Fast Transient Convolution Based Thermal Modeling Methodology for Including the Package Thermal Impact in 3D-ICs

Federica L. T. Maggioni, Herman Oprins, Eric Beyne, Ingrid De Wolf, and Martine Baelmans

***Abstract*—The relevance of accurate prediction of the thermal behavior of microelectronic systems has been increasing since the introduction of 3D-ICs. Different modeling strategies have been implemented to this scope, aiming both to increase accuracy and to reduce computational time. In this paper, a transient fast thermal model methodology for packaged 3D stacked ICs is presented. It can be considered as a multi-scale strategy whose core is constituted by a highly resolved, convolution based algorithm. This allows to compute the temperature increase due to a generic, time varying, power map in a stack configuration. On top of this, the time dependent package thermal spreading and capacitive effect is included via correction profiles. These corrections are based on the ratio between the thermal responses of the package and of the stack configurations to uniform, impulsive, power dissipation at different time steps. Validation with respect to finite element method results shows good accuracy. An error metric, to estimate a priori the need to include the package impact on top of the convolution based approach, has also been developed. Alternative but similar algorithms, which place themselves in between the solutions with and without the package impact, both from an accuracy and from a computational time point of view, are also shortly presented in this work.**

***Index Terms*— Convolution, Electronic packaging thermal management, Fast thermal model.**

I. Introduction

HIGH temperatures and high spatial and/or temporal temperature gradients represent a significant issue for the design and the fabrication of performant and reliable integrated circuits (ICs) [1]. Thermal issues are further exacerbated in 3D systems where active components are stacked on top of each other. These issues are not only due to the increased power density dissipated over the same area available for cooling, but also to the use of adhesives with low thermal conductivity for the vertical integration of the electronic components and to the reduced lateral spreading in the thinned Si chips.

Accurate thermal analysis of 3D-ICs is, therefore, crucial to ensure the reliability of the components. The numerical thermal investigation is commonly performed by finite element modeling (FEM). Although these simulations provide accurate results, they are mostly time consuming. Since computational time is a key factor in certain situations, different research groups started working on the development of compact thermal models (CTMs) or computationally fast thermal models (FTMs). By further simplifications, these methods allow to obtain relevant thermal information more quickly and more easily. Different FTM approaches have been proposed in the last decades, both for the steady state and for the transient regime. These approaches can be categorized in three main groups [2]: white, black and gray box approaches depending on whether they are physical approaches (based on RC-networks [4] or on the heat transfer equation [3]), behavioral approaches (based on experiments and/or previously simulated data [1],[6]) or a combination of them [7]-[9].

The steady state analysis is much simpler and faster than the transient one but it represents a worst case scenario. Actual devices work, indeed, in a dynamic thermal regime, with subsequently on-off switching of the cores. Considering steady state results might lead to opt for a more advanced and expensive cooling solution than is actually needed in real working conditions [10].

An important difference between the various FTMs proposed in literature concerns the level of interest of the simulation (stack, package, system,...). This choice determines which parts of the system are effectively modeled and which parts are considered as outside environment. The impact of these last parts is included by applying specific boundary conditions (BCs), which are typically insulating or convective. Although, they can provide a satisfactory estimate of the environmental thermal impact in steady state conditions, in transient simulations they cannot deal with the capacitive effect of the un-modeled parts. This might have a significant impact on the

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456".

F. L. T. Maggioni is with the Department of Mechanical Engineering, KULeuven, Celestijnenlaan 300, 3001 Leuven, Belgium and with IMEC, Kapeldreef 75, 3001 Leuven, Belgium (e-mail: maggioni@imec.be).

H. Oprins is with IMEC, Kapeldreef 75, 3001 Leuven, Belgium (e-mail: oprins@imec.be).

E. Beyne is with IMEC, Kapeldreef 75, 3001 Leuven, Belgium (e-mail: beyne@imec.be).

I. De Wolf is with the , Department of Materials Engineering, KULeuven, Kasteelpark Arenberg 44, 3001 Leuven, Belgium and with IMEC, Kapeldreef 75, 3001 Leuven, Belgium (e-mail: dewolf@imec.be).

