3D-Convolution Based Fast Transient Thermal Model for 3D Integrated Circuits: Methodology and Applications

Federica Lidia Teresa Maggioni^{1,2}, Herman Oprins², Dragomir Milojevic^{2,3}, Eric Beyne², Ingrid De Wolf^{2,4}, Martine Baelmans¹

¹ KULeuven, Department of Mechanical Engineering, Celestijnenlaan 300, 3001 Leuven, Belgium

² IMEC, Kapeldreef 75, 3001 Leuven, Belgium

³ BEAMS, Université libre de Bruxelles, Av FD Roosevelt 50, B1050 Brussels, Belgium

⁴ KULeuven, Department of Materials Engineering, Kasteelpark Arenberg 44, 3001 Leuven, Belgium
maggioni@imec.be

Abstract

A thorough thermal analysis of integrated circuits (ICs) is essential to prevent temperature driven reliability issues, which might cause the failure of microelectronic devices. The classical analysis approach is based on finite element methods (FEM). However, in the last decades, other computational methodologies have been developed with the aim to obtain results more quickly and at a reasonable accuracy. In this paper, a transient fast thermal model (TFTM) methodology for 3D-ICs based on 3D-convolution and fast Fourier transform is presented. This methodology allows to quickly and accurately predict the temporal evolution of the chip temperature distribution, due to power dissipation that can be non-uniform both in time and space, in all tiers of the 3D package.

In the first part of the paper the computational methodology is derived and described. Next, results are presented and validated with respect to conventional FEM simulations, showing good accuracy and computational time reduction. A realistic case, wherein different load switching scenarios are compared for a commercial floor-plan, is analyzed as an example of the applicability of the presented methodology. The speed of this algorithm, based on 3D-convolution, is compared with the one of previous work based on 2D-convolution and subsequent time superposition.

Keywords

Thermal analysis, fast thermal model, transient, 3D-convolution, Green's function, fast Fourier transform.

1. Introduction

Thermal analysis of ICs has always been crucial for the design of reliable chips and, with the development of new fabrication techniques, it became even more essential. New 3D die stacking approaches, in which active layers are placed on top of each other, aggravate thermal issues since the dissipated power over the same area significantly increases. Moreover, the dies are thinned down and they are vertically stacked using adhesives with low thermal conductivity [1].

Finite element methods (FEM) are commonly used for 3D die stack thermal analysis since they can provide accurate results. On the other hand, they are time consuming. In the early design phase, different designs and structures have to be tested and, therefore, faster and more user-friendly methods are preferred. This is mainly obtained by further simplifying the real structure and comes at the expense of some accuracy. A lot of research has been developed in the last years in the fast

thermal modeling area and different methodologies have been proposed [2-10].

Furthermore, real devices usually work in dynamic thermal regime, meaning that transistors are continuously switched on and off [2]. The work load is, indeed, distributed among the available cores, being some of them eligible to be kept free for a certain amount of time. This load distribution is really interesting from a thermal management point of view. The possibility to switch the load from one core, when its location reaches an excessive temperature, to another one positioned at a colder site, helps in controlling the maximum temperature within safe ranges.

The results obtained by steady state simulations, where the power is assumed to be dissipated constantly in time, represent, therefore, worst case scenarios. If the maximum allowed temperature and the amount of dissipated power are fixed for a specific device, steady state simulations could lead to opt for a more expensive cooling solution than actually needed for a real, dynamic thermal regime. These considerations, together with the need for high calculation speed, led to the development of transient fast thermal modeling (TFTM) methodologies.

TFTM methodologies have been widely studied in recent years and various strategies have been presented in literature. These approaches can be categorized in three main groups: *white-box* or physical approaches, *black-box* or behavioral approaches and *gray-box* approaches, which use combinations of physical and behavioral strategies [3].

White-box approaches are based on the physical equations governing the heat transfer phenomena and on approximation techniques. They allow, therefore, an easy explanation of the obtained results from a physical point of view but their ability to deal with unknown dynamics and relationships, as well as with complex geometries, is limited [2,4,5].

