simulation framework for single-die and emerging multi-die configurations that incorporates the interactions between temperature, supply voltage, and power dissipation. The temperature dependencies of wire resistivity and leakage power are considered, and the supply voltage dependencies of power dissipation are modeled. Starting with a reference power dissipation, the framework is capable of evaluating the temperature distribution and PDN noise simultaneously and eventually updating the power dissipation based on the thermal and supply voltage distributions. The simulation results of an example two-tier 3-D stack show that prior models considering only part of the interactions between temperature, supply voltage, and power dissipation have a maximum error of 7.66%, 9.79%, and 4.64% in IR-drop, transient power supply noise, and temperature, respectively.**

***Index Terms*— 3-D ICs, di/dt noise, IR-drop, power delivery networks (PDNs), thermal challenges.**

I. Introduction

EMERGING applications such as deep learning, data mining, and Internet of Things present performance and communication bandwidth challenges to existing integrated circuits (ICs) and computing platforms [1]. To address these challenges, novel computing fabrics have been proposed by integrating general-purpose processors with field-programmable gate array or graphics processing unit accelerators [2], [3] along with high density stacked memory [4] to increase system performance and energy efficiency. Driven by the need of these emerging computing fabrics, a number of integration solutions are proposed to provide higher bandwidth and parallelism to boost the computing speedup while increasing energy efficiency [5], [6], such as 2.5-D and 3-D integration technologies [7]. However, as the trend continues to integrate more dice in a single package, the total power density is expected to increase beyond 100 W/cm^2 [8]; the impedance of the power delivery network (PDN) will be larger, and air cooled heat sinks (ACHSs) will become incapable of cooling the whole system without keeping much of the silicon dark. Therefore, in spite of the bandwidth and power efficiency benefits brought by 2.5-D and 3-D ICs, there are thermal and power delivery challenges that are potential showstoppers.

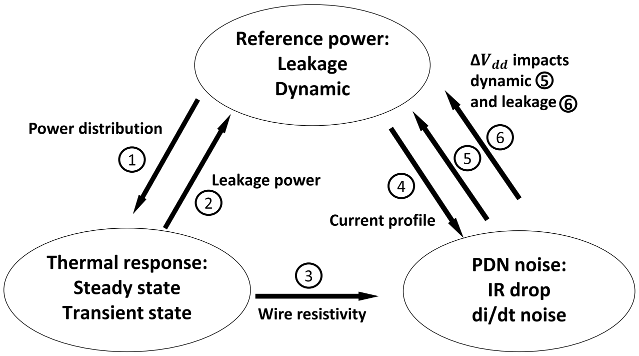


Fig. 1. Interactions between temperature distribution, PDN noise, and power dissipation.

Fig. 1 shows the dependencies between power dissipation, temperature, and PDN noise. The temperature impacts the leakage power and the power grid resistivity. Power dissipation determines the source current of the chip and is also the excitation of the PDN noise. Conversely, the power supply voltage impacts both leakage and dynamic power. Without considering the interactions between each of the components in Fig. 1, the results of the standalone models are inaccurate. For example, Su *et al.* [9] noted that the leakage power was underestimated by as much as 30% without including the impact of temperature and power supply voltage. Hence, it is essential to build a thermal and PDN noise co-simulation framework to answer "what-if" type questions for design space exploration for 2.5-D and 3-D ICs in early design stages.

Prior work focused on either developing the individual thermal [10] or PDN models [11] or studying parts of the interactions [12] shown in Fig. 1. There are no co-simulation frameworks capable of performing steady-state and transient-state analysis on thermal and PDN noise while incorporating the impact of their variation on power dissipation. Xie and and Swaminathan [13] only studied the interactions between thermal and IR-drop and did not consider the interactions with power dissipation. Although Su *et al.* [9] investigated the temperature and supply voltage dependencies of power dissipation, the thermal impact on wire resistivity and the transient-state analysis were not included.

Manuscript received January 7, 2016; revised August 4, 2016; accepted November 3, 2016. Date of publication February 14, 2017; date of current version March 14, 2017. This work was supported by the National Science Foundation under Grant CCF-1302297. Recommended for publication by Associate Editor F. H. Al-Hawari upon evaluation of reviewers' comments.

The authors are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: steven.zhang@gatech.edu; muhannad.bakir@mirc.gatech.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCPMT.2017.2657382

In this paper, we propose a framework to simultaneously study the temperature, PDN noise, power dissipation, and the interactions between them for both steady-state and transient-state analysis. The thermal model is based on finite volume method, the PDN model is based on finite difference method, and the interaction models are based on the thermal and supply voltage dependencies of power dissipation and the thermal dependencies of wire resistivity. The initial reference power is an input from *McPat* [14] and using our thermal-power and PDN-power models, we update the power dissipation until the iterations are converged. There are two loops in the framework where the outer loop iterates the thermal-power models and the inner loop iterates the PDN-power models. The major contributions of this paper are as follows.

1) To the best of the authors' knowledge, this is the first work that closes the loop shown in Fig. 1 and studies the power dissipation, thermal, and PDN simultaneously.
2) This paper quantifies the error between the models considering part of the interactions and the integrated models and demonstrates the significance of the integrated co-simulation framework of power dissipation, thermal, and PDN.
3) This is the first work that studies the full-chip thermal impact on Ldi/dt noise: prior work mainly studied IR-drop. This paper proves the validity of using a steady-state thermal profile to study Ldi/dt noise.
4) Lastly, the integrated co-simulation framework is capable of performing fast simulations to answer what-if type questions in early design stages as well as being able to conduct detailed studies on the impact of technology parameters.

The rest of this paper is organized as follows. In Section II, we present the simulation flow. Next, we discuss the detailed methodologies and the models used in the framework. Section IV studies an example 3-D stack and compares the results with the models that do not consider all the interactions. Lastly, Section V concludes this paper.

II. Simulation Flow

We propose a simulation framework for steady-state and transient-state analysis. We assume an architectural tool has already provided the reference power of the chip under uniform temperature and an ideal power supply voltage through the initial power simulations. Since our focus is the impact of supply voltage and temperature, we assume other parameters such as clock frequency remain constant. In the following iterations, the power dissipation is updated by the power models instead of calling the power simulator at every iteration. Using the initial power, we start the simulation flow and perform the thermal and power supply noise simulations. At the end of the simulations, the three metrics become consistent with each other within our interaction models. The simulator is implemented using MATLAB because dense matrix operations and calculations are required in the flow.

