Banafsheh Barabadi
Department of Mechanical Engineering,
Massachusetts Institute of Technology,
77 Massachusetts Avenue,
Cambridge, MA 02139

Satish Kumar
George W. Woodruff School
of Mechanical Engineering,
Georgia Institute of Technology,
801 Ferst Drive,
Atlanta, GA 30332

Valeriy Sukharev
Design-to-Silicon,
Mentor Graphics Corporation,
46871 Bayside Parkway,
Fremont, CA 94538

Yogendra K. Joshi[1]
George W. Woodruff School
of Mechanical Engineering,
Georgia Institute of Technology,
801 Ferst Drive,
Atlanta, GA 30332
e-mail: yogendra.joshi@me.gatech.edu

Multiscale Transient Thermal Analysis of Microelectronics

In a microelectronic device, thermal transport needs to be simulated on scales ranging from tens of nanometers to hundreds of millimeters. High accuracy multiscale models are required to develop engineering tools for predicting temperature distributions with sufficient accuracy in such devices. A computationally efficient and accurate multiscale reduced order transient thermal modeling methodology was developed using a combination of two different approaches: "progressive zoom-in" method and "proper orthogonal decomposition (POD)" technique. The capability of this approach in handling several decades of length scales from "package" to "chip components" at a considerably lower computational cost, while maintaining satisfactory accuracy was demonstrated. A flip chip ball grid array (FCBGA) package was considered for demonstration. The transient temperature and heat fluxes calculated on the top and bottom walls of the embedded chip at the package level simulations are employed as dynamic boundary conditions for the chip level simulation. The chip is divided into ten function blocks. Randomly generated dynamic power sources are applied in each of these blocks. The temperature rise in the different layers of the chip calculated from the multiscale model is compared with a finite element (FE) model. The close agreement between two models confirms that the multiscale approach can predict temperature rise accurately for scenarios corresponding to different power sources in functional blocks, without performing detailed FE simulations, which significantly reduces computational effort. [DOI: 10.1115/1.4029835]

Keywords: Joule heating, three-dimensional (3D) architecture, transient thermal analysis, reduced order modeling, proper orthogonal decomposition, progressive zoom-in approach

1 Introduction

The integrated circuits (IC) industry is driven by scaling to smaller and higher performing devices to enable lower cost and higher speed. However, major challenges exist in maintaining performance and reliability while facing fundamental scaling limitations. The current chip and package architectures are subjected to higher density of the heat dissipating elements and elevated total power generation rates. This can result in local hot-spots that are layout and/or workload dependent, leading to significant variation in the performance and leakage current of devices. Moreover, cyclic thermal events as a result of Joule heating in the metallic interconnect and transistors can lead to fatigue failure, due to the thermal expansion coefficients mismatch among different materials in the device. Thus, it is essential to develop fast and accurate multiscale models to calculate the thermal response of circuits for advanced technology nodes.

Various approaches have been proposed in the literature for predicting temperature distributions with sufficient accuracy in chips and packages [1,2]. Among these multiscale methodologies, the traditional bottom-up approaches are extensively used for transient thermal modeling. Perhaps, the best known of this class of methods is the resistance–capacitance network, which is constructed using thermal impedances [3,4]. The accuracy of the models decreases for complex geometries, complex boundary conditions, and nonlinearity in the heat conduction equation [5].

Another common bottom-up approach utilizes compact models, which can be finite volume (FV) or FE based. In a traditional FE or FV analysis, the domain is discretized in a way that each element is homogeneous. It can, however, have anisotropic thermal conductivity. Compact models do not require conventional bilinear rectangular or homogeneous elements and can have elements comprising both metal and dielectric region. Some of the first compact modeling work was done by Kreuger and Bar-Cohen in 1992 [6]. They modeled a chip package with a simplified resistor network and shorter simulation times. However, the network topography of compact models becomes complex with increase in model size, also potentially compromising the accuracy of the model [7]. Another limitation of such compact models is the difficulty in handling fluid/solid interactions. In general, these bottom-up approaches have primarily addressed the steady-state Joule heating in interconnects. However, pulsed currents and the resulting transient heat conduction in interconnect arrays remain a key concern in the design for reliability for the next generation high-performance chips.

Top-down approaches are another category of multiscale thermal modeling in microelectronics. A recent approach is behavioral thermal modeling, which is a combination of the generalized pencil-of-function (GPOF) [8,9] and subspace methods [10,11]. GPOF was developed in the communications community to estimate poles of an electromagnetic system by solving a generalized eigenvalue problem. These methods are mainly used for high-performance multicore microprocessor design. In general, they potentially suffer from a lack of predictability problems. Therefore, there is a need for the development of a new thermal simulation methodology that overcomes the challenges faced by existing thermal models.

In this study, a novel, computationally efficient, and accurate multiscale reduced order transient thermal modeling methodology is developed, which comprises two parts: (1) *progressive zoom-in* and (2) *POD*. The analyses at various length scales are integrated via the progressive zoom-in approach, which is illustrated in Fig. 1 and will be further discussed in Sec. 2.2. POD is a robust and elegant method of data analysis that provides low-dimensional but accurate descriptions of a high-dimensional system. It was first introduced by Lumley [12] in the field of turbulence; Holmes et al. [13] provided a thorough summary for applications of POD in various fields. As shown by Barabadi et al. [14], for any linear system, the method is capable of predicting

[1]Corresponding author.
Contributed by the Electronic and Photonic Packaging Division of ASME for publication in the JOURNAL OF ELECTRONIC PACKAGING. Manuscript received May 2, 2013; final manuscript received February 16, 2015; published online April 16, 2015. Assoc. Editor: Amy Fleischer.