Black-box approaches, on the other hand, are based on the study of the thermal behavior of the analyzed system. They can be based on previously fully simulated data, obtained from already established modeling techniques as FEM, or on experimental data. On the one hand, these approaches are relatively robust against unknown relationships, they give a good behavioral description and they allow to consider complex geometries. On the other hand, they may lead to physical inconsistency due to over-fitting or unstable models as well as to results that can be difficult to explain from a physical point of view [3,6,7].

Gray-box approaches, as suggested by the name, place themselves in between the *white-* and the *black-box*

methodologies. They mainly combine previous knowledge and equations together with response data [8,9,10].

In this paper, a *gray-box* TFTM methodology is presented to compute the 3D-IC transient thermal response to time-varying power inputs. The method is based on superposition and Green's function theory applied through convolution between matrices storing data concerning the system thermal responses to hot spot heat sources, and matrices containing the dissipated power information. All these matrices are three dimensional (two spatial dimensions and time). 3D-convolution is, then, employed to compute the time dependent, highly resolved, 2D temperature profiles on all the active layers. The method can be classified as *gray-box* because it is based on a physical interpretation of the phenomenon (Green's functions and superposition) but the system thermal responses are obtained from simplified FEM simulations [8,11]. Moreover, thanks to the easiness and speed, with which results for different configurations can be obtained, this TFTM can be employed in the early design phase to quickly compare different technology options.

The paper is organized as follows: in Section 2 the simplifications and assumptions of the model are presented. The mathematical derivation and the methodology itself are also described here. In Section 3, the validation of the TFTM with respect to FEM is shown together with a realistic test case, proving the applicability of the methodology. The last part of this Section presents the computational time needed by the 3D-convolution algorithm and the improvement with respect to previous published work. Conclusions are drawn in Section 4.

2. Model Construction

2.1. Assumptions, simplifications and design parameters

One of the main characteristics of fast thermal models for 3D-ICs is their ability to quickly give an estimation of the temperature increase in the 3D stack. However, to achieve computational speed some simplifications need to be introduced.

The real 3D stack, with μ bumps, TSVs, BEOLs, package,... is simplified as a vertical stack of layers with homogeneous material properties. Equivalent homogenized material properties are used for heterogeneous layers (TSVs and Si, μ bumps and underfill,...). An algorithm to include the thermal impact of the real heterogeneity in the interface layer (underfill and μ bumps) in case of two-dies stacks and steady state regime has been proposed in [12]. Another simplification in the model concerns the packaging. The environment outside the stack itself is represented by equivalent boundary conditions. Insulation is assumed on the lateral vertical boundaries of the stack while equivalent convective boundary conditions with

Figure 1: Simplification procedure: from real structure to modeled structure.

Figure 2: Model parameters.

uniform heat transfer coefficients are applied on top and bottom of the stack. Moreover, since the active layers are much thinner than the dies and the interface layers, they are supposed to be planar. Figure 1 shows a sketch of the simplifications included in the model [11].

The developed TFTM should, moreover, be suitable to be used in the design phase, when the structure of the die stack and the floorplan of the different chips have not been fixed yet, to thermally compare different possible designs. This means that relevant design parameters, definable by the end user, are needed. These are the size of the dies, the thicknesses of each layer, the material properties, the heat transfer coefficients applied on top and bottom of the stack and the power dissipation position, intensity and time variation (Figure 2).

2.2. Mathematical derivation and Green's function

The general physical equation underlying the transient conduction heat transfer phenomenon is

$$c_p(\mathbf{x})\rho(\mathbf{x})\frac{\partial T(\mathbf{x},t)}{\partial t} - \nabla \cdot [k(\mathbf{x})\nabla T(\mathbf{x},t)] = q(\mathbf{x},t) \quad (1)$$

where the material specific heat capacity $c_p(\mathbf{x})$, the mass density $\rho(\mathbf{x})$ and the conductivity $k(\mathbf{x})$, which is assumed to be temperature independent, depend on the position $\mathbf{x} = (x, y, z)$, while the temperature $T(\mathbf{x}, t)$ and the power density $q(\mathbf{x}, t)$ generated by the heat sources depend on the position $\mathbf{x}$ and also on the time t .

