Full Chip Thermal Simulation

Zhiping Yu[*], Dan Yergeau, and Robert W. Dutton
CISX 335, Center for Integrated Systems, Stanford University, Stanford, CA 94305

Sam Nakagawa, Norman Chang, Shen Lin, and Weize Xie
Computer Systems Lab (CSL), HP Co., Palo Alto, CA 94303

Abstract

A multilayer, full chip thermal analysis is presented. The design of the chip at functional-block level is directly captured to the simulator, allowing the assessment of the chip layout impact on the system performance due to the elevated operational temperature. The heat generation for each block is obtained by running the circuit-level electrical simulation separately on individual functional units. The thermal diffusion equation is then solved based on the actual structure of the chip including substrate and interconnect/insulating layers. Different thermal conductivity can be specified for each material layer. The effect of package on chip temperature distribution is modeled using thermally resistive layers as boundary between the simulated structure and surrounding environment. Proper adjustment of the boundary thermal resistance results in the correct range of simulated temperature distribution as compared to the measured data.

Both physics and implementation for the thermal simulation will be described. The code is applied to the analysis of a realistic design of CPU chip made of SOI technology with up to six metal interconnect layers. A comprehensive review of simulation results will be presented.

[*] Also with CSL, HP. E-mail: yu@ee.stanford.edu, phone: (650) 725-3644, FAX: (650) 725-7731

1 Introduction

With ever increasing chip complexity and size, and clock frequency approaching GHz range, the power dissipation of a state-of-the-art CPU chip could often reach just under 100 watts (W). Even though packaging technique has steadily been improved to “cool” the chip, the thermal effect on the chip performance and reliability is becoming an indispensable design concern.

The complete electrothermal simulation has been used for the analysis of IC devices and local chip area for sometimes [1]. Rarely is found in literature, however, the thermal analysis at the full chip level due to the formidable computation demand. The key difference lies on the scale of the problem. It is not possible to perform transistor-level electrothermal simulation at the chip level, yet the temperature distribution across the chip is a direct design concern. Especially at the placement-and-routing stage of the design, one would be interested in knowing how different placement affects the temperature distribution along critical signal/clock paths. In this work, we present a methodology which allows the full chip thermal analysis and at the meantime leaves rooms for further coupled electro-thermal analysis. In the proposed approach, the heat generation is pre-calculated at the functional block level using electrical circuit simulators, e.g., PowerMill from Synopsys. The thermal diffusion equation is then solved rigorously on the chip level taking into consideration the complete structure of the chip including interconnect layers and package. The temperature distribution simulated can be used for the analysis of signal/clock propagation delay along global paths. With further improvement, the coupled electro-thermal simulation is possible by feeding the temperature information back to the functional blocks for electrical simulation.

After this introduction, the physics involved in the thermal simulation will briefly introduced and the emphasis is on the proper handling of multilayer structure and the thermal boundary condition. A script-based simulator, PROPHET, will be introduced to solve the thermal diffusion equation for real, engineering problems. The code has been interfaced to widely available visualization tools and enhanced by post-processing programs to allow multi-dimensional rendering and probing of the simulation results. Finally, simulation results for a CPU chip using SOI technology is presented in details.

2 Physics for Thermal Analysis of Multilayer Structure

The temperature distribution in a closed structure (e.g, a chip or a device) is governed by the following thermal diffusion equation and proper boundary conditions,

$$\frac{d}{dt}cT(\boldsymbol{r}) = -\nabla \cdot (-\kappa \nabla T(\boldsymbol{r})) + g(\boldsymbol{r}) \tag{1}$$

where T is the temperature (in the context of this paper it being the lattice temperature), c is the specific heat of the material constituting the structure, κ is the thermal conductivity, and g is the heat generation rate. In the steady state, the left hand side (LHS) term in the above equation is dropped, so the only concerned quantities are κ and g. In a general, multilayer structure, κ is position-dependent, i.e., function of $\boldsymbol{r}$. Furthermore, κ is also considered temperature dependent with the following form,