A. Steady-State Analysis

For steady-state analysis, first, the reference power is used to obtain the temperature distribution. With the thermal results, the PDN grid resistance within the chip is updated. The supply voltage profile is then simulated using the updated PDN grid resistance and source current distribution. Next, the leakage and dynamic powers are updated based on the thermal and supply voltage values. For this step, the power and supply voltage form a loop and Newton method is used to accelerate the convergence rate. After this loop is complete, the thermal profile is updated and checked for whether the convergence has been reached. If not, the simulation restarts. The simulator finally returns the thermal profile, PDN noise distribution, and the updated reference power results of the system. Fig. 2 shows the whole simulation flow. Because the thermal conductivity of common materials such as copper, silicon, and silicon dioxide is usually constant in the typical temperature range of IC operation, the thermal impact on their material properties is neglected.

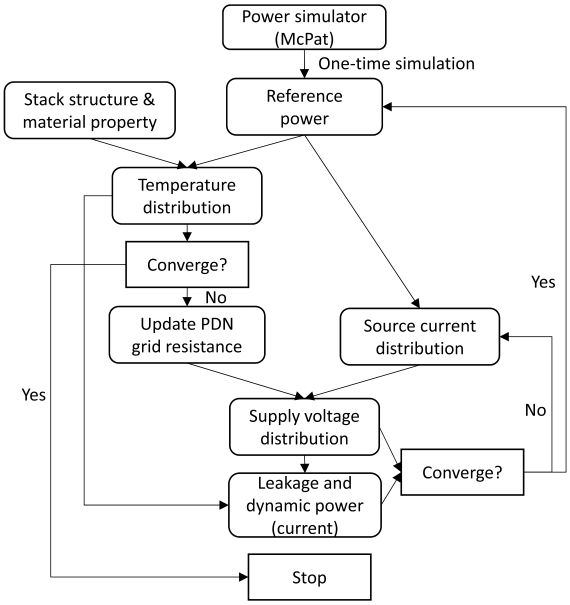

Power simulator (McPat)
One-time simulation
Stack structure & material property
Reference power
Temperature distribution
Converge?
No
Update PDN grid resistance
Source current distribution
Yes
Yes
Supply voltage distribution
No
Converge?
Leakage and dynamic power (current)
Stop

Fig. 2. Steady-state simulation flow.

B. Transient-State Analysis

The thermal response time of a typical single or multi-die package with either conventional ACHS or microfluidic heat sink is much larger than the response time of the PDN. The thermal response time is at least in the milliseconds range [15] and the response time of PDN is within the circuit switching frequency (in the nanoseconds or microseconds range). Therefore, in the PDN simulation timescale (up to microseconds [16]), the thermal profile remains constant. The validation is discussed in Section III.

Based on the above discussion, a one-time steady-state or transient thermal simulation is initially performed to obtain a temperature profile as an input. For this step, one option is to perform a steady-state simulation using a user-defined reference power representing the average power. Another option is to perform cycle-granularity transient thermal simulation from the very beginning to obtain the final thermal profile. Next, we start the transient simulation of the PDN using the

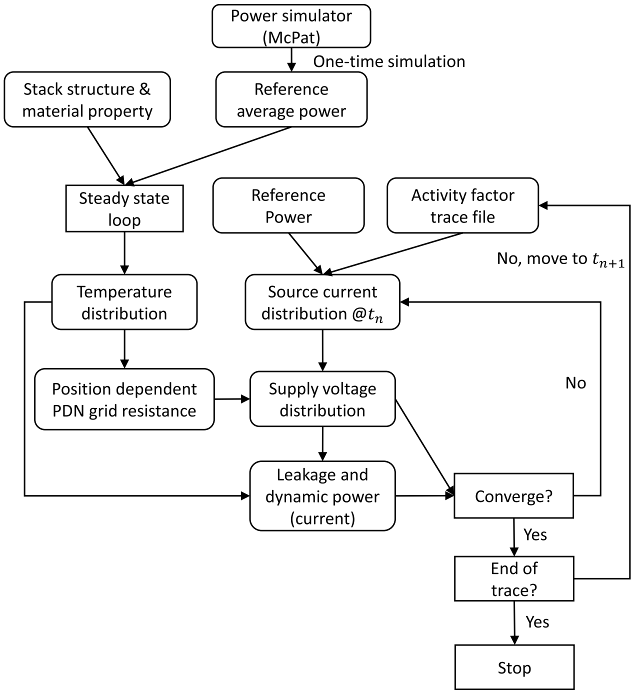

Power simulator (McPat)
One-time simulation
Stack structure & material property
Reference average power
Steady state loop
Reference Power
Activity factor trace file
No, move to t_{n+1}
Temperature distribution
Source current distribution @t_n
No
Position dependent PDN grid resistance
Supply voltage distribution
Leakage and dynamic power (current)
Converge?
Yes
End of trace?
Yes
Stop

Fig. 3. Transient-state simulation flow.

thermal inputs. In each time-step, similar power-PDN iterations are performed as in the steady-state analysis. Here, we assume leakage and dynamic powers (currents) are changing instantaneously as supply voltage changes [17]. When the loop of the current time-step is converged, the simulation of the next time-step is started until the end of the trace. The simulator finally returns the transient PDN noise distribution and updated power results. The transient analysis flow is shown in Fig. 3.

C. Algorithm

The pseudocode for the steady-state analysis is shown in Algorithm 1. Transient-state analysis is similar but includes the thermal capacitive and electrical capacitive/inductive elements, therefore we do not show it here. Because Newton method is used when solving the PDN-power loop, the power update model needs to calculate the partial derivatives over supply voltage (lines 11–15). The thermal loop (lines 20–28) performs fixed point iterations and for the PDN-power loop (line 22–25), Newton method is used.

Algorithm 1 Steady-State Simulation

```
1: function INITIALIZATION                          ▷ parse input tables
2:     $G_{elec} \leftarrow electrical\ conductance\ matrix$
3:     $G_{ther} \leftarrow thermal\ conductance\ matrix$
4:     $I_{leak} \leftarrow reference\ leakage\ current$
5:     $I_{dyna} \leftarrow reference\ dynamic\ current$
6:     $V_{node} \leftarrow V_{ref}$                         ▷ reference voltage
7:     $T_{node} \leftarrow T_{ref}$                 ▷ i.e. ambient temperature
8:     $P_{tot} \leftarrow (I_{leak} + I_{dyna}) \cdot V_{node}$
9: end function
10:
11: function PDN_SOLVER($G, I, V$)          ▷ solve $-G \cdot V = I$
12:     $G_{new} = G + \partial I_{tot}/\partial V$               ▷ Newton method
13:     $\delta I \leftarrow G \cdot V + I$
14:     $output \leftarrow V + \delta I/G_{new}$
15: end function
16:
17: procedure STEADY_STATE
18:     Initialization()
19:     $V,\ T \leftarrow V_{node},\ Therm\_solver(G_{ther}, P_{tot})$
20:     while $|T_{node} - T| > \epsilon \cdot T$ do
21:         $G_{elec\_new} \leftarrow Resistivity\_Update(T, G_{elec})$
22:         while $\left|G_{elec\_new} \cdot V + I_{tot}\right| > \epsilon$ do
23:             $I_{tot} \leftarrow Current\_Update(V, T, I_{leak}, I_{dyna})$
24:             $V \leftarrow PDN\_solver(G_{elec\_new}, I_{tot}, V)$
25:         end while
26:         $T_{node} \leftarrow T$
27:         $T \leftarrow Therm\_solver(G_{ther}, P_{tot})$
28:     end while
29:     $P_{tot} \leftarrow I_{tot} \cdot V$
30:     $output \leftarrow T, V, P_{tot}$
31: end procedure
```