The method that is used to handle the problem is Green's function based. A Green's function, $G(\xi; \xi_0)$, of a linear differential operator $L = L(\xi)$ acting on a function at a point ξ_0 , is any solution of $LG(\xi; \xi_0) = \delta(\xi - \xi_0)$ where $\delta(\xi - \xi_0)$ is the Dirac delta. For equation (1) and for a one layer, homogeneous material structure, the differential operator simplifies to $L = c_p\rho\frac{\partial}{\partial t} - k\Delta$. Under these conditions, the Green's function can be taken as a convolution operator and it can be calculated as the response of the system to a point heat source. Assuming $\xi = (x, y, z, t)$, the temperature response of the system can be calculated as

$$T(\xi) = \int G(\xi_0)q(\xi - \xi_0)d\xi_0 \quad (2)$$

which is a 4D (space and time) convolution operation between the Green's function and the power density.

The same reasoning applies for any layered structure as the one illustrated in Figure 1, provided that the temperature computation is restricted to horizontal planes and the G functions are computed accordingly. This means that, to

compute the temperature on a particular fixed z -level, 3D-convolution operations need to be solved (z is not a variable in the convolution).

2.3. Transient fast thermal model methodology

To calculate the temperature according to equation (2) via 3D-convolution, the information concerning the Green's functions and the dissipated power densities is needed. These are the two necessary ingredients for the TFTM. To distinguish the analytical quantities from the numerical ones, which are used in the model, the discretized versions of the analytical functions G and q are named hot spot responses (HSRs) and power maps (PMs), respectively.

- The *hot spot responses* ($^{\circ}\text{Cm}^2/\text{W}$) are 3D matrices storing the time dependent thermal responses of the system to hot spot power dissipations in the center of the active layers. The simulations are performed using 2D axisymmetric FEM models, whose horizontal dimension is much larger than the one of the real chip to exclude the impact of the lateral boundary conditions in these computations [11]. In the FEM, the power dissipation is localized in space but it is constant in time, after switching on at time 0. Numerical differentiation is performed afterwards to account for time impulsive power dissipation (Figure 3). This approach has been proved to have better numerical performances than the direct calculation of the impulse response by FEM. In the HSRs computation, each die is heated up separately and the temperature responses are recorded, at every time step, on all the levels of interest until steady state is reached. The *time step* is selected according to the device switching time. In case of N dies, N^2 3D HSRs are computed. The matrices are 3D since they include two horizontal spatial dimensions, whose size depends on the selected spatial resolution, and one time dimension.

Figure 3: (a) Hot spot (HS) localized in space; (b) hot spot localized in time and HSR as a function of time in the location of the HS dissipation; (c) HSR at fixed times as function of space.

- *Power maps* (W/m^2) are 3D matrices containing the time dependent information about the intensity and the location of the dissipated power. In case of N active layers, N 3D matrices are defined by the user. Each of these matrices contains k time layers, corresponding to the whole simulated time ($k = \text{simulated time}/\text{time step}$).

The HSRs are then convolved with the corresponding power maps resulting in N^2 3D-convolution operations. Due to the implementation of the convolution algorithm, the integration intervals are symmetric with respect to zero (power dissipation point) meaning that both $\text{HSR}(\cdot, \cdot, t)$ and $\text{HSR}(\cdot, \cdot, -t)$ are needed. Before power is dissipated (future time, negative values of the time variable), the temperature response is zero everywhere. For this reason, zero matrices have to be added to the HSRs to take into account the future and to allow for 3D-convolution. This means that, if m time steps are considered in the HSRs before steady state is reached, the 3D HSRs contain $2(m - 1) + 1$ time layers: one for the present, $m - 1$ for the past and $m - 1$ zero's layers for the future.

Furthermore, method of images and fast Fourier transform are exploited during 3D-convolution to, respectively, include the thermal impact of the insulating lateral boundaries and to speed up the computations [11]. Finally, results containing the temperature information referring to the same die and due to power dissipation on different layers are summed up, exploiting superposition principle, resulting in N temperature profiles. Each of them consists of k time layers that represent the thermal evolution of the system.