M. Baelmans is with the Department of Mechanical Engineering, KULeuven, Celestijnenlaan 300, 3001 Leuven, Belgium (e-mail: martine.baelmans@kuleuven.be).

calculated thermal time constant of the system.

In this paper, a gray-box FTM algorithm for package level transient simulations is presented. It allows to compute highly resolved temperature profiles on the active layers of 3D-ICs using convolution and superposition. In our previous works [10],[11], where the basic methodology has been described, only the die stack was modeled. The package and the surrounding environment were considered by means of equivalent uniform convective BCs. However, since the area of the package materials is typically larger than the area of the die stack, this simplification introduces errors due to neglecting the thermal spreading. In [12], Hériz et al. proposed a solution for steady state simulations. In this paper the methodology is extended to the transient regime, where also the capacitive effect of the package plays an important role. In [13], Pan et al. deal with a convolution based algorithm that includes the package thermal impact in the transient regime. However, in their approach, the package thermal responses need to be computed for each specific dissipated power map since they only perform time-convolution and not space-convolution. This work goes, therefore, a step further towards generalization with both spatial and temporal convolution. The developed algorithm is applied to two package configurations, which are sketched in the first row of Fig. 1. The same figure also reports the illustration of the nomenclature used throughout the paper.

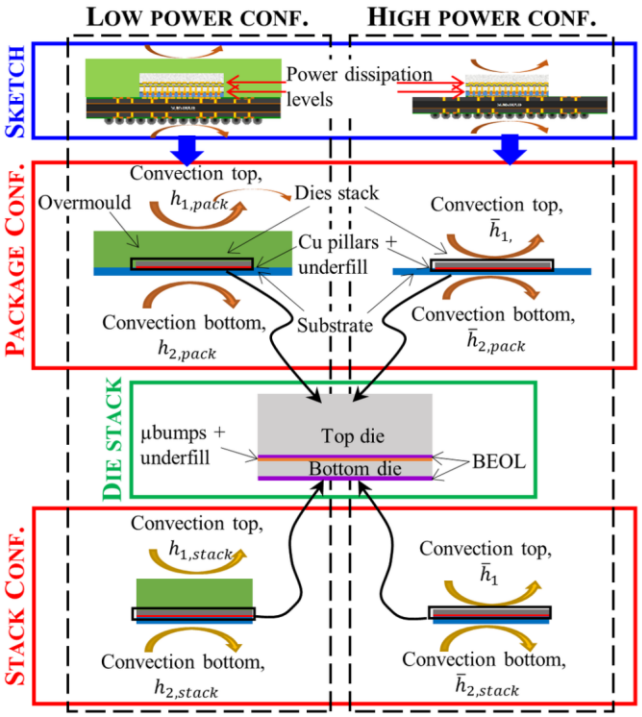

Fig. 1. Nomenclature and package configurations presented in the paper.

II. Basic Methodology

According to the Green's function theory, the temperature profile $T(\boldsymbol{\xi})$, which is the solution of the conduction equation governing the heat transfer phenomenon,

$$c_p \rho \frac{\partial T(\xi)}{\partial t} - \nabla \cdot [k \nabla T(\boldsymbol{\xi})] = q(\boldsymbol{\xi}), \tag{1}$$

can be, under certain conditions, computed as

$$T(\boldsymbol{\xi}) = G(\boldsymbol{\xi}) * q(\boldsymbol{\xi}) \tag{2}$$

where $q(\boldsymbol{\xi})$ represents the dissipated power density, $G(\boldsymbol{\xi})$ is the response of the system to a Dirac delta function and * is the convolution operator. $\boldsymbol{\xi}$ represents any variable, spatial and temporal, that is considered. The specific heat capacity, c_p, the mass density, ρ, and thermal conductivity, k, of each material can, in general, be temperature and position dependent.