III. Modeling Methodology

In this section, the thermal, PDN and power update models are presented. The formulation of the interactions between each component is described. The explicit interactions considered are temperature-wire resistivity, temperature and supply voltage-leakage power, and supply voltage-dynamic power.

A. Thermal Model

By using nonconformal grids in the chip and package substrate and the weighted thermal conductivity calculation in the chip domain, a steady-state thermal model using the finite volume method is implemented and validated against *ANSYS* with a maximum error of below 1% [18]. The backward Euler scheme is used to implement the transient thermal analysis. The implemented transient thermal model is validated against *ANSYS* with a maximum error of below 1%. The detailed validation can be found in the Appendix.

The thermal model has three inputs: first, the geometry information of the single-chip 2.5-D module or 3-D stack, second, the material property of each layer, and third, the reference power information. The power granularity can be block level or transistor level. The formulation of the thermal model is shown as

$$-G \cdot T_{n+1} + C \cdot \frac{T_{n+1} - T_n}{\delta t} = P_{n+1}(T) + H \cdot T_{\text{amb}} \quad (1)$$

where T_{n+1} and T_n are the temperatures of the current (to be solved) and the previous (known) time-steps, respectively. G is the thermal conductance matrix, C is the heat capacity matrix, and $P(T)$ is the power excitation whose leakage component is temperature dependent. H is a diagonal matrix that represents the convective boundary condition and T_{amb} is the ambient temperature. The temperature time difference term is only applicable for the transient analysis and is not applicable for the steady-state analysis.

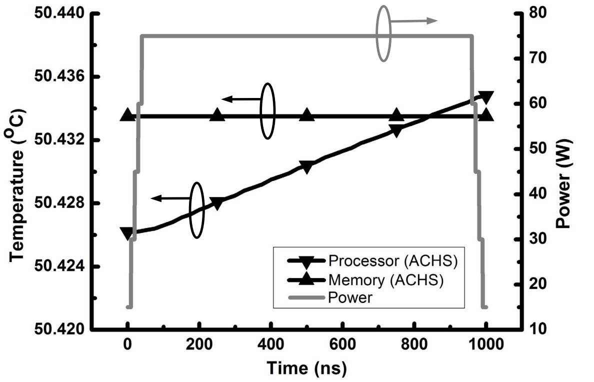


Fig. 4. Validation of stable temperature assumption in microsecond scale.

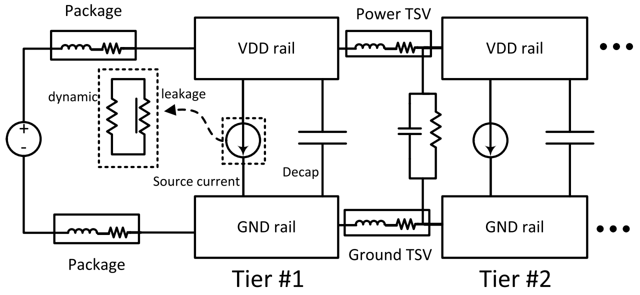


Fig. 5. Block diagram of the PDN structure. The distributed P/G rail is abstracted for visualization.

As previously stated in Section II-B, due to the large response time, the temperature profile remains constant in the timescale of microseconds even though the power experiences a sharp change. We simulated a test case consisting of a 3-D stack as shown in Fig. 7 (Section IV) to validate this. The thermal specifications can be found in Section IV. Fig. 4 shows that there is a nominal change (less than 0.1%) in the maximum temperature of the processor and memory dice when the processor power changes dramatically (memory power remains the same). The thermal profile of the chip is also checked, and it does not undergo any change.

B. PDN Model

We modified and extended the frequency domain-based PDN model described in [16]. In that work, the authors assumed a uniform distributed power dissipation and analyzed a small portion of the distributed PDN using Laplace transform. However, in reality, the power dissipation, power/ground (P/G) pads, on-chip decaps, and TSVs can be nonuniformly distributed. Therefore, we propose to use the finite difference method to evaluate the distributed on-chip PDN using nonuniform power dissipation [19]. The block diagram of the PDN network is shown in Fig. 5. Here, the distributed P/G resistance network is abstracted for visualization. The detailed structure of the VDD/GND rail is a distributed wire network. As our study focuses on the on-chip PDN modeling, we use a simplified package model where each P/G port is connected to a lumped resistor and inductor pair. The trapezoid scheme is used to formulate the transient finite difference equation [20]. The formulation is shown as follows:

$$G(T)\cdot\frac{V_{n+1}+V_n}{2}+C\cdot\frac{V_{n+1}-V_n}{\delta t} = S\cdot\frac{I_{n+1}(V,T)+I_n(V,T)}{2} \tag{2}$$

where $G(T)$ is the PDN grid conductance matrix, which is temperature dependent. C is the matrix reflecting the capacitive and inductive elements. $I(V,T)$ is the source current which is dependent on temperature (due to leakage portion) and supply voltage (due to both leakage and dynamic portions). S is the input selector matrix. The temperature-dependent PDN grid resistivity is described as

$$\rho = \rho_0(1+\alpha(T-T_0)) \tag{3}$$

where ρ_0 is the resistivity under reference temperature T_0 and α is the temperature coefficient of resistivity. The model is validated against HSPICE with a maximum error of less than 1% [19].

C. Power Update Model

In this paper, we do not aim to develop a detailed power analysis model such as *McPat* [14], but instead pursue a power update model based on distributed temperature and supply voltage. Therefore, we focus on the impact of supply voltage and temperature, while other parameters such as clock frequency are assumed to be constant. We begin with the power results from *McPat* and by considering the supply voltage and thermal variation through the whole chip, we update the value of power dissipation.