3. Results

3.1. Time dependent power map, TFTM validation

In order to validate the TFTM, a comparison with the results obtained by a commercial FEM software has been performed. The compared temperature profiles are obtained from the modeling of exactly the same structure: a die stack with homogeneous horizontal material layers (Figure 1). In the following, the absolute error, $|\text{err}| = |\text{TFTM} - \text{FEM}|$, between the TFTM and the FEM results is considered as the metric to compare the different modeling techniques.

The study of the thermal behavior of an $8.16 \times 8.16 \text{ mm}^2$ two-dies stack subjected to time varying power dissipation has been performed. The simulated time is 1.5 sec, the time step is 0.05 sec and the spatial resolution is $120 \times 120 \mu\text{m}^2$. The steady state is considered to be reached after 1 sec of constant power dissipation. Thus, $k = 30$ and $m = 20$.

Figure 4 summarizes the results for this case showing data concerning the maximum temperature increase achieved in the two dies as a function of time. The left column refers to the top die, while the right column to the bottom die. On the first row the comparison between the temperature increment computed by FEM (light blue line) and the TFTM (blue data points) is shown. The second row presents two graphs concerning the $|\text{err}|$, which results to be always smaller than $0.5 \text{ }^{\circ}\text{C}$. This shows that the accuracy of the model is really good. Moreover, the maximum error is located at the end of the process, after 1.2 sec. Considering that $m = 20$, this small error could be due to the selection of the time instant from which, in the HSRs calculation, steady state is assumed. Moreover, in this

Figure 4: Top: maximum temperature increase on top (left) and bottom (right) die computed by FEM (light blue line) and TFTM (blue data points). Bottom: $|err|$ on top (left) and bottom (right) die between the maximum temperature increase computed by FEM and by TFTM.

simulation no power is dissipated after 1.1 sec, meaning that the maximum error is located during the cooling down phase. It is, therefore, less important to have good accuracy here, since the thermally driven chip failures occur at high temperatures. During chip activities, which is the critical time, the error is less than 0.25 °C.

Together with accuracy, the other important characteristic of a compact thermal model is its computational speed. In case of FEM simulations the computational time depends on the complexity, both in time and space, of the power maps. Transient FEM can, indeed, be sped up using variable time stepping. In this way the time step is automatically adapted according to the temporal variations of the power dissipation. In particular, to accurately model fast variation in the temperature responses, smaller time steps are selected. This means that, the more the power maps vary, the higher the temperature gradients are, the smaller time steps are selected and the higher computational time is required.

In case of the TFTM, instead, the computational time is not affected by the power maps complexity but just by the number of elements in the convolved matrices. This is because the fast temperature changes have already been considered during the calculation of the HSRs. They are, therefore, already included in the HSRs and they don't need to be followed again for the specific power maps. The size of the fixed time step in the TFTM can be selected according to the shortest interval the power dissipation is kept constant in time. The improvement in computational time from FEM to TFTM simulations is, thus, greater for more complex power maps. For this simulated case, for instance, the TFTM results are obtained more than 50 times faster than the corresponding FEM ones. For power dissipations with short variations followed by long intervals with constant power maps, the improvement in computational time is less significant.

Figure 5: 3D repartitioning of the OpenSparc floorplan. The dies are 5x5 mm² in size and the spatial resolution is 100x100µm².

3.2. Test case - OpenSparc

The same developed TFTM methodology has been applied to model a different situation. This second test case, whose main aim is to show a possible application of the method, concerns the transient thermal analysis of a 3D repartitioning of the open source OpenSparc T2 floor plan [14] into two smaller stacked dies (Figure 5). To include the package thermal capacitance effect, in the computation of the HSRs, a layer of material is added on top and bottom of the stack. The cooling is mainly provided from the top side of the stack. Three different scenarios have been analyzed:

- *Scenario 1:* both the dies are 100% on for the whole simulated time;
- *Scenario 2:* the top die is always on while, for the bottom die, the load is switched from/to the four central cores (2, 3, 4 and 5) to/from the four external one (0, 1, 6 and 7) every 50 msec. When a core is on, it dissipates 100% of the power (as represented in Figure 5), while, when it is off, it dissipates 0 W;
- *Scenario 3:* this scenario is similar to Scenario 2 but the switching of the loads among the cores happens every 250 msec instead of every 50 msec.