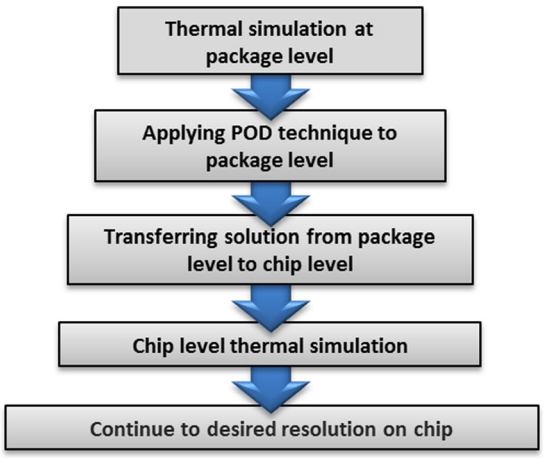

Fig. 1 Flowchart of the hybrid scheme for multiscale thermal modeling

transient temperature distribution regardless of the temporal or spatial dependence of the applied heat source. This feature provides the ability to predict temperature distributions for arbitrary heat inputs, by using a smaller sample set of applied heat sources and power maps, resulting in considerably decreased simulation time. Combining POD with the progressive zoom-in approach can further enhance the computational efficiency.

The proposed methodology has the capability of modeling several decades of length scale from package to "chip component" and potentially the "interconnect" (not included here) levels, at a significantly lower computational cost than currently available methods. This characteristic of the method also applies for time scales from seconds down to microseconds, corresponding to various transient thermal events. The suggested approach provides the ability to rapidly predict thermal responses under different power input patterns, based only on a few representative detailed simulations, while maintaining adequate spatial and temporal accuracy.

In this paper, an FCBGA package with an embedded die is considered for thermal modeling. Random dynamic power distributions were considered for the total chip power, as well as for the function blocks that compose the entire chip to demonstrate the capability of the POD method. To validate this methodology, the results were compared with an FE model developed in COMSOL [15]. It is demonstrated that the computational time is reduced by at least two orders of magnitude at every step of modeling.

2 Hybrid Scheme for Multiscale Thermal Modeling

A hybrid scheme has been developed in this paper, which combines the implementation of POD and progressive zoom-in approach, as summarized below.

2.1 Fundamentals of POD Method. POD offers an optimal set of basis functions, also known as POD modes, which are empirically determined from an ensemble of observations. These observations are obtained either experimentally or from numerical simulation, as in this study. The POD method characterizes and captures the overall behavior and complexity of a physical system by using a reduced number of degrees-of-freedom. This results in a much lower computational cost than a full-field simulation method. The most remarkable characteristic of the POD is its *optimality*, i.e., it provides the most efficient way of capturing the dominant components of an infinite-dimensional process, with only finite number of basis functions [13]. In developing the POD model, data sets are expanded for modal decomposition on empirically determined basis functions in a way that minimizes the least square error between the true solution and the truncated representation of the POD model. Therefore, it makes the POD method the most efficient method of capturing the dominant components of a large-dimensional system with a finite number of modes [16,17].

In this technique, the temperature distribution is determined from the expansion

$$T(x,y,z,t) = T_0(x,y,z) + \sum_{i=1}^{m} b_i(t)\phi_i(x,y,z) \tag{1}$$

where T_0 is the time average of temperature (i.e., the mean vector of the observation matrix), $\varphi_i(x, y, z)$ is the ith POD mode, and $b_i(t)$ is the ith POD coefficient [14]. A detailed procedure to generate a two-dimensional (2D) POD based reduced order model is provided in Ref. [14]. The primary steps to generate a POD based reduced order model are outlined below:

(1) Generating the observation matrix.
(2) Calculating basis functions (POD modes).
(3) Calculating POD coefficients, b_i.

As demonstrated in Ref. [14], the POD coefficients, b_i, can be determined by solving the discretized matrix of coupled ordinary differential equations, Eq. (2), using the sixth-order Runge–Kutta method shown below:

$$A_{ij}\dot{b}_j(t) - B_{ij}b_j(t) - (c+q)_i = 0, i, j = 1, 2, ..., m \tag{2}$$

Coefficients A_{ij}, B_{ij}, c_i, and q_i in Eq. (2) were derived and presented for 2D POD model in Ref. [14]. For this study, coefficients in Eq. (2) were determined for 3D analysis as

$$A_{ij} = \int_\Omega \varphi_j \cdot \varphi_i d\Omega \tag{3a}$$

$$B_{ij} = \int_\Omega \alpha\varphi_j \cdot \nabla^2\varphi_i d\Omega = -\int_\Omega \alpha\left(\frac{\partial\varphi_j}{\partial x}\cdot\frac{\partial\varphi_i}{\partial x} + \frac{\partial\varphi_j}{\partial y}\cdot\frac{\partial\varphi_i}{\partial y} + \frac{\partial\varphi_j}{\partial z}\cdot\frac{\partial\varphi_i}{\partial z}\right)d\Omega + \int_x \left(\alpha\varphi_j\cdot\frac{\partial\varphi_i}{\partial y}\right]_{y=y_{\min}}^{y=y_{\max}} dx + \int_y \left(\alpha\varphi_j\cdot\frac{\partial\varphi_i}{\partial x}\right]_{x=x_{\min}}^{x=x_{\max}} dy + \int_z \left(\alpha\varphi_j\cdot\frac{\partial\varphi_i}{\partial z}\right]_{z=z_{\min}}^{z=z_{\max}} dx \tag{3b}$$

$$c_j = \int_\Omega \alpha\varphi_j \cdot \nabla^2 T_0 d\Omega = -\int_\Omega \alpha\left(\frac{\partial\varphi_j}{\partial x}\cdot\frac{\partial T_0}{\partial x} + \frac{\partial\varphi_j}{\partial y}\cdot\frac{\partial T_0}{\partial y} + \frac{\partial\varphi_j}{\partial z}\cdot\frac{\partial T_0}{\partial z}\right)d\Omega + \int_x \left(\alpha\varphi_j\cdot\frac{\partial T_0}{\partial y}\right]_{y=y_{\min}}^{y=y_{\max}} dx + \int_y \left(\alpha\varphi_j\cdot\frac{\partial T_0}{\partial x}\right]_{x=x_{\min}}^{x=x_{\max}} dy + \int_z \left(\alpha\varphi_j\cdot\frac{\partial T_0}{\partial z}\right]_{z=z_{\min}}^{z=z_{\max}} dz \tag{3c}$$

$$q_j = \int_\Omega \frac{1}{\rho c_p} \varphi_j \cdot q'''(t) d\Omega \tag{3d}$$

