# Composable Thermal Modeling and Characterization for Fast Temperature Estimation

Hai Wang*, Duo Li*, Sheldon X.-D. Tan*, Murli Tirumala‡ and Ashish Gupta†
*Department of Electrical Engineering, University of California, Riverside, CA 92521
‡Intel Corporation, Portland, OR 97124
†Intel Corporation, Chandler, AZ 85226

***Abstract*—Efficient temperature estimation is critical for designing thermal efficient, low power and robust integrated circuits in nanometer regime. Thermal simulation starts from the detailed thermal structures by solving thermal diffusion equations no longer meets demanding tasks for efficient design space exploration. Compact and composable model-based simulation provides a viable solution to this difficult problem. However, building such thermal models from detailed thermal structures was not well addressed in the past. In this paper, we propose a new thermal compact modeling techniques for fast thermal analysis in the context of multi-core microprocessors design. The new approach builds the models from detailed structures for each core using finite difference method and reduces the model complexity by sampling-based model order reduction and circuit realization techniques. To improve the reduction efficiency, number of ports of thermal models are first reduced by port merging, which actually leads to coarse grids at the boundaries. The resulting thermal circuits can be simulated by general circuit simulator SPICE. Experimental results on a quad-core microprocessor architecture show that the new approach can easily build accurate thermal systems from the composite compact models. The new thermal systems lead to order of magnitude speedup over standard finite difference models in transient thermal simulation.**

## I. Introduction

Excessive on-chip temperature can cause many severe problems such as reduced reliability of chips, elevated cooling cost of the packaging [3], [8]. Thermal management and related design problems continue to be identified by the Semiconductor Industries Association Roadmap [1] as one of the five key challenges during the next decade for achieving the projected performance goals of the industry. Thus, accurate and efficient thermal modeling and analysis are vital for the thermal-aware VLSI design [10] to improve performance, reliability, power reduction as well as online temperature regulation techniques [3], [13].

Traditional thermal analysis solves the partial thermal diffusion equation directly using numerical approaches such as FEM (finite element), FDM (finite difference), and CFD (computational fluid dynamics). These approaches are accurate given the detailed thermal structures. However, the resulting equation sizes can be prohibitively large for design exploiting. Hence, thermal simulation starts from the detailed thermal structures by solving thermal diffusion equations no longer meets the demanding design tasks for efficient design space exploration. As thermal effects become first-class design constraints, efficient thermal analysis calls for much more efficient solutions. Compact and *composable* model-based simulation provides a viable solution to this difficult problem. This is similar to the model-based electronic circuit simulation (like SPICE, which no longer solves Poisson's equations directly at device level to obtain voltage and current information). In addition, just like the device models for CMOS and BJT transistors, new compact and *composable* thermal modeling technique can be easily connected to build various circuits and systems.

Many compact static and transient thermal modeling methods at different levels (parts, package, board) have been proposed in the past [5], [7], [9], [12]. One important problem, which was not well solved in the existing works, is to build *composable* thermal models for fast thermal design exploration at architecture and package levels. In this paper, we try to address this emerging problem and propose a novel composable thermal modeling approach. We demonstrate the new approach in the context of fast thermal analysis and design for multi-core microprocessors at the architecture level. The new approach builds the compact thermal models for the basic building blocks (CPU and cache cores) from the detailed thermal models generated by finite difference method. It applies sampling-based reduction technique to reduce the complexity of the model. To make the complexity reduction possible or efficient for thermal models with many ports, we propose to reduce the ports of the thermal modules by port merging technique, which leads to fine grids inside modules and coarse grids at the boundary. The course grids can be justified by smoother or smaller thermal gradients at boundary. More importantly, coarse grids can effectively reduce the number of ports for the thermal building blocks. Otherwise, it would be difficult for existing model reduction techniques because of massive numbers of ports.

Experimental results on a quad-core microprocessor architecture show that the new approach can easily build accurate thermal circuits from the compact models for different cores for fast architecture thermal analysis and optimization. The compact models lead to order of magnitude speed up over the standard finite difference method with marginal error.

