Thermal Simulation of a CPU Based on Model Order Reduction

Kayla Ruttan
Dept. of Electrical & Computer Eng.
Clarkson University
Potsdam, NY 13699-5720, USA
ruttank@clarkson.edu

Lin Jiang
Dept. of Electrical & Computer Eng.
Clarkson University
Potsdam, NY 13699-5720, USA
jiangl2@clarkson.edu

Anthony Dowling
Dept. of Computer Science
Clarkson University
Potsdam, NY 13699-5815, USA
dowlinah@clarkson.edu

Yu Liu
Dept. of Electrical & Computer Eng.
Clarkson University
Potsdam, NY 13699-5720, USA
yuliu@clarkson.edu

Ming-C. Cheng
Dept. of Electrical & Computer Eng.
Clarkson University
Potsdam, NY 13699-5720, USA
mcheng@clarkson.edu

***Abstract*—A previously developed thermal simulation technique based on model order reduction is applied to the simulation of a CPU. The approach is derived from proper orthogonal decomposition (POD) that projects the physical domain onto the POD space. It has been demonstrated that the developed approach offers an accurate thermal simulation of the CPU with a reduction in numerical degrees of freedom by several orders of magnitude compared to the direct numerical simulation (DNS). In addition, the technique has the capability of providing spatial resolution as fine as the direct numerical simulation for the CPU.**

Keywords— Model Order Reduction (MOR), Proper Orthogonal Decomposition (POD), Thermal Simulation, Thermal Circuits, CPUs, GPUs

I. INTRODUCTION

With the miniaturization of devices and multifunctional chip design, CPUs and GPUs are being produced with an extremely high density of devices and complex interconnections. According to Moore's, law the number of transistors on semiconductor chips has increased, following an exponential curve and in recent decades this number has increased drastically [1]. Along with the increase in the number of devices integrated into these chips, the power density has also increased substantially leading to undesirable temperature gradients and hot-spot formation [2]. High-temperature gradients and hot-spot formations have led to a degradation in the performance and reliability of CPU and GPU technology [3], [4]. A more effective thermal management is thus vital to assist in protecting these chips and improving their performance. Since the beginning of integrated-circuit (IC) technology, more than 60 years ago, thermal simulations for semiconductor chips have been modeled using lumped RC thermal circuits that, although offering efficient thermal analysis, do not provide accurate thermal prediction and very often fitting/scaling factors are needed to improve the accuracy [5], [6]. In addition, the poor resolution in the lumped elements of the RC thermal circuit models is incapable of capturing submicron- or nano-scale hot spots. To obtain high resolution and accurate prediction of the thermal distribution on the semiconductor chips, direct numerical simulations (DNSs) based on finite difference [7]-[9] or finite element methods [10]-[12] are usually needed. These approaches are however computationally time-consuming and are prohibitive for thermal simulation at the architecture level. We have recently developed an innovative thermal simulation approach based on model order reduction enabled by a data-driven learning process [13], [14] for semiconductor ICs. The approach is able to offer a thermal prediction with the efficiency of the RC thermal model and the accuracy of the DNS. In this work, the approach is applied to the thermal simulation of a selected CPU at the architecture level.

The developed approach utilizes proper orthogonal decomposition (POD) [15], a projection-based reduced-order model that generates the basis functions (or modes hereafter) and projects the system partial differential equation(s) onto a functional space described by these modes [13]-[18]. The POD generates orthogonal modes from solution data of the DNSs in a domain subjected to the parametric variations. For heat transfer problems, these usually include variations of spatial/temporal power sources and boundary conditions (BCs). In this study, DNSs are performed in an open-source computational platform known as FEniCS [19]. The POD approach optimizes the modes from the thermal data specifically tailored to the geometry and parametric variations of the problem using a data-driven learning process. With the heat conduction equation projected to the POD modes, the approach can significantly reduce the degree of freedom (DoF) needed to accurately predict the thermal profile in the domain.