The last two terms on the right-hand side of Eqs. (3*b*) and (3*c*) are the boundary terms. If the boundary conditions are homogeneous or insulation, these are eliminated and B_{ij} and c_i are simplified to

$$B_{ij} = -\int_\Omega \alpha \left(\frac{\partial \varphi_j}{\partial x} \cdot \frac{\partial \varphi_i}{\partial x} + \frac{\partial \varphi_j}{\partial y} \cdot \frac{\partial \varphi_i}{\partial y} + \frac{\partial \varphi_j}{\partial z} \cdot \frac{\partial \varphi_i}{\partial z} \right) d\Omega \tag{4a}$$

$$c_j = -\int_\Omega \alpha \left(\frac{\partial \varphi_j}{\partial x} \cdot \frac{\partial T_0}{\partial x} + \frac{\partial \varphi_j}{\partial y} \cdot \frac{\partial T_0}{\partial y} + \frac{\partial \varphi_j}{\partial z} \cdot \frac{\partial T_0}{\partial z} \right) d\Omega \tag{4b}$$

(4) Generating the POD temperature field.

A sufficient number of POD modes and POD coefficients need to be calculated, which can then be used in Eq. (1) for the determination of the temperature field anywhere in the domain and at any instant of time.

The number of retained POD modes is quite critical in capturing the physics of the problem. It is shown that an insufficient number of POD modes can cause significant phenomena not to be detected [18]. On the contrary, taking too many POD modes can produce unexpected behavior or make the model unstable. The last POD modes are generally associated with low energy terms in the model and have rapid localized fluctuation throughout the domain. If too many modes are considered in the POD reconstruction, the accumulation of these rapid fluctuations results in an increase in the numerical error and can potentially cause the solution to diverge [19,20]. The energy captured by the *i*th basis function in the problem is relative to its corresponding eigenvalue, λ_i. Sorting these eigenvalues in a descending order results in an ordering of the corresponding POD modes [14]. Therefore, the first POD mode captures the largest portion of energy relative to the other basis functions. To determine the truncation degree of the POD method, the *cumulative correlation energy*, E_m, captured by the first m POD modes is defined by Bizon et al. [21]

$$E_m = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{5}$$

To be able to generate a reliable POD model, in the present study, the number of POD modes is determined in such a way that the cumulative energy of the modes, calculated from Eq. (5), is larger than 99.9%.

2.2 Progressive Zoom-In Approach. The progressive zoom-in method integrates package and chip level analyses, acquiring the advantages of each. Figure 1 shows a flowchart of the approach used in this study for multiscale transient thermal modeling of a representative FCBGA package. The overall hybrid approach is outlined below:

(1) Thermal simulation at the package level: The first step is to model the entire structure, i.e., the package, including the surrounding mold, underfill, solder bumps, and substrate. This simulation is performed in the commercial code COMSOL. It is important to note that at this level, the chip is modeled as a solid block with effective material and thermal properties, without considering internal details.
(2) Applying POD technique to package level: Once the temperature distribution at package level is determined, a POD model is developed. The POD model provides the ability to predict dynamic temperature distribution for different power maps and types of power sources, without developing any further full-field FE models, which can significantly decrease computational cost and potentially be used to define a criterion for the optimal distribution of the current density in the domain.
(3) Transferring the solution from package level to the chip level: Once the temperature distribution at the package level is obtained, a combination of temperature and heat flux at the top and bottom walls of the chip is extracted and linearly interpolated on a 2D grid with higher spatial resolution. These data were then applied as boundary conditions for the chip level simulation.
(4) Chip level thermal simulation: At this level, the chip is no longer treated as a solid block. It is divided into subdomains called function blocks. Each block represents a specific component with unique functionality on the chip and consists of three sublayers: (1) top Si layer, (2) middle device layer, and (3) interconnect/dielectric multilayer (see Fig. 2). Function blocks were simulated based on the assigned power generation and calculated effective material/thermal properties for each layer within that block. At this level of thermal simulation, the spatial resolution is limited to the sublayers. Once the chip is divided into subdomains, the power map needs to be determined at any instant of time for each individual function block.
(5) Continue to the desired resolution on the chip: This method can be continued to multiple levels, such that the desired spatial resolution is achieved. Only representative results for two steps (package and chip level) are presented in this paper.

3 Results and Discussion

Figure 2(*a*) shows the schematic of the simplified FCBGA package used in this study for the package level modeling. This model is for low power portable systems, where heat sinks and forced cooling are not employed due to the compact form factor. As described in Sec. 2, the first step is to model the package for which the material properties and dimensions are required. Table 1 lists these for the die, solder bumps, underfill, mold, and substrate. These values were mainly provided by Mentor Graphics Corporation, and the rest were chosen based on Ref. [22]. Reference [23] is used as a guideline for the dimensions of the FCBGA package. Underfill is a specially engineered epoxy that fills the area between the die and the carrier surrounding the solder bumps. Effective density and specific heat of the underfill layer are calculated based on volume averaging. It is assumed that 60% of the surface area between the die and substrate is covered with underfill and 40% is solder bumps. The effective vertical ($K_{v_{eff}}$) and horizontal ($K_{h_{eff}}$) thermal conductivity values are calculated based on thermal resistor network formulation

$$K_{h_{eff}} = \frac{1}{\left(\frac{\forall_U}{\forall_{tot}} \frac{1}{K_U} + \frac{\forall_S}{\forall_{tot}} \frac{1}{K_S} \right)} \tag{6a}$$

$$K_{v_{eff}} = \left(K_U \frac{A_{hU}}{A_h} + K_S \frac{A_{hS}}{A_h} \right) \tag{6b}$$

where $\forall_{tot}$ is the entire volume, $\forall_U$ and $\forall_S$ are volumes of underfill and solder bumps, respectively. Similarly, A_{hU} and A_{hS} are the cumulative horizontal cross-sectional areas of the underfill and solder bumps. K_U and K_S are the thermal conductivities of the underfill and solder bumps, respectively. Considering that the solder bumps are made of conductive material and electrically connect the chip to the underlying substrate while underfill is an insulating material, it is expected that the vertical effective thermal conductivity of the underfill layer will be significantly higher than its horizontal value. The computed values are 20.2 and