## II. Thermal simulation problem from first principles

At the circuit, package and board level, the heat transfer phenomena is governed by the following heat differential equation [4]:

$$\rho C_p \frac{\partial T(\vec{r},t)}{\partial t} = \nabla \cdot [\kappa(\vec{r},T) \cdot \nabla T(\vec{r},t)] + g(\vec{r},t) \quad (1)$$

This work is supported in part by NSF grant under No. CCF-0448534, in part by NSF Grant under No. CCF-0902885, in part by Semiconductor Research Corporation (SRC) grant under No. 2009-TJ-1991.

which is subject to the following general thermal boundary condition (Robin's boundary condition)

$$\kappa(\vec{r},T)\frac{\partial T(\vec{r},t)}{\partial n_i} = h_i(T(\vec{r},t) - T_{amb}) \quad (2)$$

In (1), $T$ $(K)$ is the temperature, $\rho$ $(Kg/m^3)$ is the density of the material, $C_p$ $(J/m^3{\cdot}K)$ is the mass heat capacity, $\kappa$ $(W/m{\cdot}K)$ is the thermal conductivity, and $g$ $(W/m^3)$ is the heat energy generation rate. In (2), $n_i$ is the outward direction normal to the boundary condition $i$, $h_i$ $(W/m^2K)$ is the heat-transfer coefficient (for convective interface), and $T_{amb}$ is the ambient temperature surrounding the thermal systems. If $h_i = 0$, the boundary condition is adiabatic, otherwise, it is convective. Note that different materials will have different thermal conductivity $\kappa$ and it may also depend on temperature.

Finite difference or finite element methods can be used to solve (1). However, they are very expensive to solve or even prohibitive for large $n$. It will be even worse for thermal design exploration for various multi-core architectures as the spatial discretization and simulation have to be done for each architecture during the optimization steps.

## III. New thermal simulation by compact thermal modeling for multi-core systems

Instead of solving the whole thermal system in (1), a more efficient compact thermal-model based approach will be more desirable. As shown in Fig. 1(a), we first build two models for CPU core and cache core. Then, using the two models, we can build the multi-core thermal system such as the quad-core system shown in Fig. 1(b) or other multi-core thermal systems shown in the experimental section. We assume that the CPU cores are the same for simplicity, but our approach can be easily extended to different CPU cores and other functional cores. Fig. 2 is the literal structure view of a typical package for the multi-core system. Typically the heat generated at the die are conducted from its back to the heat spreader and then to the heat sink. For simplicity, we assume other sides of the die do not have heat exchange (adiabatic condition). The goal is to build compact thermal models for each building blocks (CPU, caches) so that we can quickly build different multi-core architectures for fast thermal validation instead of building the whole thermal systems from scratch via some meshing techniques.

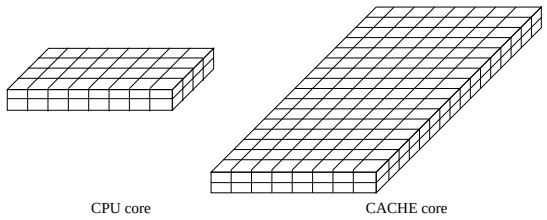
### A. Port reduction by port merging

For a structure, after the space discretization, we will end up with ordinal differential equations. Specifically, if we have $n$ discretized elements (grids) with specific boundary conditions, equation (1) becomes a linear ordinary differential equation

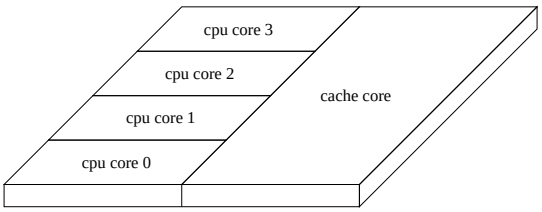
$$C\frac{dT(t)}{dt} + GT(t) = Bg(t) \quad (3)$$

where $C \in \mathbb{R}^{n\times n}$ is the thermal capacitance matrix, $G \in \mathbb{R}^{n\times n}$ is the thermal conductance matrix. $B \in \mathbb{R}^{n\times p}$ is position matrix for a total of $p$ ports including the boundary ports and power sources.

To reduce the order or complexity of such system, model order reduction techniques can be applied. However, all the exiting reduction techniques such as Krylov subspace, or truncated balanced realization methods can not deal with circuits




(a) CPU and cache cores.




(b) A quad-core architecture.

Fig. 1. CPU core, cache core and a quad-core architecture.