It is assumed that the power of each functional block consists of leakage and dynamic powers. In the PDN model, the power is converted into source current, as shown in Fig. 5. Prior work assumed that the source current (I) can be calculated by $I = P/V_{\text{dd}}$, where P is power dissipation and V_{dd} is the value of ideal supply voltage [11]. This simple method does not consider the power dependence on supply voltage and temperature. Instead, modeling the source current as two resistors (the dashed inlet box in Fig. 5) whose values are a function of supply voltage and temperature will capture the dependencies [21].

1) Leakage Power: The relationship between reference leakage power $P_{\text{leak\_ref}}$ and leakage current source $I_{\text{leak\_ref}}$, is expressed as follows:

$$P_{\text{leak\_ref}} = I_{\text{leak\_ref}}\cdot V_{\text{dd}} \tag{4}$$

where V_{dd} is the ideal supply voltage of each power grid. Based on the fitting method [9], the actual leakage current of a node, $I_{\text{leak\_act}}$, can be generalized as

$$I_{\text{leak\_act}} = I_{\text{leak\_ref}}\cdot f(V,\ T) \tag{5}$$

where $f(V,\ T)$ is the fitted function of supply voltage and temperature of the node in the chip. In this paper, we propose to use a 2-D piecewise linear model for $f(V,\ T)$. The first advantage of a 2-D piecewise linear model is to cover a wide range of voltage and temperature values ([9] only covers a small range around the reference point). Second, the partial derivative of leakage current over temperature and voltage can be easily calculated so that Newton method can be implemented to accelerate the whole simulation. For a target circuit, the voltage-temperature plane is uniformly meshed based on the number of sampling points. Next, for each sampling point, i.e., $(V_i,\ T_i)$, we run HSPICE simulations and collect the data. The leakage of an arbitrary point $(V,\ T)$ can then be calculated using $f(V,\ T)$ as

$$V_i \leq V \leq V_{i+1};\quad T_i \leq T \leq T_{i+1} \tag{6}$$

$$\xi = \frac{V - V_i}{V_{i+1} - V_i};\quad \eta = \frac{T - T_i}{T_{i+1} - T_i} \tag{7}$$

$$wt = \begin{bmatrix} \xi * \eta \\ (1-\xi) * \eta \\ (1-\eta) * \xi \\ (1-\xi) * (1-\eta) \end{bmatrix} I_{\text{node}} = \begin{bmatrix} I_{\text{leak}}(V_i,\ T_i) \\ I_{\text{leak}}(V_{i+1},\ T_i) \\ I_{\text{leak}}(V_i,\ T_{i+1}) \\ I_{\text{leak}}(V_{i+1},\ T_{i+1}) \end{bmatrix} \tag{8}$$

$$f(V,T) = \frac{wt^T \cdot I_{\text{node}}}{I_{\text{leak\_ref}}}. \tag{9}$$

Although it is difficult to use just one circuit and apply the characterization results to other circuits, the fitted $f(V,\ T)$ of an inverter array is quite accurate to use based on the results of [21]. In this paper, we use 50 stage inverter pairs (for each inverter pair, the input of the first inverter is connected to V_{dd} and the input of the second is connected to ground, and the output of both inverters is floated). PTM-MG 20 nm (HP) model [22] is used for *HSPICE* simulation and the parameter ranges for supply voltage and temperature are (0.7 and 1.0 V) and (25 °C and 110 °C), respectively, the reference voltage is 0.9 V, and the reference temperature is 100 °C, a pessimistic temperature as most of the IC design tools assume.

Fig. 6(a) shows the surface response of the model with eight sampling points. We generate 1600 random data points in the (V, T) plane for validation. Fig. 6(b) shows that the model accuracy increases as the number of sampling points increases and the error drops below 5% when the number of sampling points is larger than 13.

2) Dynamic Power: The reference dynamic power $P_{\text{dyna\_ref}}$ is expressed as

$$P_{\text{dyna\_ref}} = \alpha \cdot C \cdot f \cdot {V_{\text{dd}}}^2 \propto {V_{\text{dd}}}^2 \tag{10}$$

where α is the activity factor, f is the frequency, and C is the total capacitance. The power grid actually gets a supply voltage of $V_{\text{dd\_act}}$ instead of V_{dd} when calculating reference power, thus the dynamic power becomes

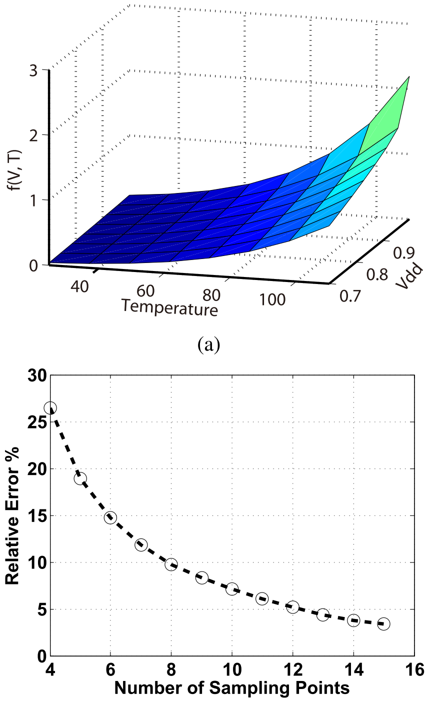
$$P_{\text{dyna\_act}} = P_{\text{dyna\_ref}} \cdot \frac{{V_{\text{dd\_act}}}^2}{{V_{\text{dd}}}^2}. \tag{11}$$

By converting the dynamic power into a current source, we have the dynamic current update model [9], [21]

$$I_{\text{dyna\_act}} = P_{\text{dyna\_ref}} \cdot \frac{V_{\text{dd\_act}}}{{V_{\text{dd}}}^2}. \tag{12}$$

IV. Model Comparison

In this section, we use a 3-D processor-on-memory stack as an example to demonstrate the capability and accuracy of the modeling work.



(a)

(b)

Fig. 6. 2-D piecewise linear model (a) response surface with eight sampling points (b) maximum error of models with different number of sampling points.

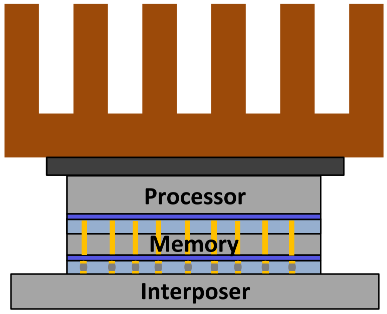


Fig. 7. 3-D-IC example: processor on memory stack.