For the Green's function theory to be valid, certain conditions need to be met. Firstly, the differential operator has to be *linear*. This means, in the analyzed case, that the material properties are temperature independent. Secondly, it has to be *translation invariant*. In other words, the system response $G(\boldsymbol{\xi})$ is independent of the location where the Dirac delta power is applied. This only holds if a configuration of infinitely large, homogeneous, stacked layers is considered and if the $G(\boldsymbol{\xi})$ functions and the resulting temperature profiles $T(\boldsymbol{\xi})$ are restricted to horizontal planes. Assuming the same footprint area for all the active layers and insulating lateral BCs, the method of images can be used to relax the requirement of infinitely large layers [11]-[12]. Moreover, to satisfy the position independence requirement, BCs with constant coefficients need to be applied on top and bottom of the stack.

The *basic* FTM methodology has been developed under these assumptions. Two ingredients provide the fundamentals for this method: the *power maps (PMs)* and the *hot spot responses (HSRs)*. The PMs are matrices storing the information about the dissipated power in a user defined resolution, while the HSRs are matrices storing, in the same resolution, the thermal responses of the system to localized heat dissipations. The latter are computed by 2D, axisymmetric FEM simulations in which hot spot power sources are subsequently dissipated on all the active layers and the temperature responses are, every time, calculated on each of them. In the steady state regime, 2D-convolution operations are performed between corresponding PMs and HSRs. The resulting temperature profiles referring to the same layer are then superposed to obtain the final response of the 3D-ICs to the applied PMs [10]. In the transient regime, also the temporal variation of PMs and HSRs needs to be considered. To this aim, the HSRs data, obtained for impulsive (i.e. one time step long) heat dissipation, are stored until steady state is reached and 3D-convolution (one temporal and two spatial variables) is used to compute the resulting time dependent temperature profiles [10].

One of the main drawbacks of this approach is that all the layers need to have the same horizontal surface. In real situations, however, the die stack is enclosed in a package, whose components have a larger area. This characteristic allows for lateral heat spreading and helps in reducing temperature. The inclusion of this spreading/package effect in the simulations can have a significant impact on the temperature profiles [11]. Hériz et al. proposed in [12], for the steady state regime, an error reduction technique to include the package impact in the convolution based methodology. The *intrinsic*

error between the stack and the package configurations (Fig. 1) is computed by comparing the system responses to uniform power dissipation in the two cases. The temperature profile for the stack geometry, $T_{stack,unif}$, is computed by the basic FTM while the one related to the package geometry, $T_{pack,unif}$, by means of FEM. The general temperature profile, $T_{FTM,basic}$, obtained with the basic FTM for non-uniform power dissipation in the stack configuration, is corrected as

$$T = \frac{T_{FTM,basic}}{1+E_r} + T_{amb}, \quad E_r = \frac{T_{stack,unif} - T_{pack,unif}}{T_{pack,unif}}. \tag{3}$$

All the temperature values are computed assuming zero ambient temperature and represent temperature increases. Rearranging of equation (3) leads to

$$T = T_{FTM,basic} \cdot C + T_{amb}, \quad C = \frac{T_{pack,unif}}{T_{stack,unif}} \tag{4}$$

where C is a position dependent *correction factor* carrying the information over the thermal spreading that is intrinsically neglected in the basic FTM for the stack configuration.

Both this package correction methodology and the one presented hereafter for the transient regime deal with the limitation of the convolution based FTM concerning the *translation invariance* of the differential operator. The system is, however, still considered to behave *linearly* and both the HSRs and the package corrections are computed assuming constant material properties.