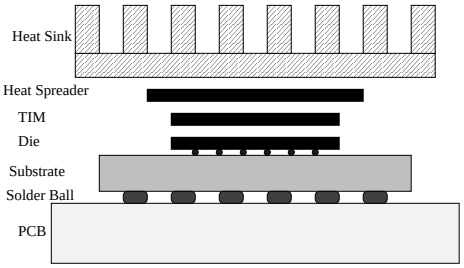



Fig. 2. The literal structure view of the multi-core system package.

with large ports [2]. As a result, port reduction becomes necessary.

In our thermal modeling program, one observation is that the temperature distribution at the boundary is smoother since the boundary is generally far away from the heat sources. As a result, we can merge some adjacent ports to form a single port. In the following, we use a $2\times2\times2$ meshed structure example (in finite difference scheme) shown in Fig. 3 to illustrate the idea.[1] For this meshed structure, There are $8$ elements (cubes) denoted by light solid circles and $24$ ports by hollow circles. Please note, in this case, every element is on the boundary and is the vertex of the cube, thus, each of them has 3 ports connected. To reduce the number of ports, we can merge 4 adjacent ports into 1 port shown in Fig. 3 as dark solid circle, and the number of ports will be reduced to 6.

### B. Boundary conditions for composibility

The thermal models need to be constructed such that system connected by these modules is the same as the system constructed directly from the original structure. It turns out that to make the thermal model composable, the *adiabatic* thermal conditions (special Neumann's boundary condition) should be

[1]For simplisity, we ignore the internal power source in this example. Also, in Fig. 3, capacitor at each node is not displayed and only the ports on the front three faces are shown.

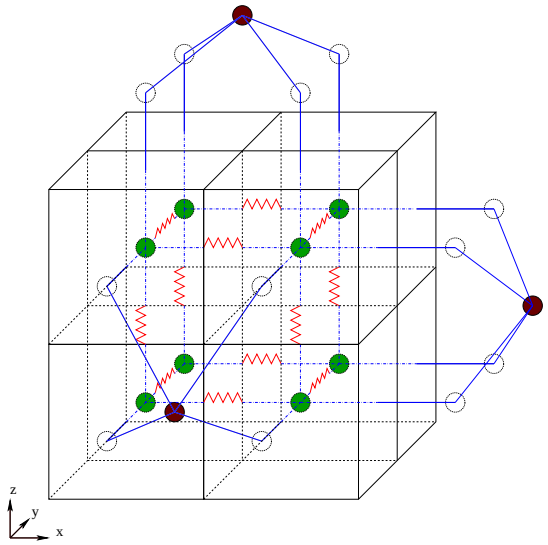



Fig. 3. A $2 \times 2 \times 2$ meshed structure case where the boundary faces (ports) are merged.

added at the thermal modules.

$$\kappa(\vec{r}, T)\frac{\partial T(\vec{r}, t)}{\partial n_i} = 0 \tag{4}$$

Specifically, for the *adiabatic* condition, there is no thermal exchanges between the boundary nodes and outside. However, when two modules are connected together, the connected boundary nodes of the two modules will become one node in the connected system. Heat will flow via normal diffusion as there is no boundary between the two modules. The interface with *adiabatic* thermal conditions is called *composable interface* in this paper.

For other boundaries, which interact with outside directly via convection or other heat exchange mechanism, a proper thermal conditions (Robin's boundary conditions) in terms of equivalent thermal resistances and independent sources can be added at the ports [4].

### C. Model complexity reduction and realization

Reducing the complexity of linear dynamic systems by means of model order reductions have been studied intensively for reducing parasitic electronic circuits in the past [2]. For compact thermal modeling, Krylov subspace based approaches have been applied to reduce the large models [5], [6]. In this paper, we apply more accurate sampling-based reduction technique, which is based on global accurate truncated balanced realization (TBR) reduction scheme. Sampling-based methods [11], [14] try to mitigate the high computational cost of standard TBR method, where the Gramians are approximated using Monte-Carlo sampling approach.

After the reduction, we want the reduced models can be realized into SPICE-friend circuits. This can be done by further diagonalizing the reduced system matrices via generalized eigen-decomposition, and realizing into RC circuits with controlled sources. Then, the realized reduced system is simulated by SPICE-type simulators [5].