$$\kappa = \kappa_0 (T/300)^{-\alpha} \tag{2}$$

where T in units of Kelvin and as examples, $\kappa_0 = 1.45$ and 0.014 W/K·cm for silicon and oxide, respectively, and $\alpha = 1.2$ for both materials. The source of heat generation depends on the nature of the circuit operation. At the device simulation level, it is the local Joule heat ($\boldsymbol{J} \cdot \boldsymbol{E}$, where $\boldsymbol{J}$ and $\boldsymbol{E}$ are current density and electric field, respectively), and at the block level, it can be assumed that the power consumption for the functional block under the typical signal pattern is the source for the entire block. By averaging the power over the "volume" of the functional block – it is easier to estimate the volume of the block for SOI technology because the thickness of the active silicon thin film can be naturally taken as the depth for the block while for the bulk CMOS technology one has to assume a "skin" depth within which the heat generation is to occur uniformly. Since SOI technology holds great potential in improving the chip performance (by reducing the power consumption and raising the clock frequency) for the same feature size as compared to bulk CMOS technology [2] and the thermal issue for SOI is a bigger concern, in this work we choose the CPU chip built on SOI as the example for analysis.

As an example, consider a hypothetical floating point unit (FPU), which consumes power of 6 W from the circuit simulation. The unit has a layout area of $1800 \times 2600\,\mu$m. If the thickness of the top silicon thin film in an SOI technology is $0.2\,\mu$m, then the volume for the unit is $9.36 \times 10^{-7}\,\mathrm{cm}^3$. Considering the overall block power consumption, the strength (i.e., the rate) of the heat generation source is translated to $6.41 \times 10^6\,\mathrm{W/cm}^3$.

Another critical factor in determining the temperature distribution is the boundary condition of the structure under simulation. For a chip, the environment temperature (say, the room temperature) is imposed through the package. To include the package literally into the simulation is a daunting task due to its complexity and irregular structure. In this simulation, we apply thermally resistive layers, called capping layers, on both the top and bottom surfaces of the chip to serve as the thermal resistance between the simulated chip and the environment (considered as the thermal reservoir due to the constant temperature being maintained). For the side of the chip, the reflective boundary condition is assumed, i.e., no thermal flow out of the chip through side walls. The simplification of this boundary condition is justified considering the fact that the area of side walls (a typical wafter thickness is around $500\,\mu$m vs. the lateral dimension of chip, which easily surpasses 1 cm on each side of the chip) is much smaller than the flat surfaces (top and bottom) of the chip. Both the thermal conductivity and the thickness of capping insulating layers can be used as adjustable parameters to model the effect of the package (i.e., thermal resistance). In this case, the thickness is used as the fitting parameter to have the simulated temperature range fall in what is measured through a test chip, which uses embedded resistors to monitor the on-chip temperature distribution.

The thermal diffusion equation is solved for each material region in the structure, and if there are multilayers in the structure, one needs to specify the constraint for temperature across the material interface. The simplest constraint is to assure the temperature continuous at the interface. We will give a complete description of the system to be solved in the next section.

3 Script-Based IC Process/Device Simulation

In contrast to the existing TCAD (Technology CAD) tools where only material/physical parameters are allowed to change but not the physical system outside of the built-in library, Stanford and Bell Labs of Lucent Technologies have been developing a script-based approach to TCAD in recent years. In this approach, a physical system (i.e., PDEs[1]) is described by an assembly of mathematical terms. Each such term consists of a geometric opertor, which has been implemented in the library in a discretized form (using either finite difference or finite element method) and a physical operand (or called flux) upon which the operator act. There is a library in the code for all these geometric operators and physical operands. If a physical system to be solved is not available in the code, it can be assembled using the "pre-fabricated" terms. This script approach expands the capabilities of TCAD tools tremendously (ability of specifying the system instead of mere variation of physical parameters).