TABLE I
Parameters for Thermal Model

	Conductivity W/mK	Thickness μm
TIM	3	25
Memory die	149	100
Underfill layer	0.9	25
Processor die	149	100
Micro-bump	60	25
Interposer	149	200
Copper	400	N/A
SiO_2	1.38	N/A

A. Specification

The stack we evaluate is shown in Fig. 7. The processor is placed on top of the memory for thermal considerations. The thickness of each layer and thermal conductivity of each material are shown in Table I. The reference power maps of the memory and processor dice are shown in Fig. 8(a) and (b) [18].

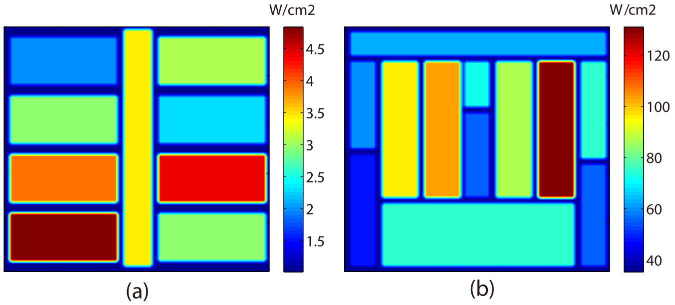


Fig. 8. Reference power maps. (a) Memory die (2.82 W). (b) Processor die (74.49 W).

TABLE II
PARAMETERS FOR PDN MODEL

	value
P/G TSV diameter	5 μm
on-die decap %	10% of the area
# of P/G TSV	676/625
Package inductance	0.1 nH
Package resistance	0.0107 Ω
P/G wire thickness	5 μm
P/G wire width	3.33 μm
P/G wire pitch	30 $\mu\ m$

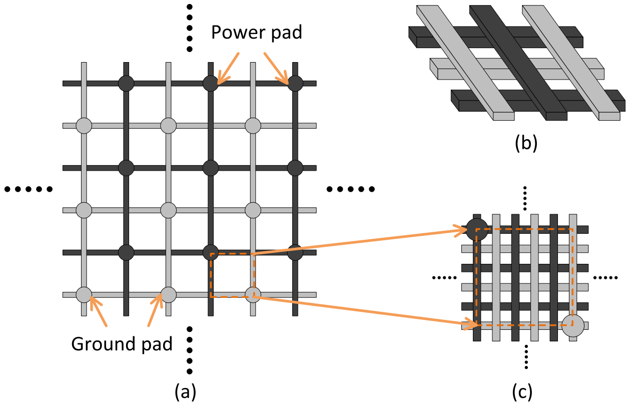


Fig. 9. On-die PDN structure. (a) P/G pads with wires. (b) Interleaved structure of P/G wires. (c) Dense PDN wires between P/G pads.

The reference temperature is 100 °C, and the supply voltage is 0.9 V for 22-nm multigate ICs [22]. Based on the simulations from *McPat* [14], under the reference temperature, 20% of the total power is leakage. The chip size is assumed to be 1 × 1 cm^2 and the interposer size is assumed to be 2 × 1.5 cm^2. The heat spreader is assumed to be 4.5 × 3.5 cm^2 and the ACHS is converted into a boundary condition of 0.24 W/K [23]. All the other surfaces are adiabatic. The ambient temperature is assumed to be 38 °C.

The parameters for the PDN model are shown in Table II. In this paper, we focus on the global on-die PDN and it is assumed to consist of the top two metal layers. Each metal layer is assumed to be 5 μm thick. Fig. 9 shows the detailed geometry and configuration of the interleaved global PDN [19]. Besides P/G TSVs, we assume there are 10 000 signal TSVs with 5 μm diameter and 0.5-μm-thick liner. The TSVs are assumed to be uniformly distributed because of the high-power processor. The reference source current is calculated using (5) and (12). For the transient PDN analysis, we consider a worst case scenario where both dice switch from full sleep mode to peak power dissipation (power map shown in Fig. 8) in a rise time of 100 ps and remain in peak power mode for 50 ns. Based on Section III, when performing transient simulation, the thermal maps do not change over the timescale of the PDN analysis. For this case, we assume both dice have been in the peak power state long enough so that the thermal profiles with the maximum temperatures are reached.

B. Modeling Scenarios

In this section, we compare a number of models to establish the benefits of the proposed work. The first model (denoted as standalone model) provides fixed input power maps for the thermal and PDN models. Standalone thermal and PDN analysis are performed (constant reference temperature is used throughout the chip for PDN analysis). In the second model, we consider the interactions between power dissipation and PDN only (PDN-power model). The thermal effects on power dissipation and PDN wires are not included (same as in the standalone model), but in this case, the final updated power distribution is used to perform thermal analysis.

In the third model, we add the thermal impact on PDN wires to the PDN-power case (denoted as PDN-therm model) [24]. In the fourth model, we consider the thermal impact on both wire resistivity and leakage power, but the interactions between PDN and power dissipation are not included (PDN-therm-leak model) [12].

In the fifth model, we add the interactions between thermal and leakage power to PDN-power model, but the thermal impact on grid resistivity is not included (partial-therm model). Lastly, we include all the interactions shown in Fig. 1 (full-model). The six models are summarized in Table III. In practice, there are usually thermal, power, and PDN constraints for a system. If under a configuration any of the metrics are out of bound, the configuration should be changed to meet the constraints such as lowering the target frequency. To model this, it is necessary to include an integrated power module in the framework. The authors are aware of these constraints, but in this paper, our focus is to present a way to co-simulate these metrics.

C. Model Results

The simulation results are shown in Table IV. To easily compare them, the metrics of each model are normalized to those of the standalone model and are shown in the parentheses.