III. Package Correction in Transient Regime

As presented in [10], the time dependent temperature profiles in transient regime for the stack configuration can be computed either by 3D-convolution or by 2D-convolution with subsequent time superposition. The related formulas are

$$T_{stack}(\boldsymbol{\xi}) = \int G(\boldsymbol{\xi_0}) q(\boldsymbol{\xi} - \boldsymbol{\xi_0}) d\boldsymbol{\xi_0} = G(\boldsymbol{\xi}) *_{3D} q(\boldsymbol{\xi}) \tag{5}$$

$$T_{stack}(\boldsymbol{x}; t) = \int [G(\boldsymbol{x}, t_0) *_{2D} q(\boldsymbol{x}, t - t_0)] dt_0 = \int \bar{T}(t_0; \boldsymbol{x}, t)\, dt_0 \tag{6}$$

where $\boldsymbol{\xi} = (x, y, t) = (\boldsymbol{x}, t)$ is the space-time variable, $*_{3D}$ and $*_{2D}$ are, respectively, the 3D- and 2D-convolution operators and $\bar{T}(t_0; \boldsymbol{x}, t)$ is the impulsive *partial* temperature increase at time t. It is *partial* because it represents just the temperature increase due to power dissipated at $t - t_0$. The main difference between the 3D and the 2D approach is that, in equation (5), the solution is computed, at the same time and by means of one single operation, for all the possible values of both the spatial and the temporal variables. When the approach in equation (6) is selected, on the other hand, the results refer to a fixed point in time t. Moreover, even to obtain $T(\boldsymbol{x}; t)$ at fixed time t, multiple 2D-convolutions operations are needed, one for each past power dissipation time, $t - t_0$. As a consequence, since the convolution operations can be sped up by the application of the fast Fourier transform, the approach in equation (5) is computationally much faster than the one in equation (6), when the number of considered time steps is large enough [10].

Considering the formulas from a numerical point of view, the temperature increase, for a stack configuration, on the vertical level j due to power dissipation on level i, can be computed as

$$T_{stack,ij}(\boldsymbol{\xi}) = HSR_{ij}(\boldsymbol{\xi}) *_{3D} PM_i(\boldsymbol{\xi}) \tag{7}$$

$$T_{stack,ij}(\boldsymbol{x}; t) = \sum_k \left[HSR_{ij}(\boldsymbol{x}, t_k) *_{2D} PM_i(\boldsymbol{x}, t - t_k)\right] \Delta t_k = \sum_k \bar{T}_{ij}(t_k; \boldsymbol{x}, t) \Delta t_k \tag{8}$$

where Δt_k is the time step.

In order to apply the package correction procedure in the transient regime, the temporal sequence of the PMs has to be considered. If the same power p' is dissipated at time $t - t_0$ rather than at $t - t_1$, with $t_0 \neq t_1$, the system response at time t is different. For this reason, each impulsive partial temperature increase profile $\bar{T}_{ij}(t_k; \boldsymbol{x}, t)$ needs to be corrected individually, depending on the value of $t - t_k$. In this way, the different time constants of the different parts of the package are taken into account. This is why the transient FTM with package correction is implemented via 2D-convolution and time superposition. The 3D-convolution algorithm does not allow the direct access to the partial temperature increase profiles.

Since the HSRs are computed for impulsive heat dissipation and, therefore, $\bar{T}_{ij}(t_k; \boldsymbol{x}, t)$ are the partial temperature profiles at time t due to impulsive power dissipation at time $t - t_k$, the transient correction profiles need to be computed accordingly. Analogously to the steady state algorithm, they are defined as

$$C_k = \frac{T^k_{pack,unif}}{T^k_{stack,unif}}. \tag{9}$$

Since, in the discrete domain, impulsive means one time step long, $T^k_{pack,unif}$ and $T^k_{stack,unif}$ are the temperature profiles at time step k obtained for uniform, impulsive power dissipation in the time interval $[0, \Delta t)$ for, respectively, the package and the stack configuration. The temperature increase is computed as

$$T_{ij}(\boldsymbol{x}; t) = \sum_k \left[HSR_{ij}(\boldsymbol{x}, t_k) *_{2D} PM_i(\boldsymbol{x}, t - t_k)\right] C_k \Delta t_k = \sum_k \bar{T}_{ij}(t_k; \boldsymbol{x}, t) C_k \Delta t_k. \tag{10}$$