In the new approach, the thermal modules and their reduced models are realized into SPICE compatible format using SPICE *.subckt* command. After this, we can build different multi-core architectures (their thermal circuits) on top of these basic thermal building-block modules in SPICE netlists.

## IV. Experimental results

The proposed method has been implemented in MATLAB. First, we build the finite difference models and their reduced composable models for single CPU core and single cache core using two-grid discretization based finite difference method and sampling based model reduction technique. Then, we compose quad-core systems using the original and the reduced composable CPU cores and cache cores. Finally, thermal transient simulation is performed using HPSICE on Linux server with Intel Quad-core CPU and 16GB memory to obtain the temperature distribution for both original and reduced composite thermal systems.

To build composable model for CPU and cache cores, we set up the size of the discretization grid as $32 \times 16 \times 8$ for CPU core ($8mm \times 4mm \times 2mm$) and $64 \times 32 \times 8$ for cache core ($16mm \times 8mm \times 2mm$), in order to keep both CPU and cache cores sharing the same discretization step value $\Delta x = \Delta y = \Delta z = 0.25mm$. This is because the length and width of cache core are twice of the CPU core while the height is the same. Here, we choose thermal conductivity $\kappa = 149W/(m°C)$, material density $\rho = 2300Kg/m^3$ and specific heat $c_p = 700J/(Kg°C)$.

Fig. 4 shows temperatures at the internal power sources in 3D formats. Between the power sources, the temperatures are unknown due to the reduction, as a result, they are marked as zero.

The temperature distributions at the top surface are shown in Fig. 5. As we can see the highest temperatures are in the left hand side. This is expected as more CPU cores are put into the left hand side.

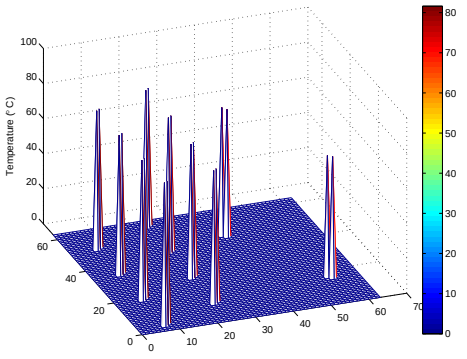



Fig. 4. 3-D temperature distribution at the power sources for the quad-core architecture.

Fig. 6 compares the transient simulation accuracy between the finite difference method and the composable modeling method. The two transient results are almost the same.

The CPU time results for different configurations of cores are shown in Table I. In the table, xx *org* means the original models from finite difference method and xx *mod* means the reduced systems with composable models. It can be seen that with the reduced thermal system, we can achieve about two

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| Circuit | #Node | #Elem | Run time (s) | | Memory | Speed up | |
| | | | Trans | Total | (mb) | Trans | Total |
| 2-core org | 11825 | 46428 | 40.5 | 78.3 | 27 | | |
| 2-core mod | 881 | 1756 | 0.11 | 0.33 | 2.5 | 368 | 237 |
| 4-core org | 23649 | 92852 | 85.2 | 237.0 | 49 | | |
| 4-core mod | 1761 | 3508 | 0.23 | 1.0 | 4.7 | 370 | 237 |
| Quad-core chip org | 45713 | 183290 | 219.6 | 724.2 | 100 | | |
| Quad-core chip mod | 5765 | 11508 | 1.2 | 31.7 | 16.5 | 183 | 23 |

TABLE I
CPU TIME COMPARISONS BETWEEN THE ORIGINAL MODELS (XX *org*) AND REDUCED THERMAL SYSTEMS USING COMPOSABLE MODELS (XX *mod*)

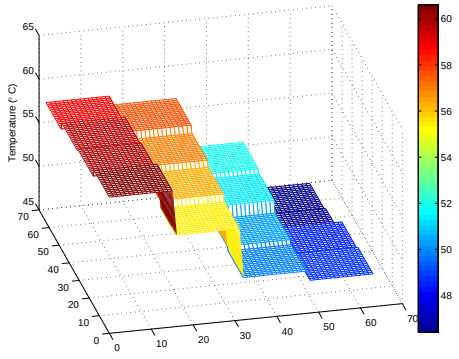



Fig. 5. 3-D temperature distribution at the convective surface for the quad-core architecture.