First, we analyze all the steady-state analysis results (IR-drop, temperature, and power). Comparing standalone and PDN-power model, we observe that the standalone model overestimates all the metrics by about 3%–6%. This is because when the IC is operating, not a single power grid receives the ideal power supply voltage due to IR-drop. Based on (5) and (12), with a lower supply voltage, the actual source

TABLE III

SIMULATION MODEL

Model	Description	Arrows included in Fig. 1
standalone	Power results from *McPat*, individual thermal and PDN simulation	① ④
PDN-power	Interactions between power dissipation and PDN are added to standalone models	① ④ ⑤ ⑥
PDN-therm	Thermal impact on wire resistivity is added to PDN-power case	① ③ ④ ⑤ ⑥
PDN-therm-leak	Impact of PDN and thermal on leakage power and the thermal impact on wire resistivity are added to standalone models	① ② ③ ④ ⑥
partial-therm	Interactions between temperature and leakage power are added to PDN-power case	① ② ④ ⑤ ⑥
full-model	Thermal impact on wire resistivity is added to partial-thermal case	① ② ③ ④ ⑤ ⑥

TABLE IV

RESULTS FOR DIFFERENT DETAILED MODELS

Model	die	noise(mV)		temperature	power(W)	
		IR-drop	Transient	($^\circ C$)	dynamic	leakage
Standalone	Processor	38.79	118.07	92.74	59.59	14.90
	Memory	5.32	83.85	91.31	2.26	0.56
PDN-power	Processor	37.11 (4.33%)	104.35 (11.62%)	89.70 (3.28%)	56.63 (4.97%)	13.91 (6.64%)
	Memory	5.13 (3.57%)	75.19 (10.33%)	88.51 (3.07%)	2.24 (0.88%)	0.56 (0.00%)
PDN-therm	Processor	36.02 (7.14%)	103.36 (12.46%)	89.81 (3.16%)	56.74 (4.78%)	13.95 (6.38%)
	Memory	5.10 (4.14%)	75.47 (11.19%)	88.61 (2.96%)	2.24 (0.88%)	0.56 (0.00%)
PDN-therm-leak	Processor	34.71 (10.52%)	107.57 (8.89%)	87.01 (6.19%)	59.59 (0.00%)	8.31 (44.23%)
	Memory	4.90 (7.89%)	76.50 (8.77%)	85.71 (6.13%)	2.26 (0.00%)	0.31 (44.64%)
partial-therm	Processor	34.41 (11.28%)	97.25 (17.63%)	85.37 (7.95%)	56.89 (4.53%)	7.42 (50.20%)
	Memory	4.74 (10.90%)	69.80 (16.76%)	84.22 (7.76%)	2.24 (0.88%)	0.30 (46.43%)
full-model	Processor	33.05 (14.80%)	96.02 (18.68%)	85.51 (7.80%)	57.02 (4.31%)	7.47 (49.87%)
	Memory	4.71 (11.47%)	70.18 (16.30%)	84.35 (7.62%)	2.24 (0.88%)	0.30 (46.43%)

The relative percentage change in the parentheses is normalized to the results of the standalone model.

current becomes smaller than the reference value, resulting in lower power and as a consequence, the simulated temperature is smaller.

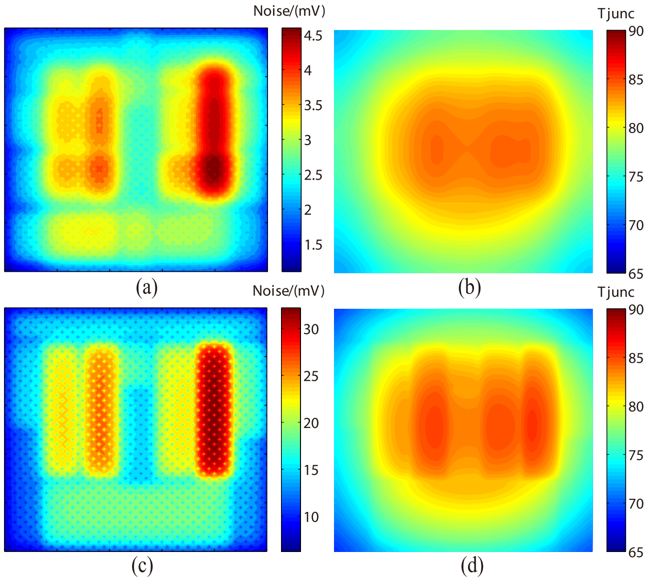
The PDN-therm model is similar to the PDN-power model with the only difference being that the thermal impact on wire resistivity is included. For the PDN-power model, the wire temperature is the reference temperature (100 °C) while the temperature for the PDN-therm model is about 90 °C. Based on (3), the wire resistivity changes by 3.93% for a 10 °C temperature change (temperature coefficient of copper wire is $3.9 \cdot 10^{-3}$/°C). The difference in IR-drop of the processor die between PDN-Power and PDN-thermal models has good agreement with this number.

The PDN-therm-leak model adds the thermal impact on both wire resistivity and leakage power to the standalone model. Due to the significant impact of temperature on leakage power, the leakage estimation becomes more accurate. Thus, the IR-drop and temperature are also closer to the full-model results.

Partial-thermal model includes the thermal impact on leakage power as well as the PDN-power interactions. With these effects adding up, the results become smaller than the first four models: the IR-drop decreases by 11.28%, the temperature decreases by 7.95%, and the dynamic power decreases by 4.53% compared to the standalone model. However, for leakage, it almost drops by half because the actual temperature is lower than the reference temperature (100 °C) and in this temperature range, the leakage has an exponential relationship with temperature, resulting in severe errors. For example, the leakage current at 80 °C is only 53.23% of that at 100 °C based on the leakage power model described in Section III-C.1.

For the full-model, the temperature impact on wire resistivity is included (compared to partial-therm model). As a result, the IR-drop of the full-model becomes slightly lower because of the lower PDN impedance at the simulated temperature (versus the reference temperature). Fig. 10 shows the thermal and IR-drop profiles of both dice using the full-model simulation. There is strong thermal and IR-drop coupling between the two dice due to the uniformly distributed TSVs.

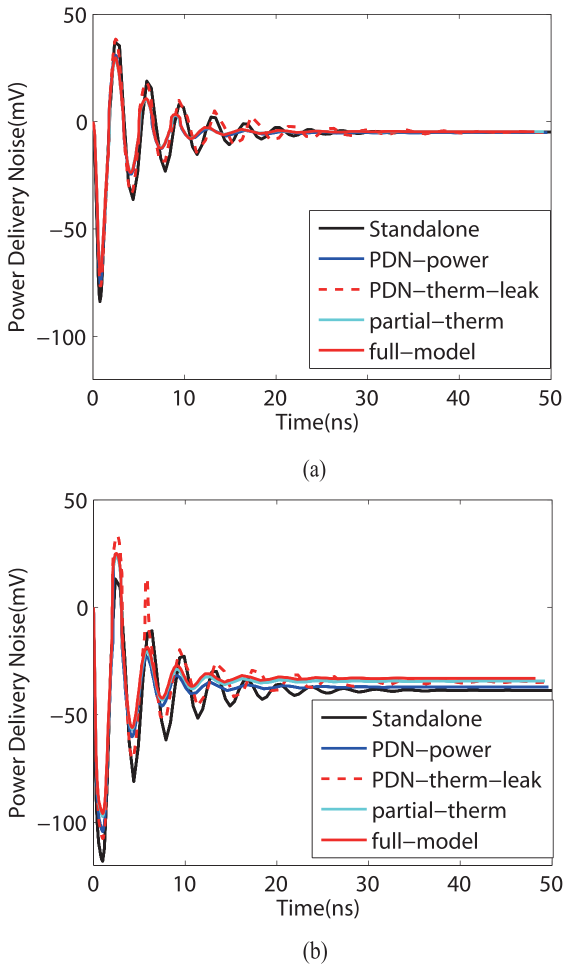
A similar trend is found for the maximum transient power supply noise: the standalone model is 11.62%, 12.46%, 8.89%, 17.63%, and 18.68% higher than PDN-power, PDN-therm, PDN-therm-leak, partial-therm, and full-model models, respectively. To understand the transient PDN noise, we plot the maximum noise over time, as shown in Fig. 11. Not only is the difference in maximum noise large, so is the noise profile. The models that include the interaction between power dissipation and supply voltage predict a relatively faster damping profile (PDN-power, PDN-therm, partial-therm, and full models). We define the damping rate as the amplitude of the second noise valley divided by that of the first noise valley.