In Fig. 2, the half-diagonals of the correction profiles C_k are plotted for the two package configurations considered in this paper. In both cases, the die stack is attached to a package substrate by a layer of Cu pillars and underfill, for which equivalent in-plane and out-of-plane material properties are used. In the low power (LP) configuration the die stack is overmolded while, in case of the high power (HP) configuration, the chip backside is exposed and the cooling is directly applied on top of it (Fig. 1). It is important to stress that these are just two possible examples of package configurations, chosen to show the capabilities of the model. The considered HP configuration is, in particular, a bare die configuration and not a typical HP one, which normally has a lid that acts as heat spreader. The methodology is, however, more general and can be applied to different scenarios. In Fig. 2, the time dependency

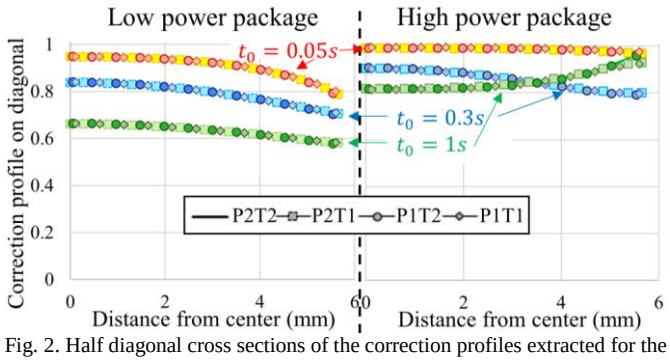

Fig. 2. Half diagonal cross sections of the correction profiles extracted for the LP (left) and the HP (right) package configurations at different times: at the end of the HS dissipation (0.05 s, red), at 0.3 s (blue) and at 1s (green). Different marker's types indicate different dissipation and temperature response levels in which the correction profiles are computed. In the legend PxTy stands for power dissipated on die x and temperature computed on die y; 1 indicates top die and 2 bottom die.

of C_k is clearly visible. To be noted from the same plot is the lack of normalization of the profiles. This is because the BCs applied to the stack configuration have been calculated in such a way that $T_{stack,unif} \approx T_{pack,unif}$ at steady state. However, while at steady state the application of the correction profiles compensates just for the difference in thermal resistance due to lateral spreading in the package, in the transient regime the capacitive capability of the different package materials needs also to be taken into account. Since, in the stack configuration, parts of the package are neglected, the correction profiles need to account for their missing capacitive effect. This is why the maximum of C_k is time dependent.

IV. Temperature Profiles for Uniform Power Dissipation and Package Configuration

Since $T^k_{pack,unif}$ are obtained by FEM, their computation could be computationally expensive and could represent a bottleneck for this methodology. Different approaches have been tested, at steady state, to obtain $T_{pack,unif}$ avoiding FEM. They were based on the conformal mapping strategy [14] and on analytical solutions [15] but none of them, even without considering time dependency, revealed to be able to deliver acceptable results. For this reason, possible simplifications of the FEM have been analyzed to allow for computational time reduction. This comes, of course, at the expense of the details included in the model. However, these temperature profiles are used as correction factors to be applied on top of the basic FTM results for specific PMs. This approach can, therefore, be seen as a multi-scale strategy in which the correction is located on the lower level of accuracy. For this reason a coarse mesh, together with a quarter symmetry, is used.

A. Uniform Power Dissipation for the Stack Configuration

In case of uniform power dissipation on the horizontal layer i of a stack configuration, the system temperature response on level j, $T^k_{stack,unif,ij}$, is constant in space but variable in time. Its thermal evolution can, therefore, be described with *one single value* per time step. The simple direct application of the convolution based FTM to compute $T^k_{stack,unif,ij}$ would, however, result in *highly resolved uniform* temperature fields.