II. BRIEF OVERVIEW OF PROPER ORTHOGONAL DECOMPOSITION

POD generates a set of modes from spatial/temporal thermal data accounting for the parametric variation. This is done by maximizing the mean square inner product of the thermal data and the modes in the domain Ω,

978-1-7281-7571-3/20/$31.00 ©2020 IEEE

$$\left\langle \left(\int_\Omega T(\vec{r},t)\,\varphi\, d\Omega\right)^2\right\rangle \Big/ \int_\Omega \varphi^2 d\Omega, \tag{1}$$

where $\langle\cdot\rangle$ indicates an average. Each mode derived from this maximization process contains the maximum least squares (LS) information of the thermal behavior described by the thermal data [20], [21]. The generated modes constitute an orthogonal functional space and can offer the best LS solution provided by the thermal data with a very small number of modes (i.e., DoF).

Applying variational calculus to (1), this problem is reformulated to the Fredholm equation,

$$\int_{\vec{x}'} \boldsymbol{R}(\vec{r},\vec{r}')\vec{\varphi}(\vec{r}')d\vec{r}' = \lambda\vec{\varphi}(\vec{r}), \tag{2}$$

where $\boldsymbol{R}(\vec{x},\vec{x}')$ is a two-point correlation tensor given by

$$\boldsymbol{R}(\vec{r},\vec{r}') = \langle T(\vec{r},t) \otimes T(\vec{r}',t)\rangle, \tag{3}$$

and λ is the POD eigenvalue of **R** and represents the mean squared temperature captured by the corresponding POD mode. This decomposition process leads to an eigenvalue problem represented by (2) for **R.** Once the POD modes are found, the temperature $T(\vec{x},t)$ can be represented by a linear combination of the POD modes,

$$T(\vec{r},t) = \sum_{j=1}^{M} a_j(t)\,\varphi_j(\vec{r}) \tag{4}$$

where M is the selected number of modes for the temperature solution and a_j is the time-dependent coefficient for each mode. To generate POD modes and eigenvalues more efficiently, the method of snapshots [13], [14], [18] is applied.

Using the Galerkin projection method, a set of equations for a_j is derived by projecting the heat conduction equation onto an eigenspace,

$$\int_\Omega \left(\varphi\frac{\partial\rho CT}{\partial t} + \nabla\varphi\cdot k\nabla T\right)d\Omega$$
$$= \int_\Omega \varphi P_d(\vec{r},t)\,d\Omega - \int_S \varphi(-k\nabla T\cdot\vec{n})dS, \tag{5}$$

where k is the thermal conductivity, P_d the power density, ρ the density, C the specific heat, S the boundary surface, $\vec{n}$ the outward normal vector of the surface, and (-$k\nabla T$) the heat flux on the surface. With a selected number of modes M, the spatial integrals in (5) can be pre-evaluated to construct a set of M ordinary differential equations (ODEs) for a_j,

$$\sum_{j=1}^{M} c_{i,j}\frac{da_j}{dt} + \sum_{j=1}^{M} g_{i,j}a_j = P_{pod,i},\ i = 1\ \ to\ \ M, \tag{6}$$

where $c_{i,j}$ and $g_{i,j}$ are elements of the thermal capacitance and conductance matrices in the POD space and given by

$$c_{i,j} = \int_\Omega \rho C\varphi_i\varphi_j\,d\Omega\ \ \text{and}\ \ g_{i,j} = \int_\Omega k\nabla\varphi_i\cdot\nabla\varphi_j\,d\Omega. \tag{7}$$

$P_{pod,i}$ represents the projected power density in Ω and heat flux across S along the ith POD mode and is given by

$$P_{pod,i} = \int_\Omega \varphi_i P_d(\vec{r},t)\,d\Omega - \int_S \varphi_i(-k\nabla T\cdot\vec{n})\,dS. \tag{8}$$

The crucial but time-consuming steps for developing a POD thermal model for thermal simulation of a CPU include (i) collection of thermal data from DNSs of the selected CPU, (ii) generation of the POD modes given in (2), and (iii) evaluation of the POD model parameters in (7).