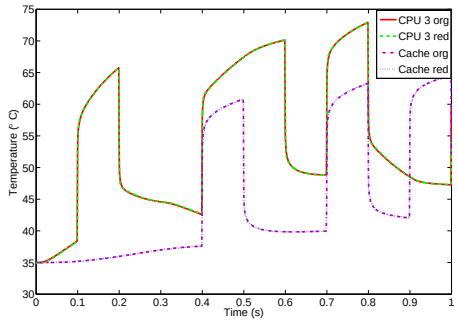



Fig. 6. Accuracy comparison with the full model at some power nodes for the quad-core architecture.

order of magnitude speed up for the transient simulation of the multi-core systems. In addition, the reduced thermal systems lead to much smaller memory footprint than the original models.

## V. CONCLUSION

In this paper, we have proposed a new thermal compact modeling techniques for fast thermal analysis. The new approach builds the models from detailed structures by the finite difference method for each modules and reduces the model complexity by sampling based model order reduction technique. To improve the complexity reduction efficiency, port reduction by adjacent port merging has been proposed, which leads to coarse grids at the boundary. This scheme also makes thermal modules more composable for building large thermal systems. We also studied the boundary conditions for composition of models and circuit realization techniques for easy model generation and simulation. Experimental results on a quad-core microprocessor architecture show that the new approach can easily build accurate thermal circuits from the composable compact models for different cores. The reduced composite models lead to order of magnitude speedup over standard finite difference models.

## REFERENCES


[1] "International technology roadmap for semiconductors(ITRS)," 2009 update, http://public.itrs.net.

[2] A. C. Antoulas, *Approximation of Large-Scale Dynamical Systems*. The Society for Industrial and Applied Mathematics (SIAM), 2005.

[3] D. Brooks and M. Martonosi, "Dynamic thermal management for high-performance microprocessors," in *Proc. of Intl. Symp. on High-Performance Comp. Architecture*, 2001, pp. 171–182.

[4] Y.-K. Cheng, C.-H. Tsai, C.-C. Teng, and S.-M. Kang, *Electrothermal Analysis of VLSI Systems*. Kluwer Academic Publishers, 2000.

[5] L. Codecasa, D. D'Amore, and P. Maffezzoni, "An arnoldi based thermal network reduction method for electro-thermal analysis," *IEEE Tran. on Components and Pacakaging Technologies*, vol. 26, no. 1, pp. 186–192, March 2003.

[6] ——, "Boundary condition independent compact models of dynamic thermal networks with many heat sources," in *Thermal and Thermomechanical Phenomena in Electronic Systems*, 2006, pp. 685–689.

[7] Y. C. Gerstenmaier and G. Wachutka, "Rigorous model and network for transient thermal problems," *Mircroelectronics Journal*, vol. 33, pp. 719–725, September 2002.

[8] S. Gunther, F. Binns, D. Carmean, and J. Hall, "Managing the impact of increasing microprocessor power consumption," in *Intel Technology Journal*, First Quarter 2001.

[9] C. Lasance, H. Vinke, H. Rosten, and K.-L. Weiner, "A novel approach for the thermal characterization of electronic parts," in *Proceedings of the IEEE 11th Annual Semiconductor Thermal Measurement and Management Symposium*, 1995, pp. 1–9.

[10] M. Pedram and S. Nazarian, "Thermal modeling, analysis, and management in VLSI circuits: Principles and methods," *Proc. of the IEEE*, vol. 94, no. 8, pp. 1487–1501, Aug. 2006.

[11] J. R. Phillips and L. M. Silveira, "Poor man's TBR: a simple model reduction scheme," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 1, pp. 43– 55, 2005.

[12] M. Rencz, G. Farkas, V. Székely, A. Poppe, and B. Courtois, "Boundary condition independent dynamic compact models of packages and heat sinks from thermal transient measurements," in *Proceedings of the 5th Electronics Packaging Technology Conference*, 2003, pp. 479–484.

[13] K. Skadron, M. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature aware microarchitecture," in *Proc. IEEE International Symposium on Computer Architecture (ISCA)*, 2003, pp. 2–13.

[14] K. Willcox and J. Peraire, "Balanced model reduction via the proper orthogonal decomposition," *AIAA Journal*, 2002.