For this reason, a methodology similar to the one originally presented in [11] for *steady state* is proposed. This approach is basically a simplification of the 2D-convolution, valid in case one of the two matrices (the PM in this case) is uniform. Under this circumstance, indeed, each value of the matrix resulting from a convolution, can be computed as $T = \sum_{\bar{l},\bar{m}} PM \cdot HSR_{\bar{l}\bar{m}}$, where $\bar{l}, \bar{m}$ are row and column indices. Moreover, since the HSR has circular symmetry, the calculations can be further simplified considering a 1D-HSR vector, whose elements are function of the distance from the hot spot (HS). Doing so, care should be taken on how many terms in the sum refer to the same value in the 1D-HSR vector. In other words, we need to know how much *area* of the original PM refers to each single 1D-HSR value. Exploiting the circular symmetry property and multiplying each term in the sum by the corresponding circular annulus area, the following equation is obtained:

$$T = \sum_{\bar{l}} PM \cdot [1D\text{-}HSR]_{\bar{l}} \cdot \pi(a_{\bar{l}+1} - a_{\bar{l}})^2 \qquad (11)$$

where $a_{\bar{l}}$ is the middle point between the locations to which $[1D\text{-}HSR]_{\bar{l}-1}$ and $[1D\text{-}HSR]_{\bar{l}}$ refer.

For the transient regime, exactly the same procedure can be applied, considering each time step separately. The only difference is that the result for each time step is in units of ºC/s and, therefore, it has to be multiplied by the corresponding time step (cf. eq. (8)).

B. Levels of the Correction Profiles

In accordance with the FTM methodology, the correction profiles should depend on the levels in which power is dissipated and on which temperature is computed. In case of stacks with N dies, for example, N^2 correction profiles should be calculated, one for each [power dissipation level – temperature computation level] combination. However, as shown in Fig. 2, the difference between the correction profiles calculated on different levels is not significant neither in the LP (left) nor in the HP (right) configuration. For this reason, in the following, a single correction profile is used, reducing, in this way, the number of required FEM simulations.

C. Equivalent Material Properties

Another simplification that can be implemented while computing $T^k_{pack,unif}$, consists in neglecting the layered structure of the die stack (Si, µbumps and BEOL). This information is, indeed, already included in the stack configuration and, therefore, in the uncorrected results obtained by the basic FTM. For this reason, and also because the dependence of the correction profiles on the power dissipation layer is negligible, the die stack can be assumed to be of one uniform material while computing $T^k_{pack,unif}$ by FEM. This uniform material block should, however, mimic the thermal behavior of the original layered die stack. Appropriate equivalent material properties are, therefore, assigned to it.

The equivalent orthotropic thermal conductivity is computed, at steady state, by means of equivalent resistance networks for the stack configuration. The *complete* vertical heat

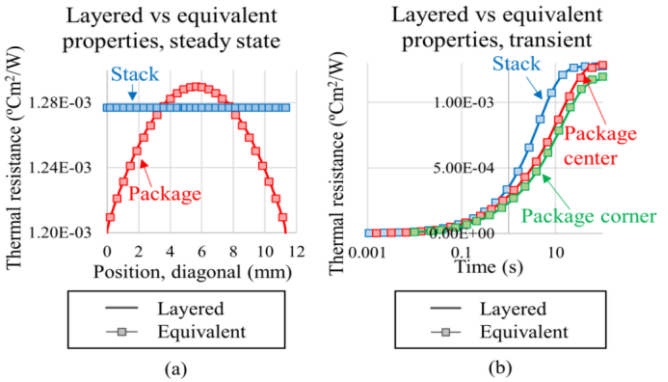

Fig. 3. Normalized temperature profiles for uniform power dissipation obtained by using equivalent properties (markers) and by using the more realistic layered structure for the die stack (full line). (a): Diagonal cross sections in steady state regime for package (red) and stack (blue) configuration. (b): Logarithmic time scale temperature evolution for the stack (blue) and the package configuration, in the center (red) and in the corner (green) of the die.

path is considered for computing k_z, while a horizontal heat path through the die stack is used to calculate the equivalent value of $k_{x,y}$. On the left hand side of Fig. 3, the really good agreements between the normalized temperature profiles, on the diagonal of the die, obtained by using equivalent properties and by using the more realistic layered structure for the die stack, are shown for the stack and the package configuration.