Figure 6 shows the time evolution of the temperature increase in the places indicated, in Figure 5, with a blue cross (labeled by “center” and represented by darker colors and dash lines in Figure 6) and with a green cross (labeled by “lateral” and represented by lighter colors and full line in Figure 6). The temperature values referring to the central location are slightly higher than the lateral ones because the former point is located in the middle of two active cores while the latter one is in the middle of one active core (the other one is on the opposite corner).

The curves referring to Scenario 1 (circular markers) show an increasing temperature value until steady state is reached. This is, of course, because the power maps are static. The other curves, instead, show a periodic behavior, the only difference being in the period: for Scenario 2 (star-shape markers) it is 50 msec while, for Scenario 3 (square markers), 250 msec. It is important to note that the TFTM is able to detect the thermal effect of these changes in power dissipation. It can, therefore, be exploited in similar cases when the load can be switched between different cores and an appropriate switching time needs to be calculated.

Figure 6: Temperature evolution as a function of time in the points indicated with a blue (label “center”) and green (label “lateral”) cross in Figure 5. Top die values on the top and bottom die values on the bottom of the graph.

3.3. Computational time

In previous work [8], a similar approach based on Green’s function has been presented. The inference of the TFTM methodology from the analytical equations has been, however, a bit different. The basic equation (2) can, indeed, be rewritten as

$$T(\xi) = \int G(\xi_0)q(\xi - \xi_0)d\xi_0 = \int (\int G(x_0, t_0)q(x - x_0, t - t_0)dx_0)dt_0 \approx \sum \bar{T}(x, t; t_{0,i})\Delta t_{0,i}. \quad (3)$$

The inner integral in the third term, $\bar{T}(x, t; t_0)$, which is the temperature at time t due to power dissipated at time $t - t_0$, is computed by spatial convolution (2D-convolution once restriction to planes is applied). The external time integral is, then, solved via numerical integration resulting in the right hand side of the equation ($\Delta t_{0,i}$ is the time step).

This approach is, therefore, based on 2D-convolution in space combined with superposition in time and the extra zero’s matrices representing the future time don’t have to be added.

Figure 7: Comparison between the computational time needed for 3D-convolution and for 2D-convolution with subsequent time superposition.

However, the 3D-convolution methodology, which we present in this paper, proved to be computationally faster when the number of considered time layers k is large enough. The exact number depends on the spatial resolution but it is relatively low (around 6 and 8). In most realistic situations, larger simulation intervals are considered for which, therefore, 3D-convolution is faster. The overall speed-up depends on the dimensions of the HSRs and PMs matrices.

Figure 7 shows a comparison between the computational time needed by the 3D-convolution algorithm and by the 2D-convolution plus subsequent time superposition one. The *tic-toc* Matlab command has been used to obtain the computational time. The data refer to a two-dies stack in which each active layer is discretized in a grid of 50 x 50 cells. A 50 msec time step is considered and the comparison has been performed for different simulated times ($2 \leq k \leq 30$). The time length m of the HSRs is always equal to the number k of time layers in the PMs. As we can see from the graph, for small k values the performances of the two algorithms are comparable but, for larger simulated times, the 3D-convolution approach outperforms the 2D one. A five time improvement is reached, for example, considering 30 time layers, which correspond, in this case, to 1.5 sec of simulated chip activity. It is worth noting that the computational time doesn’t grow linearly or quadratic with the number of time layers, k . This is related to the convolution algorithm whose execution time depends on the prime factors of the involved matrices dimensions [15].

4. Conclusions

A methodology for the fast calculation of the transient thermal response of 3D-ICs based on Green’s function theory has been developed. The selected methodology provides temperature profiles with high spatial resolution. The time resolution is normally chosen as the shortest device switching time (50 msec is used as an illustration in this paper, but this is not a limitation). Compared to previous work on the same topic, the implementation of 3D-convolution, instead of 2D-convolution and subsequent time superposition, results in lower computational time in most practical situations.