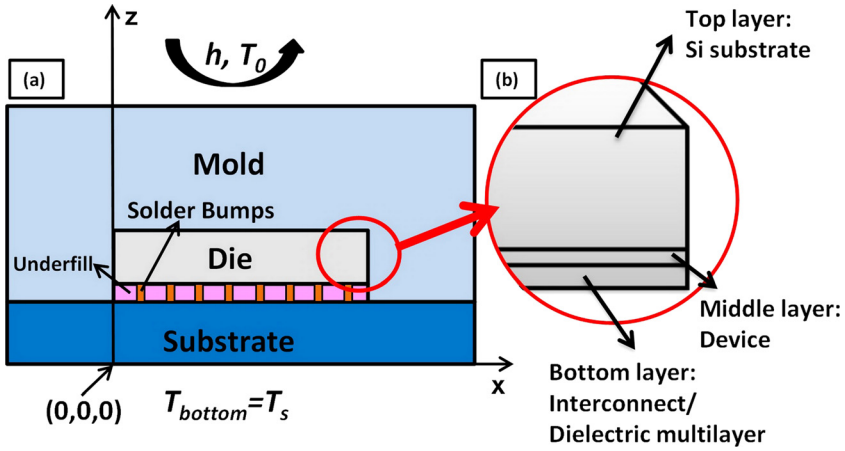

Fig. 2 (*a*) Schematic of a simplified FCBGA for package level modeling and (*b*) zoomed-in schematic of the die layer used in chip level modeling

Table 1 Material properties and dimensions of the package

	Thermal conductivity (W/m·K)	Density (kg/m^3)	Specific heat capacity (J/kg·K)	Dimension (mm^3)
Die	98.4	2300	721	$10 \times 10 \times 0.266$
Underfill	0.6	1820	236	$10 \times 10 \times 0.1$
Solder bumps	50	8510	183	40% of the die surface area
Effective underfill	1.47 (horizontal) 20.24 (vertical)	4496	214.8	60% of the die surface area
Mold	0.5	1820	236	$37 \times 37 \times 1.9$
Substrate	0.7	1700	920	$37 \times 37 \times 1.1$

1.47 W/m·K, respectively. The package dimensions, also listed in Table 1, were provided by Mentor Graphics Corporation.

Natural convection boundary condition is imposed on the top surface and vertical boundaries of the package with a heat transfer coefficient of $h = 15\ \mathrm{W/m^2\ K}$ in the typical range for air cooling [24]. A constant temperature boundary condition is applied to the bottom surface. The initial temperature and the surrounding temperature were assumed to be equal to the room temperature $T_{\mathrm{amb}} = 300\ \mathrm{K}$.

A detailed FE model is developed in COMSOL using a time step of $dt = 0.05$ s. The convergence of the FE model is verified with respect to the solver type, time step, and time integration method. The FE model of the package consists of 75,919 elements, of which 343 are for the chip (die). This grid size is determined after performing mesh independence analysis. For the grid independence study, the mesh resolution of the model is continuously refined until there is less than 1% difference in the computed temperatures. This analysis indicated that the grid size of 75,919 elements is sufficient. Total chip power is $Q = 3 \sin 2\pi t + 3$ (W), which is applied for 1 s. The temperature rise in the simulation domain is represented by ΔT (K) throughout the paper. Figure 3(*a*) shows the spatial distribution of the temperature rise in the FCBGA package extracted from the FE model after 1 s. The temperature rise of the chip is plotted separately in Fig. 3(*b*). Table 2 demonstrates the numerical solution parameters and specifications used in package level FE model, POD technique, and chip level FE model.

After obtaining the transient temperature field at the package level, the POD model is developed using the algorithm demonstrated in Sec. 2.1. Twenty-six observations of the transient temperature solution were taken in the first 0.5 s using the package level FE model. These observations correspond to the temperature solutions obtained at different time instants using total chip power of $Q = 3 \sin 2\pi t + 3$ (W). It is important to note that the observations are generated only for this case, and results for any different power dissipation are calculated without any new observations. In fact, the POD solutions of these transient thermal scenarios are independent of the initial observations. Essentially, for any linear system, once the solution to a sample case of chip total power is obtained, there is no need to generate new observations or perform full-field FE simulations. The ability of the POD method to predict other cases based on a smaller sample set can significantly

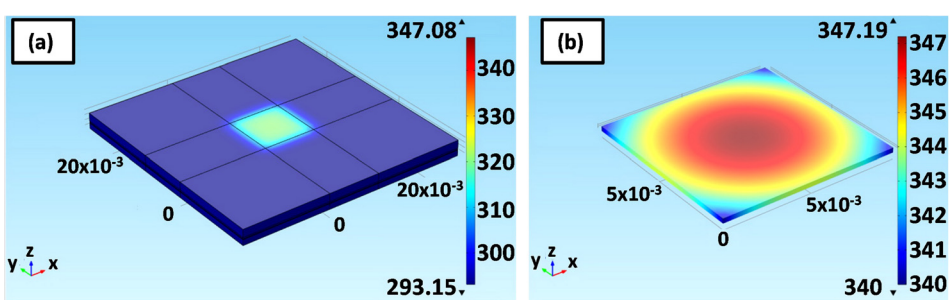

Fig. 3 Spatial distribution of temperature rise extracted from FE method after 1 s for (*a*) FCBGA package and (*b*) chip

Table 2 Parameters of numerical solution

	FE package level model	POD package level model	FE chip level model
Solver type	Crank–Nicolson time integration scheme and conjugate gradient iterative solver	Sixth-order Runge–Kutta method	Crank–Nicolson time integration scheme and conjugate gradient iterative solver
Time step (s)	0.05	0.05	0.05
Number of element	75,919	75,919	268,033
Mesh independent	Yes	Yes	Yes
Simulation time	23.7 mins	40 s (first run) 15 s (other runs)	26.27 mins

decrease computational cost. After the observations have been generated, the POD basis functions (POD modes) are calculated. In order to build a reliable but fast reduced order model, only four POD modes are used in the present model. This is chosen such that the *cumulative correlation energy*, E_m, Eq. (5), is greater than 99.9%. The first two modes alone capture over 96% of the energy. The results will not have the desired accuracy if the number of initial observations, n, is less than the minimum required POD modes (four in the present case).