Besides of the specification of PDEs themselves, users also need to describe the boundary condition and possibly the constraints across the material interface. The following is an example of specifying a

[1] PDE: partial differential equation

system of thermal diffusion for a multilayer chip.

```
system name=thermal
+ sysvars=tl
+ tmpvars=kappa
+ term0=dirichlet.default_dirichlet(0|tl)@{exposed/nitride,nitride/bulk}
+ term1=box_div.diffusion(kappa,tl|tl)@{silicon,oxide,nitride}
+ term2=constraint.continuity(tl|tl)@{silicon/oxide,silicon/nitride,oxide/nitride}
+ term3=-1*nodal.self(hg|tl)@{silicon}
+ nterm=4
+ func0=kappa(tl|kappa)@{silicon,oxide,nitride}
+ nfunc=1
```

The detailed syntax can be found in Stanford's web site: `http://www-tcad.stanford.edu/~prophet/` or refer to paper [3]. A few explanations are provided below:

The equation (thermal diffusion) itself is described using `term1` (divergence term in Eq. (1)) and `term3` (heat generation), and the Dirichlet boundary condition (i.e., constant temperature, `tl`) is specified for top (`exposed`) and bottom (`bulk`) of the thermal resistive layers (named `nitride` but with different κ). The continuous condition for temperature across the material interface is specified in `term2`. The function dependence of κ on T (Eq. 2) is specified through `func0`. The complete script will be given in separate journal paper to be submitted.

4 Example

A CPU chip fabricated using SOI technology is analyzed. Under the typical signal pattern and clock frequency it consumes about 65 W. In the simulation, seventeen blocks with total power consumption of 56 W are included. In Figure 1, the heat generation rate is shown across the chip as view from the top of the SOI substrate. It can also be seen that the region where the heat generation rate is the highest is located in the block (the coordinates for the upper-right corner of the block is about $X = 16000, Y = 13000\mu$m) which is FPU.

The simulation is carried out on the entire chip structure (including six layers of interconnect and corresponding insulating layer in between) capped with two thermal resistive layers (`nitride` in the script above) on both top and bottom surfaces. The chip itself consists of (from bottom up) silicon

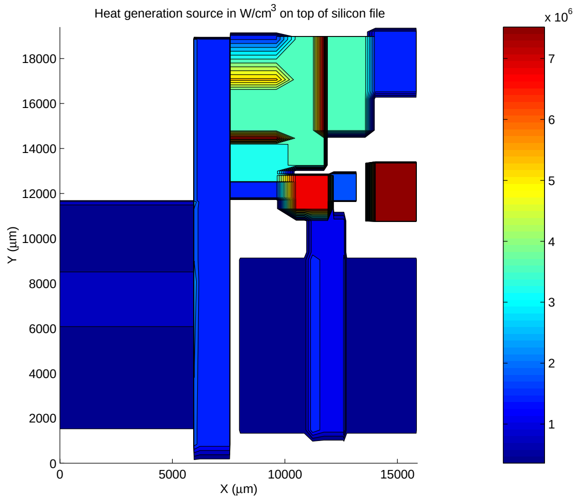
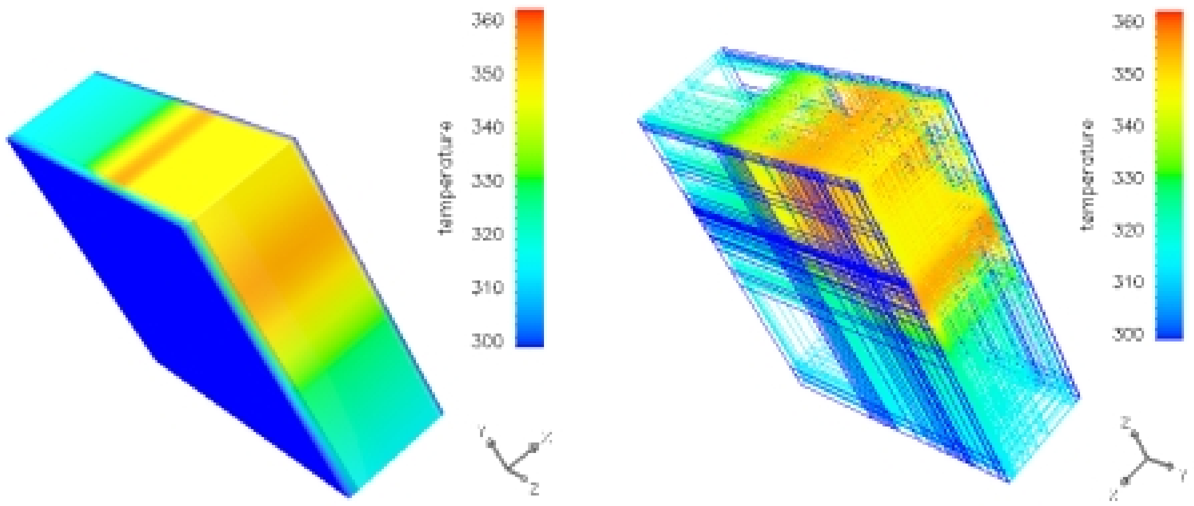