III. Development of POD Models for Thermal Simulation of a CPU

A. Thermal data collection

The Alpha EV6 processor shown in Fig. 1 [22] is selected for this work. Dynamic thermal simulations of the selected CPU are performed from a finite element simulator FEniCS [19] to collect thermal data that are needed for generating (or training) the POD modes. Adiabatic boundary conditions are applied to the surfaces of the chip except for the bottom of its substrate where the Robin boundary condition is used with ambient temperature taken as 0°C. In Fig. 1, A, B and C are the locations of the functional units where the uniform power pulses are applied. Each power source with a period of 3.333µs represents an average, over 10k cycles at 3 GHz. Thermal simulation of the CPU structure with meshes of 129×129×7 is performed in FEniCS for 3ms.

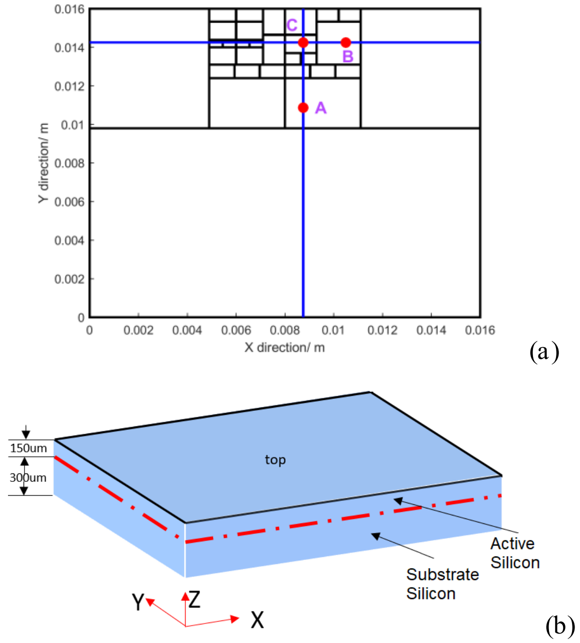

Fig. 1. (a) Floorplan of the CPU chip (Alpha EV6 processor) with dimensions and (b) its 3D structure.

B. Generation of POD mode and evaluation of model parameters

Once the thermal data are collected from the FEniCS simulation of the CPU, the method of snapshots [13], [14], [18] is applied to solve (2) for POD modes and eigenvalues. The eigenvalue λ of each mode solved from (2) represents the mean squared temperature variations captured by the mode and thus reveals the importance of the mode. The eigenvalue for the collected data is displayed in Fig. 2 which shows that the eigenvalue decreases by 4 orders of magnitude from the first POD mode to the 3rd mode and nearly 6 orders to the 5th mode. This strongly indicates that the POD approach would provide very good accuracy with 3 to 5 modes. The eigenvalue curve becomes nearly flat beyond the 17th mode due to the machine precision.

With the modes determined from (2), the model parameters, $c_{i,j}$ and $g_{i,j}$ in (7) and power density in the POD space $P_{pod,i}$ in (8) need to be evaluated to solve a_j from the ODEs given in (6). The spatial and dynamic temperature on the CPU chip can then be determined from (4). It should be noted that the resolution of the temperature solved from the POD approach is determined by the POD modes φ_i whose resolution is identical to the DNS.