Since the POD modes are three-dimensional, for better visualization, 2D contours of the first four POD modes at height $z = 1.33$ mm across the center of the die are illustrated in Fig. 4. This height is chosen because it has the highest temperature gradient, due to material inhomogeneity and the application of power source only to the die. The POD modes are normalized with the total sum of the modes for a more accurate comparison.

To have a realistic and accurate thermal simulation, a detailed dynamic power map of the embedded chip is required. However, one of the major challenges in microelectronics is the determination of the dynamic power dissipation in the chip, since power values and temperature distribution are coupled in an electrothermal loop. A randomly generated function is assumed for the dynamic chip power in this study to illustrate the application of the POD formulation. Figure 5(*d*) shows the randomly generated power distribution for the chip for the first 1 s. The minimum and maximum allowed values for the power were chosen to be 3 W and 18 W, respectively. In essence, there are three changes in the nature of the previously used power source ($Q = 3 \sin 2\pi t + 3$ (W)) and the current random chip power:

(1) The first case is only applied for 0.5 s, whereas the second case used for POD approach models the entire 1 s.
(2) The magnitude of the maximum value for the second case is 18 W versus 6 W for the initial FE simulation.
(3) The temporal behavior of the power has changed from a well-defined sinusoidal function to a randomly generated step function.

The benefit of using the POD model to predict the transient thermal profile for a different power source than the original one is that no new observation or full-field simulation is required. The POD coefficients were calculated as functions of time using the method of Galerkin projection [14].

Once the POD modes and the b-coefficients are calculated, the transient temperature field can be determined using Eq. (1). Figure 6(*a*) displays the 3D spatial distribution of temperature extracted from the POD model at 1 s. For higher precision, the domain is sliced vertically along the XZ plane and four of these slices are presented. The right-most slice is the A–A cross section across the center of the die (Fig. 6(*b*)). To validate the results of the POD model, a full-field FE model with a time step of 0.05 s is developed in COMSOL using the same grid points and elements used in the POD model. The results are shown in Fig. 6(*c*). It can be inferred that the POD model closely predicts the transient thermal behavior of the system, not only for the given time domain but also for projected future time (>0.5 s) using just a few POD modes. The mean absolute error between the POD and FE model is 7.2% over the entire space and time domain. Required computation time for the full-field FE simulation is 23.7 mins versus 40 s for the POD simulations. The first POD simulation run-time is 40 s, while additional simulations with different power sources take 15 s each. The computations are performed on a workstation using an Intel(R) Core (TM) i7 @ 2.20 GHz with 8 GB RAM.

For a more comprehensive comparison between POD and FE results, the time-dependent temperature rise at four different points in the FCBGA package (center of the mold, die, underfill, and substrate) is considered (Fig. 7(*a*)). The maximum error occurs at the center of the die between $t = 0.833$ and $t = 1$ s. As illustrated in Fig. 7(*b*), this is the time period when the maximum

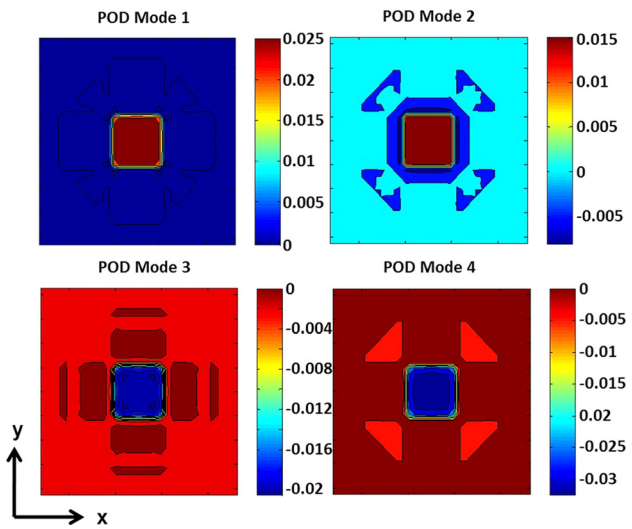

Fig. 4 2D contour plots of the first four POD modes at $z = 1.33$ mm from the bottom of the package; this plane crosses the center of die

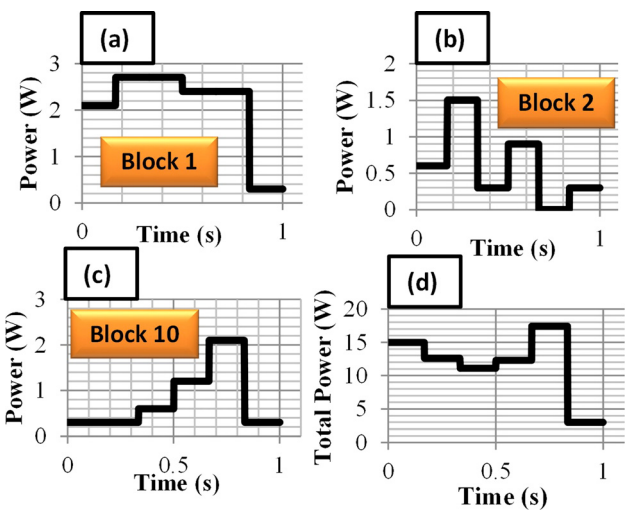

Fig. 5 Dynamic randomly generated power profile for function blocks 1, 2, 10 (*a–c*) and randomly generated total chip power (*d*)