Figure 1: Heat generation distribution viewed from the top of the SOI substrate. The units for the generation rate is W/cm^3.

substrate, buried oxide, and active silicon thin film. The 3D rendering of the simulated structure and temperature distribution is shown in Figure 2(a) together with the mesh (see-through) specification (Figure 2(b)).

The temperature distribution inside of the structure can be probed using 2D or 1D plots. The temperature distribution on the top of the SOI substrate is shown on Figure 3(a) and that on the wafer cross-section as cut along the line of $Y = 12000\,\mu$m is shown in Figure 3(b). It can be seen that the highest temperature spot is about 55 °C above the environment temperature (300K), which is about the same as measurement indicated. It can also be seen that the temperature profile generally follows the pattern of heat generation source (Figure 1) but not exactly due to the thermal diffusion and reflective boundary condition on the side walls. Further tuning of the thermal resistance requires more elaborate measurement setting which is beyond the scope of this work.

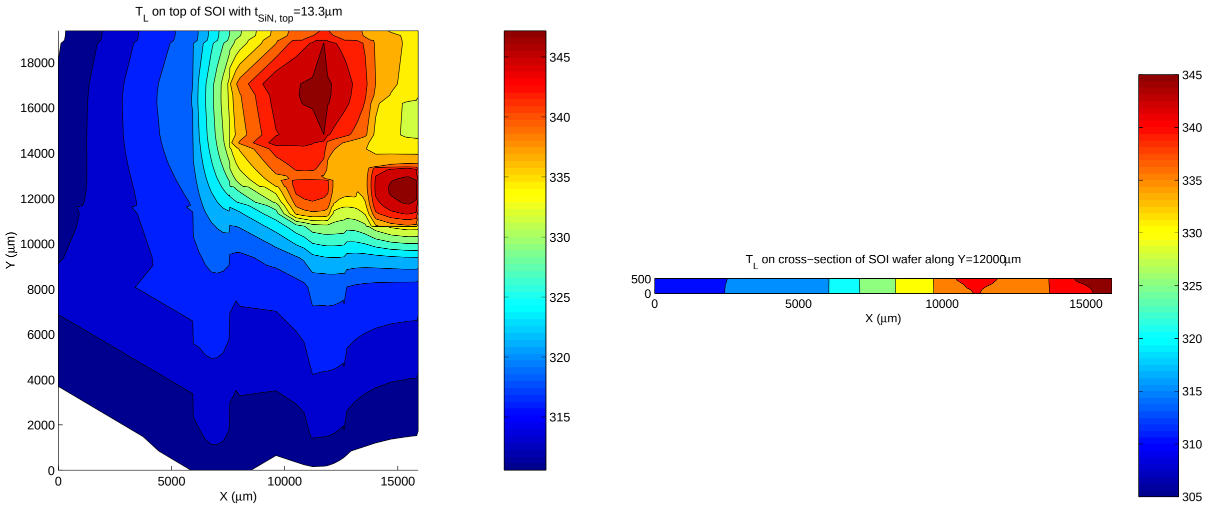
A final note is that in order to assess the impact of the temperature distribution on the electrical performance of the chip, it will be useful to look at the temperature profile along a particular path, say a data path. The temperature variation will inevitably cause the different behavior in signal delay and crosstalk. In Figure 4, it is shown that the temperature profile along a global data path ($X = 7000\,\mu$m in Figure 3(a)), which runs top down on the chip. It can be seen the temperature difference in the same data path can be as large as 25 °C.