Concerning the computational time, it is much less than in case of FEM models of the same structure (stack of homogeneous layers) and, opposite to the FEM, it is independent of the temporal and spatial complexity of the power maps. For the test case shown in this document, the computational time of the TFTM is more than 50 times lower than the one of the FEM model with a maximum absolute error within 0.5 °C (out of 60 °C temperature increase). It has also been demonstrated that the model is capable to predict the thermal impact of different switching scenarios and, therefore, it can be used to select an appropriate switching time to move the load among different cores, which can help in avoiding overheating.

References

1. Szabo, P., Poppe, A., Farkas, G., Szekely, V., Courtois, B., & Rencz, M. (2006, May). "Thermal characterization and compact modeling of stacked die packages". In *Thermal and Thermomechanical Phenomena in Electronics Systems, 2006. ITherm'06. The Tenth Intersociety Conference on* (pp. 251-257). IEEE.
2. Christiaens, F., Vandeveld, B., Beyne, E., Mertens, R., & Berghmans, J. (1998). "A generic methodology for deriving compact dynamic thermal models, applied to the PSGA package". *Components, Packaging, and Manufacturing Technology, Part A, IEEE Transactions on*, 21(4), 565-576.
3. Liu, Z., Tan, S. X. D., Wang, H., Hua, Y., & Gupta, A. (2014). "Compact thermal modeling for packaged microprocessor design with practical power maps". *Integration, the VLSI Journal*, 47(1), 71-85.
4. Sridhar, A., Vincenzi, A., Ruggiero, M., Brunschweiler, T., & Atienza, D. (2010, November). "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling". In *Proceedings of the International Conference on Computer-Aided Design* (pp. 463-470). IEEE Press.
5. Gerstenmaier, Y. C., & Wachutka, G. K. (2005). "Transient temperature fields with general nonlinear boundary conditions in electronic systems". *Components and Packaging Technologies, IEEE Transactions on*, 28(1), 23-33.
6. Barabadi, B., Joshi, Y. K., & Kumar, S. (2012, December). "Rapid multi-scale transient thermal modeling of packaged microprocessors using hybrid approach". In *Electronics Packaging Technology Conference (EPTC), 2012 IEEE 14th* (pp. 157-164). IEEE.
7. Schweitzer, D. (2009). "A fast algorithm for thermal transient multisource simulation using cubic spline interpolation of Zth functions". *Components and Packaging Technologies, IEEE Transactions on*, 32(2), 478-483.
8. Park, J. H., Shakouri, A., & Kang, S. (2008, March). "Fast evaluation method for transient hot spots in VLSI ICs in packages". In *Quality Electronic Design, 2008. ISQED 2008. 9th International Symposium on* (pp. 600-603). IEEE.
9. Gerstenmaier, Y. C., & Wachutka, G. K. (2004). "Efficient calculation of transient temperature fields responding to fast changing heat sources over long duration in power electronic systems". *Components and Packaging Technologies, IEEE Transactions on*, 27(1), 104-111.
10. Beneventi, F., Bartolini, A., Tilli, A., & Benini, L. (2013). "An effective gray-box identification procedure for multicore thermal modelling".
11. Maggioni, F. L. T., Oprins, H., Beyne, E., De Wolf, I., & Baelmans, M. (2014). "Fast convolution based thermal model for 3D-ICs: Methodology, accuracy analysis and package impact". *Microelectronics Journal*, 45(12), 1746-1752.
12. Maggioni, F.; Oprins, H.; Beyne, E.; De Wolf, I. and Baelmans, M., "Convolution based steady state compact thermal model for 3D integrated circuits: methodology for including the thermal impact of die to die interconnections", *Proceedings of the 15th International Heat Transfer Conference, IHTC-15. Kyoto*, 10-15 August 2014.
13. Maggioni, F. L. T., Oprins, H., Beyne, E., De Wolf, I., & Baelmans, T. (2013, September). "Convolution based compact thermal model for 3D-ICs: methodology and accuracy analysis". In *Thermal Investigations of ICs and Systems (THERMINIC)*, 2013 19th International Workshop on (pp. 152-157). IEEE.
14. OpenSparc T2, URL: <http://www.oracle.com/technetwork/systems/opensparc/opensparc-t2-page-1446157.html>
15. Matlab, TheMathworks, URL: <http://nl.mathworks.com/help/matlab/ref/fft.html>