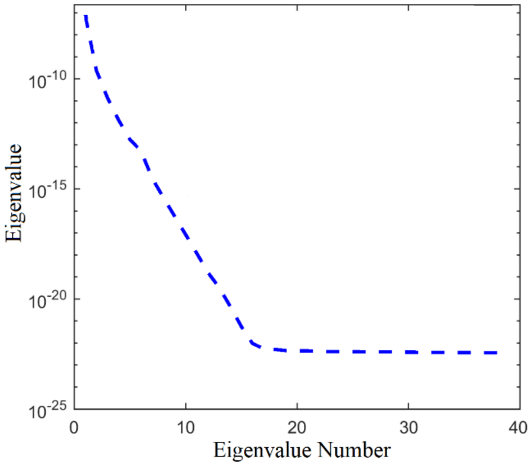

Fig. 2. The eigenvalue spectrum for the thermal data generated from the FEniCS thermal simulation of the CPU.

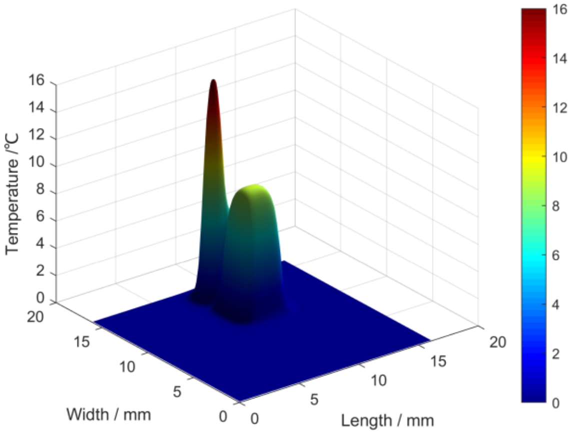

Fig. 3. Thermal distribution of the CPU subjected to 3 power sources indicated in Fig. 1.

IV. POD Thermal Simulation Of The Selected CPU

The developed POD approach presented in Section III is applied to the thermal simulation of the CPU structure given in Fig. 1. The results are compared to the thermal FEniCS simulation of the same CPU. For a meaningful comparison, power sources and BCs are identical between these 2 approaches. Power sources are located in the same functional units shown in Fig. 1. The thermal profile of the CPU at t = 3ms (see Fig. 4(a)) on the device layer, where the power sources are located, is illustrated in Fig. 3.