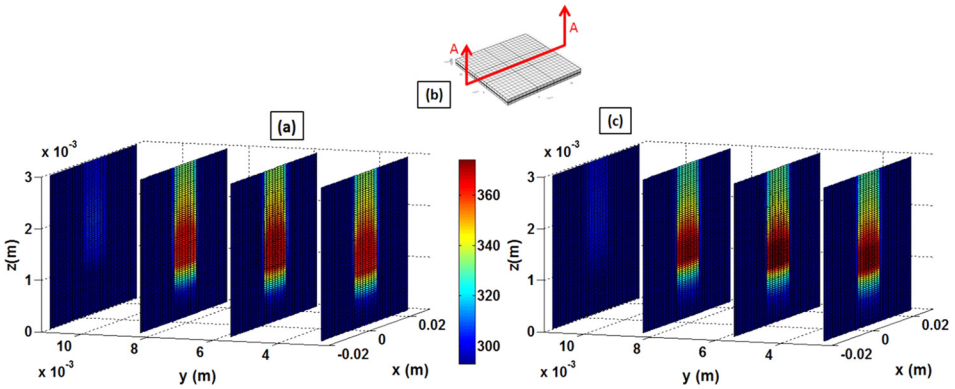

Fig. 6 Spatial distribution of temperature rise at 1 s extracted from the POD model (*a*) and FE simulation (*c*). The domain is sliced vertically along *XZ* plane. The right-most slice is the A–A cross section (*b*).

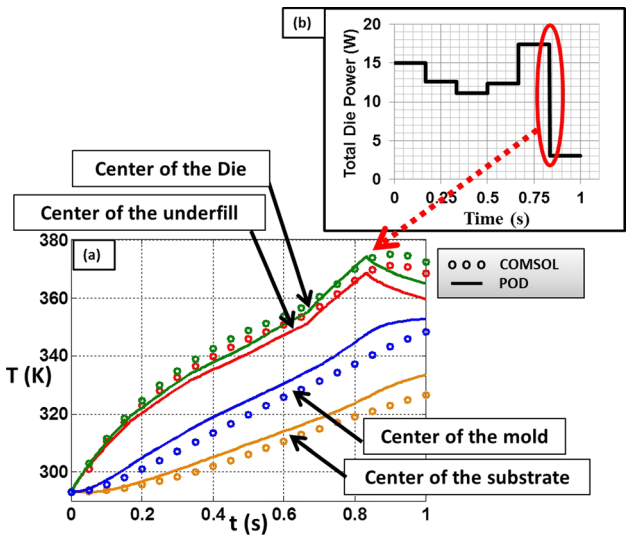

Fig. 7 Comparison of temporal dependence of temperature rise between FE (markers) and POD (solid lines) models at four different points (*a*) and the corresponding randomly generated total chip power (*b*)

jump in the total chip power occurs. The dotted arrow in Fig. 7 points to the time of this maximum jump in the temperature plot.

After obtaining the transient thermal solution at the package level and with the POD model, the next step in the hybrid scheme is to transfer the solution to the chip with the higher spatial resolution in the form of boundary conditions. Due to the transient nature of this analysis, temperature on the top surface and heat flux on the bottom surface of the die are extracted at ten different time intervals between 0 and 1 s (every 0.1 s). The extracted data are then applied as temporal boundary conditions for the chip level model. The four side walls of the die are assumed to be adiabatic considering the high aspect ratio of the die. The solution is linearly interpolated on a 2D grid with much higher spatial resolution at this level (268,033 elements to model the chip at this level versus 343 elements to model the chip at the package level).

At the chip level simulation, the die is no longer treated as a solid block. It is segmented into ten subdomains called function blocks. In practical applications, each block represents a specific component with unique functionality on the chip. In this study, the blocks were artificially created for illustration of the proposed methodology [25]. As demonstrated in Fig. 2(*b*), each block has three layers: (1) top Si layer with the thickness of 0.249 mm, (2) middle layer which is a 5 μm-thick device layer, and (3) interconnect/dielectric multilayer at the bottom with the thickness of 16.72 μm. The third layer consists of 21 sublayers including ten metal layers.

Due to the high level of geometrical complexity, a combination of directional volume and surface averaging methods was used to determine the effective properties of the functional blocks. Table 3 indicates the calculated material properties of the blocks at the chip level simulations. Density and specific heat are calculated using the volume averaging method. In-plane thermal conductivity is determined based on the ratio of the volume of the interconnects to the total volume, due to the fact that the in-plane thermal transport is governed mainly by the interconnects. On the other hand, vias are the dominant paths of through-plane heat transfer in each block. Therefore, for the vertical thermal conductivity, the values are calculated based on the ratio of the volume of the vias

Table 3 Properties and dimensions of the function blocks for chip level simulation

		Vertical thermal conductivity (W/m·K)	Horizontal thermal conductivity (W/m·K)	Density (kg/mm^3)	Specific heat capacity (J/kg·K)
Interconnect/ dielectric layer	Block 1	0.48	3.53	1512.17	742.01
	Block 2	0.48	3.53	1512.16	742.01
	Block 3	0.49	3.56	1512.73	741.99
	Block 4	0.48	3.49	1511.44	742.05
	Block 5	0.49	3.66	1514.61	741.90
	Block 6	0.47	3.41	1509.90	742.12
	Block 7	0.49	3.58	1513.21	741.96
	Block 8	0.49	3.63	1514.20	741.91
	Block 9	0.48	3.49	1511.46	742.05
	Block 10	0.49	3.56	1512.82	741.98
Device layer		34	34	2320	678
Si layer		130	130	2329	700

