Fast Placement-Dependent Full Chip Thermal Simulation

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

Sam Nakagawa and Jeff Deeney[‡]
Computer Systems and Technology Lab (CSTL), HP Labs, HP, Palo Alto, CA 94303
[‡]Business Critical Computing (BCC), HP, Ft. Collins, CO 80525

Norman Chang, Shen Lin, and Weize Xie, Apache Design Solutions, Palo Alto, CA 94306

Abstract—**A general purpose semiconductor device/process simulator, PROPHET [1], has been adapted for full chip thermal analysis and is capable of quickly (~1 minute CPU time) assessing the impact of functional block placement on chip temperature distribution. The key to fast simulation is a new algorithm which maps the heat generation of functional blocks (FPU, e.g.) to a coarse mesh while maintaining conservation of heat generating sources. Design of two high-performance CPU chips based on bulk CMOS and SOI technologies, with total power consumption of 100 and 60 watts, respectively, has been evaluated for thermal performance using this approach. Excellent results have been achieved in terms of benchmarked accuracy and computational efficiency.**

Up to seven interconnect layers have been included in the simulation; effects of packaging are modeled using two capping thermally resistive layers on top and bottom of the chip. Considering the extremely non-uniform nature of interconnect/interleaving-insulating layers, anisotropic thermal conductivity is a critical factor in modeling thermal properties and has been implemented in the simulator [2].

The potential benefit of using pure silicon (Si-28), which has a higher thermal conductivity than that of natural silicon (1.6 times as big at the room temperature), in reducing the peak chip temperature has also been studied. It is shown that with a power consumption level of 100 watts, the peak temperature can be lowered by about 10% (from 136 to 123 °C).

Introduction

With increased number of transistors packed on a single chip, the rapid increase of clock frequencies (already above 1GHz) can easily push chip power consumption over 100 watts in spite of voltage scaling and low-power circuit innovations. Limiting peak operating temperature has become a dominant issue for improving chip reliability and performance. A simulation tool which can simultaneously address heat generation/dissipation and block layout is invaluable in assessing the thermal performance at early stages of design (e.g., floor planning).

Approach

The power consumption for each functional block is calculated using either transistor- or logic-level electrical simulation, and the power is considered to be uniformly distributed within the active silicon film in SOI structure or a thin skin surface layer for bulk CMOS. Thus, the heat generation profile can be obtained based on block placement. The number of blocks in the functional decomposition is typically tens to hundreds. If these blocks are directly used to guide generation of a conforming mesh, the complexity of multiple interconnect and interleaving insulating layers would require tremendous amounts of memory (on the order of millions of unknowns) and CPU time to fully simulate the chip. Instead, a novel scheme is proposed, which maps the heat generation sources to a pre-specified mesh. In order not to omit contributions from any functional block, the power dissipation of each block is assigned to one or more of the neighboring nodes according to its position in the mesh, and the heat generation intensity at each mesh point is weighted by the discretization's control volume. When using this scheme, a rather coarse mesh can be used (e.g. 25-by-25 grid for a 2-by-2 cm chip) without losing significant accuracy. The impact of placement changes on the temperature distribution can readily be observed with very little computational effort (one minute or less in CPU time).

The thermal model includes the temperature dependence of nonlinear thermal conductivities and the multi-layer nature of the structure is rigorously treated. Effect of packages are modeled by two thermally resistive capping layers. Boundary conditions for sidewalls of the chip (thickness around 500 μm) are assumed reflective (i.e., thermally insulated).

Physical Models

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, 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. 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 [3],

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

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. Expression (2) implies that as the environment temperature is elevated, the material thermal conductivity becomes poorer. 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

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

0-7803-6412-0/01/$10.00©2001 IEEE.

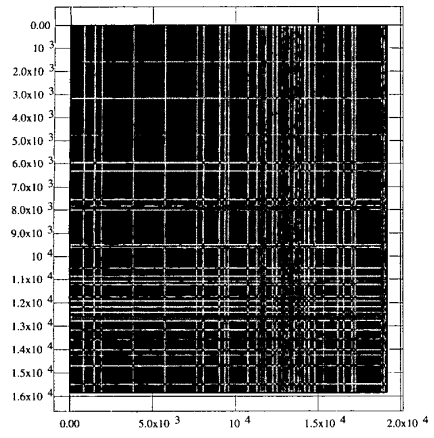

Fig. 1. Non-uniform tensor-product mesh tracking the topology of functional blocks on top plan of a CPU chip