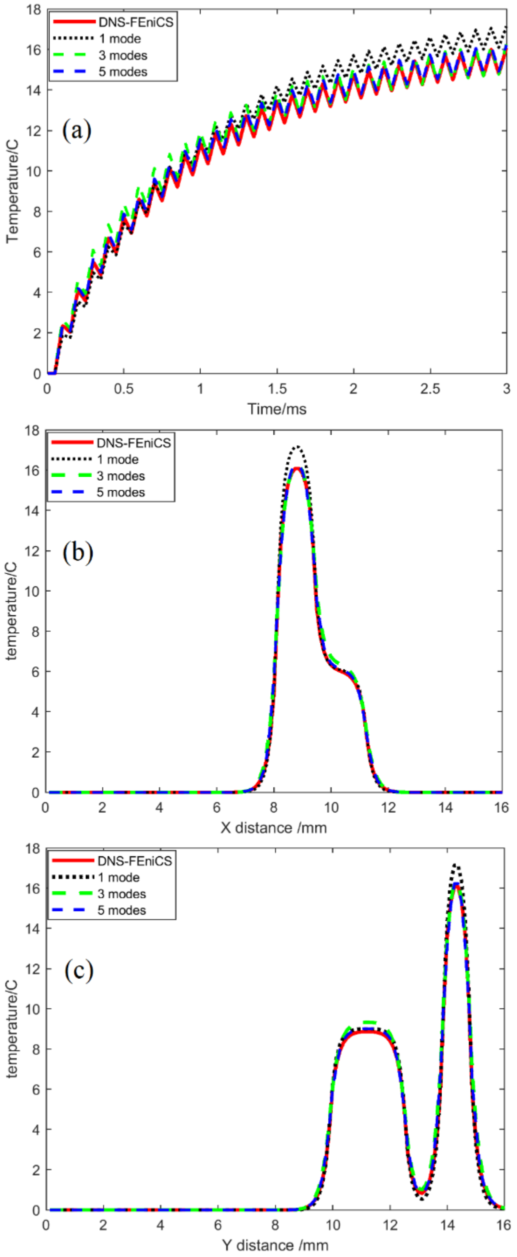

Fig. 4. Temperature distributions derived from the POD models, each with a different number of modes, compared to FEniCS simulation. (a) Dynamic temperature at the hot spot, in the Functional unit C, shown in Fig. 1. (b) The temperature distribution in the *x*-direction at t = 3*ms* through the hot spots C and B. (c) The temperature distribution in the *y*-direction at t = 3*ms* through the hot spots, C and A. The temperature values, included in these figures, are given as the number of degrees above the ambient temperature.

POD simulation results with different numbers of modes are illustrated in Figs. 4(a)-4(c) compared to the finite element simulations. The dynamic temperature evolution in the functional unit C (see Fig. 1) is shown in Fig. 4(a). Although the one-mode POD provides the dynamic thermal prediction with a large LS error, the inclusion of 3 modes in the POD approach offers a good agreement with the FEniCS result. Results from these 2 approaches are indistinguishable when 5 modes are included. Such a small DoF can be achieved by the POD approach is clearly indicated by the substantial decrease in the eigenvalue shown in Fig. 2.

Spatial profiles of temperature on the device layer along the *x* and *y* directions (see Fig. 1) reveal a similar capability to that of the POD approach, as shown in Figs. 4(b) and 4(c) when comparing the POD and FEniCS simulation results. That is, the 3 -mode POD model provides an accurate prediction, compared to the FEniCS simulation. Also, the 5-mode POD results are in excellent agreement with FEniCS simulation results. In addition, the resolution derived from the POD models is as fine as that in FEniCS simulation.

V. CONCLUSION

A reduced-order thermal model based on POD has been applied to the thermal simulation of a CPU structure subjected to several dynamic power sources. It is found the POD model with only 3 modes presents a thermal prediction of a CPU with good accuracy, compared to DNSs using the finite element method from FEniCS. With 5 modes included in the POD model, an excellent agreement with FEniCS can be achieved. This amounts to a reduction of numerical DoF by more than 4 orders of magnitude. In cases where finer resolutions are needed, the reduction in the DoF will be even greater. With such a small DoF, even smaller than the lumped thermal circuit model, the POD not only offers thermal prediction as accurate as that of the DNSs, but also provides resolution as fine as that of the DNSs. This is compared to the conventional thermal simulations at the architecture level where only a lower resolution can be achieved due to the simulation efficiency. The developed POD technique offers an innovative approach that will offer accurate and efficient thermal simulations for CPUs and GPUs with a finer resolution that has been unattainable in the past.

ACKNOWLEDGMENT

This work is supported by National Science Foundation under Grant Nos, OAC-1852102 and ECCS-2003307.

REFERENCES

[1] https://github.com/karlrupp/microprocessor-trend-data

[2] Sultan, Hameedah, Anjali Chauhan, and Smruti R. Sarangi. &quot;A survey of chip-level thermal simulators.&quot; ACM Computing Surveys (CSUR) 52.2 (2019): 1-35.

[3] A. Heinig, R. Fischbach, and M. Dietrich, "Thermal analysis and optimization of 2.5D and 3D integrated systems with wide I/O memory," in Proc. Conf. Therm. Thermomech. Phenom. Electron. Syst. (ITherm), Orlando, FL, USA, May 2014, pp. 86–91.

[4] J. Zhou, J. Yan, K. Cao, Y. Tan, T. Wei, M. Chen, G. Zhang, X. Chen, S. Hu, "Thermal-aware correlated two-level scheduling of real-time tasks with reduced processor energy on heterogeneous MPSoCs", J. Systems Architecture, 82, 1-11, 2018.