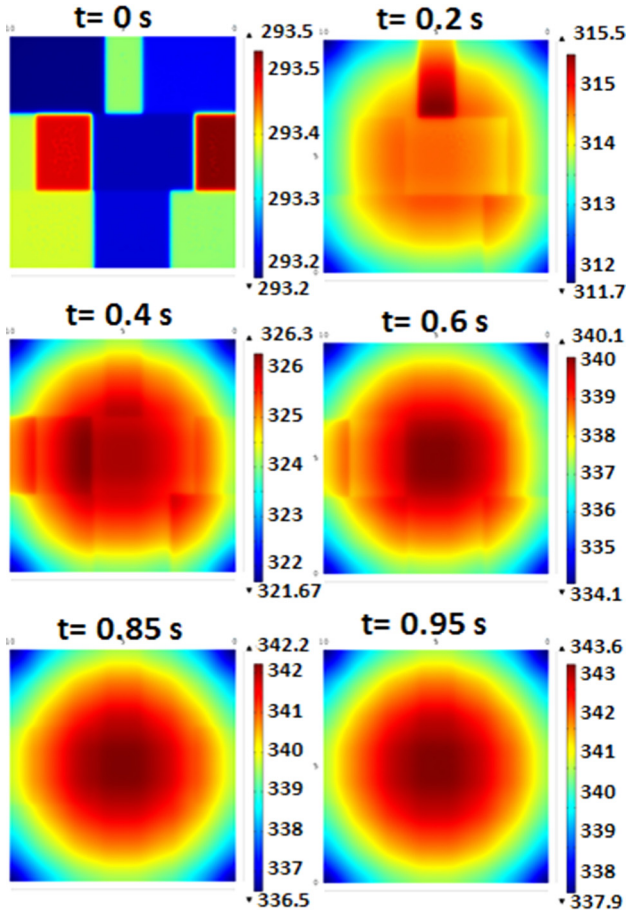

Fig. 8 Transient temperature distribution at the interface of device and interconnect/dielectric layers: $t = 0$, 0.2, 0.4, 0.6, 0.85, and 0.95 s

to the entire volume of each block. At this stage, the spatial resolution is limited to the sublayers of the blocks.

Once the chip is divided into subdomains, the dynamic power grid needs to be assigned to individual function blocks. For this study, the Joule heating produced in the third layer is neglected and the only powered layer is the device layer. Using the same method as described earlier, ten random power sources with minimum and maximum values of 0 and 3 W were generated between 0 and 1 s. The power sources for blocks 1, 2, and 10 are presented in Figs. 5(*a*)–5(*c*) as representatives. The block power sources are generated in such way that their sum will equal the total chip power used for the package level simulation as shown in Fig. 5(*d*).

After allocating the power sources to the function blocks, an FE model is developed using the time step of $dt = 0.05$ s for the final step of the hybrid scheme. As mentioned, the model consists of 268,033 elements. The computational time to run the transient simulation for 1 s is 26.27 mins. Figure 8 displays the 2D spatial distribution of temperature rise extracted from the FE solution at various times between 0 and 1 s at height $z = 16.72\ \mu$m, which is the plane between the device layer and interconnect/dielectric multilayer (plane between layers 2 and 3). Based on the one-dimensional simplified resistance-network model of the chip, it can be seen that the majority of heat generated at the device layer will be dissipated through the top silicon layer and only about 5% of the heat is dissipated through the underlying interconnect/dielectric multilayer; i.e., $(R_{\text{through Si layer}}/R_{\text{through interconnet/dielectric layer}}) \sim 0.06$.

4 Summary and Conclusion

In this study, a computationally efficient and accurate multiscale reduced order transient thermal model is developed which has the capability of modeling several decades of length and time scales at a considerably lower computational cost, while maintaining satisfactory accuracy. In particular, by using the proposed model, the computational time is reduced by at least two orders of magnitude at every step of zooming into the geometry. It is also shown that the hybrid scheme accurately predicts the transient thermal behavior of the system for not only the time domain considered for the initial observations, but also for time outside of the specified initial time domain. The mean absolute error between the proposed and FE model is 7.2% over the entire space and time domain.

A distinct benefit of the proposed method is that, for any linear system, the POD solution is independent of the transient power profile. In other words, once the solution to a sample power input is obtained, there is no need to generate new observations or full-field FE simulations. This important feature can drastically decrease computational cost for parametric numerical simulations, making POD a fast and robust method for reduced order model of transient heat conduction in microelectronic devices. An additional unique characteristic of this model is that the initial observations can be obtained experimentally, which creates the ability of modeling a potentially complex system without generating any numerical model.

The hybrid scheme proposed in this study is not limited to the two levels considered in the present study and can potentially extended from package to "interconnect level." One of the strengths of this method is that the algorithm can be scaled to multiple levels and can be used to simulate more detailed structures on the chip, while taking advantage of the capabilities of POD method to avoid any further full-field simulation. In essence, without losing the desired resolution, the hybrid scheme proposes a new approach to further decrease the computational cost by orders of magnitude.

The integration of the proposed method into the commercially available software packages can create a powerful tool for both academic and industry applications. It will address the lack of physical models for multiscale thermal problems, relating potential performance variation to critical layout parameters. Another possible application of this method would be in the IC design industry. A POD model can be developed for a specific chip structure using output signals of the embedded temperature sensors on the chip as the original observations. By incorporating this model into a closed-loop on-chip control system, the possible locations of hot-spots can then be predicted and potentially avoided.

Acknowledgment

The research is partially funded by the Semiconductor Research Corporation (SRC) under Task 1883.001 and Design-to-Silicon division of Mentor Graphics Corporation.

Nomenclature

A = cross-sectional area (m^2)
A_{ij} = coefficient in Eq. (2)
b_i = ith POD coefficient (K)
B_{ij} = coefficient in Eq. (2)
c_i = coefficient in Eq. (2)
dt = time step (s)
E_m = cumulative correlation energy
h = heat transfer coefficient (W/m^2 K)
K = thermal conductivity (W/m·K)
m = number of POD modes used
n = number of observations
Q = total chip power (W)
q_i = coefficient in Eq. (2)
t = time (s)
T = temperature (K)
T_0 = time averaged temperature (K)
z = height (mm)

Greek Symbols

λ_i = ith eigenvalue
φ_i = ith POD mode

Subscripts

amb = room/ ambient
hS = solder bumps (horizontal)
hU = underfill (horizontal)
h_eff = effective horizontal value
S = solder bumps
tot = total
U = underfill
v_eff = effective vertical value

References

[1] Gurrum, S. P., Joshi, Y. K., King, W. P., Ramakrishna, K., and Gall, M., 2008, "A Compact Approach to On-Chip Interconnect Heat Conduction Modeling Using the Finite Element Method," ASME J. Electron. Packag., **130**(3), p. 031001.