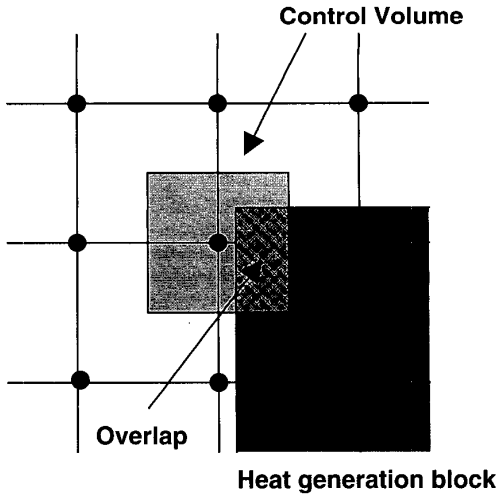

Fig. 2. The contribution of the heat generation block to the node's control volume is scaled by the ratio of the overlap area to that of the control volume.

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.

Numerical Algorithm

Because of the size of the chip (typically 1.5 to 2 cm on each side for the lateral dimensions) and the number of layers (SOI, interconnect, and capping layers), to capture the fundamental feature of the structure and to have a reasonable simulation accuracy, the grid count can easily top half million. The main reason for that type of huge mesh is that a simple approach would require at least four grids for each function block (a rectangle in floorplanning) in order to resolve the distribution of heat generation sources. The number of such functional blocks in modern CPU chips ranges from a few tens to more than one hundred. A more rigorous meshing scheme would need even more grids to make sure the heat generation source won't "spill" out of the block and this requires an encasing frame around the topology of the block, thus doubling the mesh count. An example is shown in Fig. 1. The practice has shown that, with this meshing scheme, only around twenty blocks could be included in the simulation before machine with a 1GB main memory is exhausted. Thus, not all the heat sources can be included during the simulation, leading to inaccurate temperature distribution.

To overcome the above limitation, a scheme of mapping ALL the heat generation sources to a coarse mesh is proposed, which greatly reduces the grid count (and CPU time) while keeping the integrity of heat generation profile. This mapping scheme is described as follows.

For the finite difference method, FDM, in spatial discretization, there is a control volume associated with each grid. A heat generation intensity (in units W/cm^3) is assigned to each node and assumed the same value everywhere within the control volume. The total heat generated from this volume is its volume integration. So after the mesh is fixed, to distribute the heat source from each functional block one has to decide which case it is and make corresponding decision:

1. If the block, which is a 3D volume with thickness determined by the user's choice such as the thickness of the silicon thin film in SOI, falls entirely within a certain control volume, then the heat intensity for this volume including the grid is added by an amount of power dissipation (in units watts) for this block divided by the volume of this control volume.
2. If the block straddles among several control volumes associated with different grids, the block has to be partitioned into pieces depending on the intersection of the block and a particular volume (Fig 2). The heat from that piece of block is then assigned to that control volume, which will be in turn averaged out over the entire volume to obtain the heat generation intensity to be assigned to the grid. This procedure can be described in the following formula:

$$g_{ij} = P_j \frac{V_{ij}}{V_j} \frac{1}{V_i} \tag{3}$$

where g_{ij} is the heat generation intensity contributed to node i by block j, and V_i and V_j are volumes of the control volume associated with node i and block j, respectively; V_{ij} is the intersection between volumes i and j.

The above scheme guarantees the heat generation source would be precisely conserved no matter how coarse the mesh might be. Also the distribution of the heat source is preserved to the extent of the coarseness of the mesh. And the scheme is general independent upon the dimensionality and the shape of the mesh elements.

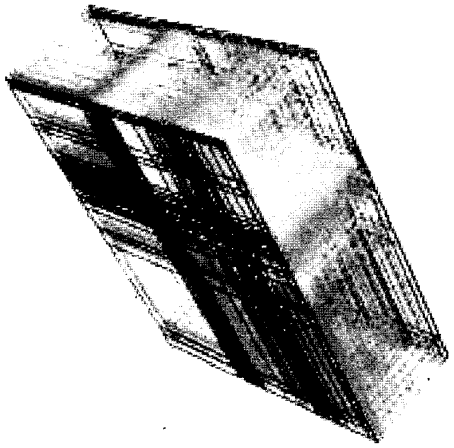

Fig. 3. 3D Mesh of simulation for a CPU chip using SOI technology

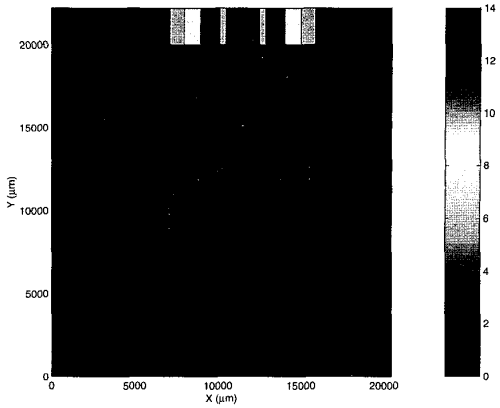

Fig. 5. Block placement and corresponding heat generation distribution (in units normalized for this application) for the bulk CMOS CPU chip.