(a) (b)

(c) (d)

Fig. 10. Steady-state analysis result of full-model case. (a) IR-drop of memory (4.71 mV). (b) Thermal of memory (84.35 °C). (c) IR-drop of processor (33.05 mV). (d) Thermal of processor (85.51 °C).



(a)

(b)

Fig. 11. Transient power supply noise comparison. (a) Memory die. (b) Processor die. PDN-therm is very similar to PDN-power, thus it is omitted for better visualization.

The relative difference in the damping rate between standalone and PDN-power, PDN-therm, partial-therm, and full models is 20.03%, 20.10%, 18.86%, and 18.89%, respectively. This difference results from the power-PDN negative loop that makes the noise damp faster.

TABLE V

RESULTS AFTER DIFFERENT NUMBER OF ITERATIONS

# iteration	IR drop (mV)	Temperature ($^\circ C$)	dynamic	leakage
1	34.79	92.86	56.90	9.21
2	33.33	86.78	57.00	7.74
3	33.10	85.71	57.01	7.51
4	33.06	85.54	57.02	7.48
5	33.06	85.52	57.02	7.47
6	33.05	85.51	57.02	7.47
7	33.05	85.51	57.02	7.47

In summary, thermal-leakage and PDN-power interactions have a significant impact on steady-state results, and the PDN-power interaction affects the transient PDN noise greatly. However, thermal-PDN interaction has a relatively smaller impact.

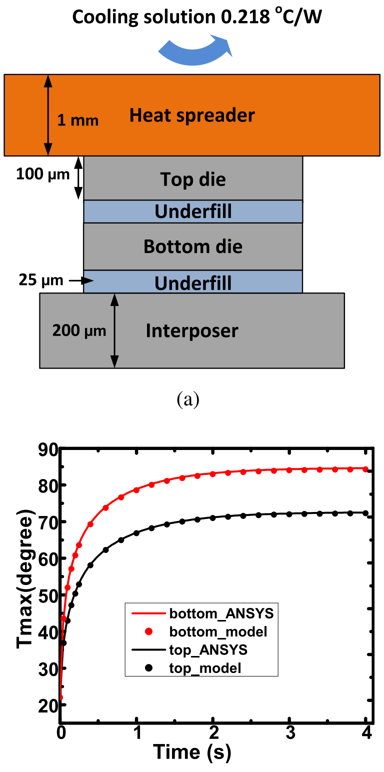
D. Accuracy Improvement Compared to Prior Work

Compared to the PDN-therm model, which only considers PDN-thermal interaction, the full-model achieves an accuracy improvement of 7.66%, 6.22%, and 4.64% for IR-drop, transient PDN noise, and maximum temperature, respectively; compared to PDN-therm-leak model, which includes both PDN-thermal and thermal-leakage interactions, the full-model achieves an accuracy improvement of 4.26%, 9.79%, and 1.61% for IR drop, transient PDN noise, and maximum temperature, respectively.

Several leakage power estimation efforts have proposed a dc analysis framework similar to the partial-therm model [9], [12]. Nevertheless, these two efforts focused on the leakage power. Moreover, a coarse thermal model was used in [12] to reduce the integration complexity, and as a result, the thermal map was not full-chip scale; only one iteration of the integrated analysis was performed in [9], since more iterations did not increase the estimation accuracy of leakage power significantly. However, we find one iteration is not adequate for obtaining accurate results due to the large change of leakage power in the temperature range between 85 and 100 °C. Table V shows the full-model results of the processor die after several iterations. It is observed that at least three iterations are necessary to achieve a relative error less than 1%.

V. CONCLUSION

In this paper, we present a thermal and PDN co-simulation framework incorporating the interactions between temperature, PDN noise, and power dissipation. First, compared to prior work, the proposed models show that when we do not consider the interactions, there is a maximum error of 7.66%, 9.79%, and 4.64% in IR-drop, transient noise, and temperature, respectively. Second, the integrated simulator is capable of performing fast simulations to answer what-if type questions in early design stages as well as being able to conduct detailed studies such as the impact of a wide range of technology parameters and different power delivery architectures. The modeling framework will benefit the architecture and packaging research communities.



(a)

(b)

Fig. 12. Transient thermal validation experiments (a) 2-die 3-D stack (b) transient thermal validation results.

Appendix
Transient Thermal Validation

Fig. 12(a) shows a 3-D stack with two dice used for transient thermal validation against *ANSYS*. Nonuniform power maps with average power densities of 36 and 43.8 W/cm^2 are assigned to the top and bottom dice, respectively. The power maps are similar to Fig. 8(b) with appropriate scaling. The heat spreader size is assumed to be 3 × 3 cm^2 with a cooling of 0.218 °C/W added to the top surface. Other faces of the stack are assumed to be adiabatic. The interposer size is assumed to be 2 × 1.5 cm^2, and the chip size is set as 1 × 1 cm^2. The thickness is labeled in Fig. 12(a). The starting temperature of the whole stack is assumed to be the ambient temperature, which is 22 °C. We add the power excitation from time equal to 0 s and perform the transient thermal analysis from 0 to 4 s. The results of the maximum temperature of each die are shown in Fig. 12(b). The maximum temperature of both dice in each time point matches *ANSYS* results with an error of less than 1%. Moreover, the thermal profiles of each time point are compared, and the maximum error is also less than 1%.

References

[1] Intel. *Developing Solutions for the Internet of Things*, accessed on Feb. 3, 2017. [Online]. Available: http://www.intel.com/content/www/us/en/internet-of-things/white-papers/developing-solutions-for-iot.html

[2] A. Putnam *et al.*, "A reconfigurable fabric for accelerating large-scale datacenter services," in *Proc. 41st ACM/IEEE ISCA*, Jun. 2014, pp. 13–24.