[5] K. Skadron, M.R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, D. Tarjanacm, "Temperature-Aware Microarchitecture: Modeling and Implementation", ACM Trans. Architecture & Code Optimization, Vol. 1, No. 1, March 2004.

[6] W. Huang, K. Sankaranarayanan, R.J. Ribando, M.R. Stan, K. Skadron, "An Improved Block-Based Thermal Model in HotSpot 4.0 with Granularity Considerations", WDDD'07, San Diego, CA, June 2007.

[7] Zhang, Runjie, Mircea R. Stan, and Kevin Skadron. &quot;Hotspot 6.0: Validation, acceleration and extension.&quot; University of Virginia, Tech. Rep (2015).

[8] Elsawaf, M. A., H. A. Fahmy, and A. L. Elshafei. &quot; CPU dynamic thermal management via thermal spare cores.&quot; 2009 25th Annual IEEE Semiconductor Thermal Measurement and Management Symposium. IEEE, 2009.

[9] Jiang, Zhong Hua, Ning Xu, and Chun Xiang Wu. &quot;Thermal floorplan base on conjugate gradient solver in hotspot.&quot; Applied Mechanics and Materials. Vol. 608. Trans Tech Publications Ltd, 2014.

[10] Chen, Chyi-Tsong, Ching-Kuo Wu, and Chyi Hwang. &quot;Optimal design and control of CPU heat sink processes.&quot; IEEE Transactions on Components and Packaging Technologies 31.1 (2008): 184-195.

[11] Al-Rashed, Mohsen H., et al. &quot;Investigation on the CPU nanofluid cooling.&quot; Microelectronics Reliability 63 (2016): 159-165.

[12] Liu, Yong Zhen, et al. &quot;Application of Thermal Contact Resistance in Simulation of Heat Dissipation by CPU Heat Sinks Based on ANSYS.&quot; Advanced Materials Research. Vol. 941. Trans Tech Publications Ltd, 2014.

[13] W. Jia, B. T. Helenbrook, and M.-C. Cheng, "Thermal modeling of multi-fin field effect transistor structure using proper orthogonal decomposition," IEEE Trans. Electron Devices, vol. 61, no. 8, pp. 2752–2759, Aug. 2014.

[14] W. Jia, B. T. Helenbrook, and M.-C. Cheng, "Fast Thermal Simulation of FinFET Circuits Based on a Multiblock Reduced-Order Model," IEEE Trans. CAD ICs. Syst, vol. 35, no. 7, pp. 1114–1124, Jul. 2016.

[15] J. L. Lumley, "The structure of inhomogeneous turbulent flows," Atmos. Turbulence Radio Wave propag., pp. 166–178, 1967.

[16] T. Borggaard, "Proper orthogonal decomposition for reduced order control of partial differential equations," Ph.D. dissertation, Dept. Math., Virginia Polytechnic Inst., Blacksburg, VA, USA, 2000.

[17] A. Chatterjee, "An introduction to the proper orthogonal decomposition," Current Sci., vol. 78, no. 7, pp. 808–817, 2000.

[18] L. Sirovich, "Turbulence and the dynamics of coherent structures. I–Coherent structures. II–Symmetries and transformations. III–Dynamics and scaling," Quart. Appl. Math., vol. 45, pp. 561–571 and 573–590, Oct. 1987.

[19] https://fenicsproject.org/

[20] T. Hummel and A. Pacheco-Vega, "Application of Karhunen-Lo`eve expansions for the dynamic analysis of a natural convection loop for known heat flux," J. Phys. Ser., vol. 395, no. 1, 2012, Art. ID 012121.

[21] V. Algazi and D. Sakrison, "On the optimality of the Karhunen-Lo`eve expansion (Corresp.)," IEEE Trans. Inf. Theory, vol. 15, no. 2, pp. 319–320, Mar. 1969.

[22] http://lava.cs.virginia.edu/HotSpot/