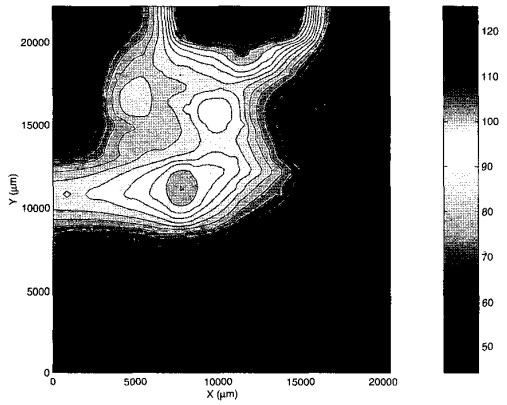

Fig. 4. Simulated temperature distribution on top of silicon substrate (bulk CMOS technology) with 75-by-75 mesh on the lateral dimensions. The power consumption for the entire chip is 100 W.

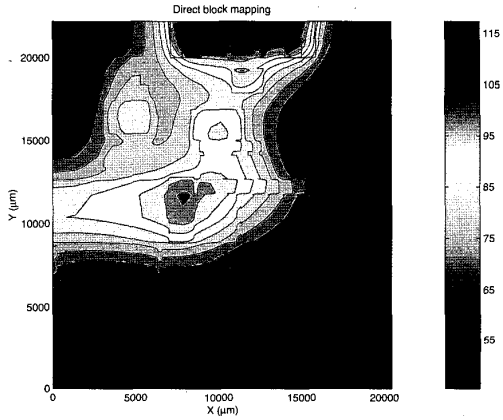

Fig. 6. Temperature distribution for the same chip as in Fig. 4 with conforming mesh to the block placement. Due to memory limitation, only 36 out of 115 blocks included and peak temperature is lower.

Results

Two CPU chip designs based on both bulk CMOS and SOI technologies have been investigated using 3D simulation capabilities of PROPHET (Fig. 3). Tensor product mesh is used throught the simulation region. The number of grids in the vertical direction is fixed to 42 and the mesh on the lateral plane varies. The temperature distribution on a 75-by-75 grid across the chip of dimension about 2×2 cm^2 is shown in Fig. 4. The chip is made of bulk CMOS and contains 115 functional blocks; the total power consumption is 100 W. It can be seen that the peak temperature is about 130 °C; detailed features (hot spots) are well captured in the simulation. To validate the accuracy, a denser and more precise mesh is used which conforms exactly with block size and location. Due to excessive numbers of grid required to conform to all functional blocks, only 32 blocks are included, but they constitute 85% of the total power consumption (heat distribution in Fig. 5). The results are shown in Fig 6. One can see the temperature distribution and peak temperature are similar, yet the computation time in the uniform mesh (100-by-100) is less than one half of that for the conforming mesh (Table 1). To further reduce the CPU time and to maintain reasonable accuracy, two more meshes are tested: one with 50-by-50 and another with 25-by-25 grid (Fig. 7), which demonstrates both acceptable accuracy and substantially less time to simulate.

Next we look at the peak temperature for different designs. Using the same coarse mesh, the second chip design with power consumption of 60 watts is simulated and the results are shown in Fig. 8. A noticeable feature of this particular design is that the peak temperature is located on the corner of the chip. From the thermal performance point of view, this type of design is less favorable in remov-

Mesh size	CPU time (s)
Conforming	10,456
100×100	4,233
50×50	403
25×25	68
25×25 (w/ loaded mesh)	27

TABLE I
CPU TIME (IN 440MHz HP-PA8500) COMPARISON FOR DIFFERENT MESH SIZE.