(a) Solid rendering of temperature distribution. The $X-$axis points from the surface to the back of the substrate.

(b) Wire representation of structural meshing. Note that the orientation is different from figure on the left.

Figure 2: (a) The simulated structure and 3D contour plot for temperature under the specified heat generation condition in Figure 1; (b) Wireframe plot of mesh used in the simulation and the temperature distribution.



(a) Temperature distribution on substrate top

(b) Temperature distribution at wafer X-section

Figure 3: Contour plots for temperature distribution: (a) on the top of the SOI substrate; (b) at the cross-section of the wafter along the line of $Y = 12000\,\mu$m in Figure 1

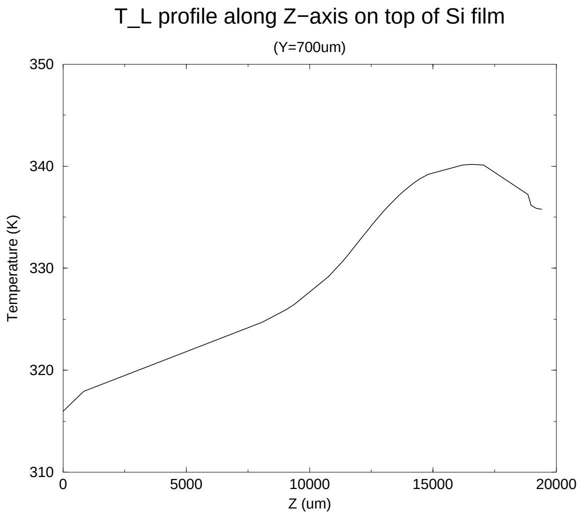


Figure 4: Temperature distribution along the global data path ($X = 7000\,\mu$m in Figure 1).

5 Summary

The methodology and implementation for the full chip thermal simulation is demonstrated using an SOI CPU analysis. The meshing and the distribution of heat generation source of the chip essentially follows the block diagram of the chip, allowing direct assessment of the effect of placement and routing on the thermal distribution. The temperature profile obtained from the simulation is viable for further electrical simulation in such analysis as signal delay and clock skewing.

Acknowledgement: The authors from Stanford would express their great appreciation over the research grant from Interconnect Focus Center.

References

[1] B.H. Krabbenborg, A. Bosma, H.C. de Graaff, and A.J. Mouthaan, “Layout to circuit extraction for three-dimensional thermal-electrical circuit simulation of device structure,” *IEEE Tr. CAD IC and S*, Vol. 15, No. 7, p. 765, July 1996.

[2] E. Leobandung, *et al.*, “Scalability of SOI technology into 0.13μm 1.2V CMOS generation,” *IEDM* ’98 p. 403, San Francisco, Dec. 1998.

[3] C.S. Rafferty, Z. Yu, B. Biegel, M.G. Ancona, J. Bude, and R.W. Dutton, “Mu lti- dimensional quantum effect simulation using a density-gradient model and script- level programming techniques,”

SISPAD '98 p. 137, Lueven, Belgium, Sept. 1998.