[3] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, and D. Glasco, "GPUs and the future of parallel computing," *IEEE Micro*, vol. 31, no. 5, pp. 7–17, Sep. 2011.

[4] J. Jeddeloh and B. Keeth, "Hybrid memory cube new DRAM architecture increases density and performance," in *Proc. Symp. VLSI Technol. (VLSIT)*, Jun. 2012, pp. 87–88.

[5] J. Cong, K. Guruaj, M. Huang, S. Li, B. Xiao, and Y. Zou, "Domain-specific processor with 3D integration for medical image processing," in *Proc. IEEE Int. Conf. Appl.-Specific Syst., Archit. Processors (ASAP)*, Sep. 2011, pp. 247–250.

[6] J. Power, Y. Li, M. D. Hill, J. M. Patel, and D. A. Wood, "Implications of emerging 3D GPU architecture on the scan primitive," *ACM SIGMOD Rec.*, vol. 44, no. 1, pp. 18–23, Mar. 2015.

[7] D. Liu and S. Park, "Three-dimensional and 2.5 dimensional interconnection technology: State of the art," *ASME J. Electron. Packag.*, vol. 136, no. 1, Feb. 2014, Art. no. 014001, doi: 10.1115/1.4026615.

[8] ITRS. (2013). *International Technology Roadmap for Secmiconductors*. [Online]. Available: http://www.itrs.net/

[9] H. Su, F. Liu, A. Devgan, E. Acar, and S. Nassif, "Full chip leakage estimation considering power supply and temperature variations," in *Proc. Int. Symp. Low Power Electron. Design*, Aug. 2003, pp. 78–83.

[10] A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza, "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2010, pp. 463–470.

[11] R. Zhang, K. Wang, B. H. Meyer, M. R. Stan, and K. Skadron, "Architecture implications of pads as a scarce resource," in *Proc. 41st ACM/IEEE Int. Symp. Comput. Archit. (ISCA)*, Jun. 2014, pp. 373–384.

[12] Y. Liu, R. P. Dick, L. Shang, and H. Yang, "Accurate temperature-dependent integrated circuit leakage power estimation is easy," in *Proc. Conf. Design, Autom. Test Eur.*, Apr. 2007, pp. 1526–1531.

[13] J. Xie and M. Swaminathan, "Electrical-thermal co-simulation of 3D integrated systems with micro-fluidic cooling and Joule heating effects," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 1, no. 2, pp. 234–246, Feb. 2011.

[14] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2009, pp. 469–480.

[15] A. Fourmigue, G. Beltrame, and G. Nicolescu, "Efficient transient thermal simulation of 3d ics with liquid-cooling and through silicon vias," in *Proc. Design, Autom. Test Eur. Conf. Exhibit. (DATE)*, Mar. 2014, pp. 1–6.

[16] G. Huang, M. S. Bakir, A. Naeemi, and J. D. Meindl, "Power delivery for 3-D chip stacks: Physical modeling and design implication," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 2, no. 5, pp. 852–859, May 2012.

[17] S. R. Sarangi, G. Ananthanarayanan, and M. Balakrishnan, "LightSim: A leakage aware ultrafast temperature simulator," in *Proc. 19th Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2014, pp. 855–860.

[18] Y. Zhang, Y. Zhang, and M. S. Bakir, "Thermal design and constraints for heterogeneous integrated chip stacks and isolation technology using air gap and thermal bridge," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 4, no. 12, pp. 1914–1924, Dec. 2014.

[19] L. Zheng, Y. Zhang, and M. S. Bakir, "Full-chip power supply noise time-domain numerical modeling and analysis for single and stacked ICs," *IEEE Trans. Electron Devices*, vol. 63, no. 3, pp. 1225–1231, Mar. 2016.

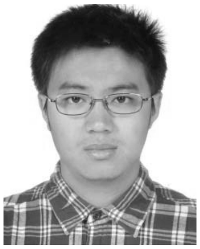
[20] H. Zhuang, S.-H. Weng, J.-H. Lin, and C.-K. Cheng, "MATEX: A distributed framework for transient simulation of power distribution networks," in *Proc. ACM Design Autom. Conf.*, Jun. 2014, pp. 1–6.

[21] X. Zhang, T. Tong, S. Kanev, S. K. Lee, G.-Y. Wei, and D. Brooks, "Characterizing and evaluating voltage noise in multi-core near-threshold processors," in *Proc. Int. Symp. Low Power Electron. Design*, Sep. 2013, pp. 82–87.

[22] ASU. *Predictive Technology Model*, accessed on Feb. 3, 2017. [Online]. Available: http://ptm.asu.edu/

[23] Intel. *Intel CoreTM i7 Processor Families for the LGA2011-0 Socket, Thermal Mechanical Specification and Design Guide*, accessed on Feb. 3, 2017. [Online]. Available: http://www.intel.com/content

[24] J. Xie and M. Swaminathan, "Electrical–thermal cosimulation with nonconformal domain decomposition method for multiscale 3-D integrated systems," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 4, no. 4, pp. 588–601, Apr. 2014.



Yang Zhang (S'13) received the B.S. degree in microelectronics and math (double major) from Peking University, Beijing, China, in 2012. He is currently pursuing the Ph.D. degree in electrical engineering with the Georgia Institute of Technology, Atlanta, GA, USA.

Muhannad S. Bakir (SM'12) is currently a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. His current research interests include 3-D electronic system integration, advanced cooling and power delivery for 3-D systems, biosensors and their integration with CMOS circuitry, and nanofabrication technology.

Dr. Bakir was a recipient of the 2013 Intel Early Career Faculty Honor Award, the 2012 DARPA Young Faculty Award, and the 2011 IEEE CPMT Society Outstanding Young Engineer Award. In 2015, he was elected by the IEEE CPMT Society to serve as a Distinguished Lecturer for a four-year term. He and his research group received more than 20 conference and student paper awards including six from the IEEE Electronic Components and Technology Technology Conference, four from the IEEE International Interconnect Technology Conference, and one from the IEEE Custom Integrated Circuits Conference. His group was awarded the 2014 Best Paper of the IEEE TRANSACTIONS ON COMPONENTS, PACKAGING, AND MANUFACTURING TECHNOLOGY in the area of advanced packaging. He is an Editor of the IEEE TRANSACTIONS ON ELECTRON DEVICES and an Associate Editor of the IEEE TRANSACTIONS ON COMPONENTS, PACKAGING, AND MANUFACTURING TECHNOLOGY.