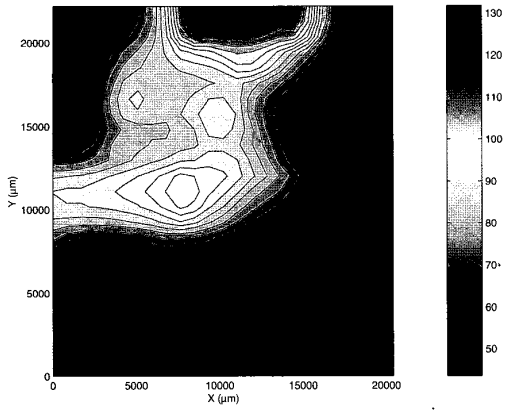

Fig. 7. Temperature distribution simulated for 25-by-25 mesh with other conditions the same as in Fig. 4.

ing the heat accumulated at the corners.

The Case of Silicon-28

Recently, serious efforts have been made in finding an alternative material with a better thermal conductivity than natural silicon (with three isotopes: ^{28}Si (92%), ^{29}Si (5%), ^{30}Si (3%)) used for CMOS IC substrate. It turns out that the pure silicon (^{28}Si), called silicon-28, has a thermal conductivity enhancement of almost 60% at the room temperature compared to natural silicon [4]. Simulation has been performed on the CPU chip with 100 W power consumption to assess the benefits of using silicon-28 in terms of thermal performance. In Fig. 9, the temperature distribution for the chip with silicon-28 substrate is shown. The peak temperature under the same power dissipation for the chip is reduced from 136 °C for natural silicon wafer (Fig. 7) to 123 °C, a reduction of almost 10%. Note that the temperature dependent thermal conductivity in silicon-28 decreases faster ($\alpha = 1.664$ in Eq. (2)) than that for natural silicon when the temperature is elevated.

Conclusion

A fast turn-around, large-scale thermal simulation approach using a device simulator has been developed to address the full chip temperature distribution. The impact of floorplanning design on the chip thermal performance has been studied. The simulated temperature profile and its spatial gradient can be used to further investigate the thermal impact on chip electrical performance such as delay and electromigration. It is further found through this work that using silicon-28 as the wafer, the peak temperature can be lowered as much as 10% as compared to that using natural silicon.

Acknowledgment: DARPA funding (DABT63-95-C-0090) to Stanford is greatly appreciated.

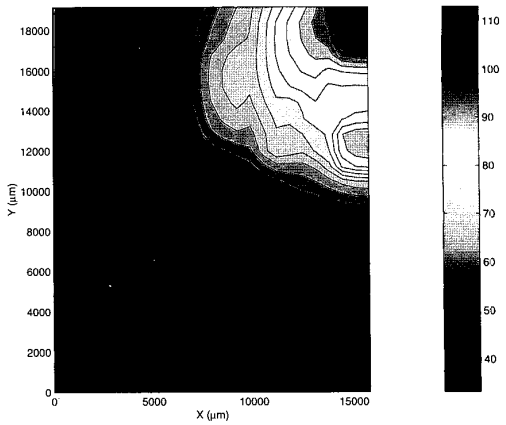

Fig. 8. Temperature distribution simulated for the SOI CPU chip design with lower power consumption than the first one. The peak temperature can be seen to locate on the corner of the chip.

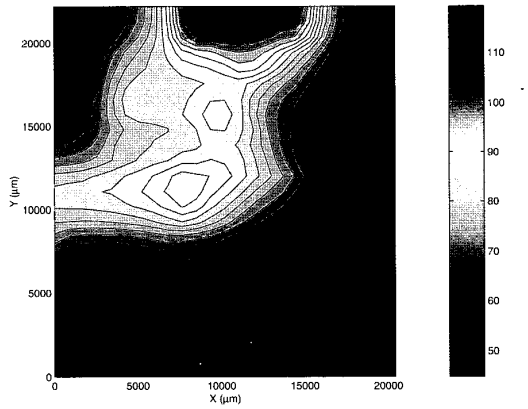

Fig. 9. Temperature distribution for silicon-28 wafer

REFERENCES

[1] C.S. Rafferty, Z. Yu, B. Biegel, M.G. Ancona, J. Bude, and R.W. Dutton, "Multi-dimensional quantum effect simulation using a density-gradient model and script-level programming techniques," *SISPAD* '98 p. 137, Lueven, Belgium, Sept. 1998.

[2] O. Tornblad, P.G. Sverdrup, D. Yergeau, K.E. Goodson, Z. Yu, and R.W. Dutton, "Modeling and simulation of phonon boundary scattering in PDE-based device simulators," *SISPAD* '00, p. 58, Seattle, WA, Sept. 2000.

[3] Z. Yu, *et al.*, PISCES-2ET manual, Stanford University, 1994.

[4] W.S. Capinski, *et al.*, "Thermal conductivity of isotopically enriched silicon," *Applied Physics Letters,* vol. 71, no. 15, p. 2109 (1997).