[2] Joshi, Y., 2012, "Reduced Order Thermal Models of Multiscale Microsystems," ASME J. Heat Transfer, **134**(3), p. 031008.

[3] Christiaens, F., Vandevelde, B., Beyne, E., Mertens, R., and Berghmans, J., 1998, "A Generic Methodology for Deriving Compact Dynamic Thermal Models, Applied to the PSGA Package," IEEE Trans. Compon., Packag., Manuf. Technol., Part A, **21**(4), pp. 565–576.

[4] Lasance, C., Vinke, H., Rosten, H., and Weiner, K. L., 1995, "A Novel Approach for the Thermal Characterization of Electronic Parts," Eleventh Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM XI), San Jose, CA, Feb. 7–9.

[5] Gerstenmaier, Y., and Wachutka, G., 2002, "Rigorous Model and Network for Transient Thermal Problems," Microelectron. J., **33**(9), pp. 719–725.

[6] Krueger, W., and Bar-Cohen, A., 1992, "Thermal Characterization of a PLCC-Expanded R_{jc} Methodology," IEEE Trans. Compon., Hybrids, Manuf. Technol., **15**(5), pp. 691–698.

[7] Celo, D., Xiao Ming, G., Gunupudi, P. K., Khazaka, R., Walkey, D. J., Smy, T., and Nakhla, M. S., 2005, "Hierarchical Thermal Analysis of Large IC Modules," IEEE Trans. Compon. Packag. Technol., **28**(2), pp. 207–217.

[8] Hua, Y., and Yu, Z., 1989, "Generalized Pencil-of-Function Method for Extracting Poles of an EM System From Its Transient Response," IEEE Trans. Antennas Propag., **37**(2), pp. 229–234.

[9] Hua, Y., and Yu, Z., 1991, "On SVD for Estimating Generalized Eigenvalues of Singular Matrix Pencil in Noise," IEEE Trans. Signal Process., **39**(4), pp. 892–900.

[10] Zao, L., Tan, S. X. D., Hai, W., Quintanilla, R., and Gupta, A., 2011, "Compact Thermal Modeling for Package Design With Practical Power Maps," International Green Computing Conference and Workshops (IGCC), Orlando, FL, July 25–28.

[11] Duo, L., Tan, S. X. D., Pacheco, E. H., and Tirumala, M., 2009, "Architecture-Level Thermal Characterization for Multicore Microprocessors," IEEE Trans. VLSI Syst., **17**(10), pp. 1495–1507.

[12] Lumley, J. L., 1967, "The Structure of Inhomogeneous Turbulent Flows," *Atmospheric Turbulence and Radio Wave Propagation*, Nauka, Moscow, pp. 166–178.

[13] Holmes, P., Lumley, J. L., and Berkooz, G., 1998, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, Cambridge, UK.

[14] Barabadi, B., Joshi, Y., and Kumar, S., 2011, "Prediction of Transient Thermal Behavior of Planar Interconnect Architecture Using Proper Orthogonal Decomposition Method," ASME Paper No. IPACK2011-52133.

[15] COMSOL, 2011, "COMSOL Version 4.2," Comsol Multiphysics, Inc., Burlington, MA, http://www.comsol.com

[16] Chatterjee, A., 2000, "An Introduction to the Proper Orthogonal Decomposition," Curr. Sci., **78**(7), pp. 808–817.

[17] Berkooz, G., Holmes, P., and Lumley, J., 1996, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry* (Cambridge Monographs on Mechanics), Cambridge University Press, Cambridge, UK, pp. 1200–1208.

[18] Graham, M. D., and Kevrekidis, I. G., 1996, "Alternative Approaches to the Karhunen–Loeve Decomposition for Model Reduction and Data Analysis," Comput. Chem. Eng., **20**(5), pp. 495–506.

[19] Rowley, C. W., Colonius, T., and Murray, R. M., 2001, "Dynamical Models for Control of Cavity Oscillations," AIAA Paper No. 2001-2126.

[20] Samadiani, E., Joshi, Y., Hamann, H., Iyengar, M. K., Kamalsy, S., and Lacey, J., 2012, "Reduced Order Thermal Modeling of Data Centers Via Distributed Sensor Data," ASME J. Heat Transfer, **134**(4), p. 041401.

[21] Bizon, K., Continillo, G., Russo, L., and Smula, J., 2008, "On Pod Reduced Models of Tubular Reactor With Periodic Regimes," Comput. Chem. Eng., **32**(6), pp. 1305–1315.

[22] Incropera, F. P., Bergman, T. L., Lavine, A. S., and Dewitt, D. P., 2011, *Fundamentals of Heat and Mass Transfer*, Wiley, Hoboken, NJ.

[23] Chang, K. C., Li, Y., Lin, C. Y., and Lii, M. J., 2004, "Design Guidance for the Mechanical Reliability of Low-K Flip Chip BGA Package 1," 37th International Microelectronics and Packaging Society (IMAPS) Topical Workshop and Exhibition on Flip Chip Technology, Long Beach, CA, Nov. 14–18, pp. 21–24.

[24] Tang, L., and Joshi, Y. K., 2005, "A Multi-Grid Based Multi-Scale Thermal Analysis Approach for Combined Mixed Convection, Conduction, and Radiation Due to Discrete Heating," ASME J. Heat Transfer, **127**(1), pp. 18–26.

[25] Barabadi, B., Kumar, S., Sukharev, V., and Joshi, Y. K., 2012, "Multi-Scale Transient Thermal Analysis of Microelectronics," ASME Paper No. IMECE2012-89864.