For simulations in the transient regime, the equivalent capacitance value is also needed. Since the thermal capacitance is defined as a volumetric integral, its equivalent value is computed by means of volume average. The right hand side of Fig. 3 shows the comparison between the normalized transient thermal responses of a system modeled using equivalent properties and using the more realistic layered structures. The two graphs in Fig. 3 proved the possibility to substitute, in the package FEM simulations, the different layers in the die stack with a single material block to which equivalent properties are assigned. Computational time is, consequently, reduced.

D. Temperature Profile Extraction

Another step in the algorithm, in which computational time can be saved, is the extraction of the temperature profiles from the packaged FEM model. The correction procedure consists in point-by-point multiplications between the temperatures obtained by the basic FTM, $T_{FTM,basic}$, and the correction profiles. For this reason $T^k_{pack,unif}$, which are extracted from the FEM coarse model, need to have the same high resolution as $T_{FTM,basic}$. The common way to achieve this aim, is through space interpolation of $T^k_{pack,unif}$ at each individual time step.

Another possibility is to take into account what each single correction profile is applied to. The temperature increase at each time step is, indeed, computed in the FTM as the sum of the temperature profiles due to impulsive power dissipations in the past. The further ago an impulse has been dissipated, the less its contribution is on the final temperature increase. This means that the accuracy, with which $T^k_{pack,unif}$ are extracted, can be lowered according to how long ago the generating impulsive power has been dissipated.

After several tests on different package structures, it has been found that the difference in *shape* between $T^k_{pack,unif}$ at different time steps can be, in most cases, neglected. Their achieved maximum/minimum values do, however, change. For this reason, the fine grid interpolation of two profiles, followed by a surface scaling in order to match the maximum/minimum values predicted at each specific time step by the coarse FEM, suffices.

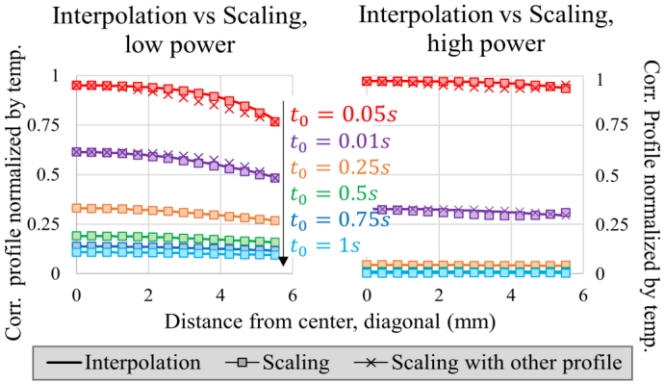

Fig. 4. Normalized correction profiles at different time steps for the LP (left) and the HP (right) packages. Full lines represent results obtained by interpolation at each single time step while square markers the ones got by the scaling approach. Crosses are for results obtained by interpolating, and then scaling, just one temperature profile instead of two.

These two profiles are the first two in chronological order (at the end of the impulsive power and one time step later). These are the ones with the highest impact and, for which, therefore, higher accuracy is more appropriate. The creation of the highly resolved temperature profiles for all the other time steps is performed by scaling the second interpolated surface, the one obtained during the cooling down phase. In this way, a gain in computational time, which is proportional to the amount of time steps to steady state, is obtained. For 20 time steps, for example, this step of the algorithm is approximately 10 times faster.

Normalized correction profiles at different time steps are shown in Fig. 4 for the LP package, on the left, and for the HP package, on the right. The normalization is performed multiplying each real correction profile by the ratio between the maximum temperature obtained at that specific time step and the one at the end of the impulsive power dissipation. From the plots it is clear that the loss in accuracy introduced by the scaling approach is negligible.

The crosses in Fig. 4 denote the results that would be obtained by interpolating just one temperature profile instead of two. The crosses referring to the normalized correction at $t = 0.05s$ are obtained by scaling the temperature profile for $t = 0.1s$ and vice versa. For the HP package, the interpolation of two separate temperature profiles could be avoided but this is not the case for the LP configuration. This is mainly due to the lower cooling rate and higher spreading resistance in the latter situation. However, for general situations, the interpolation of two temperature profiles proved to be better.

E. Error Metric

In order to include the package thermal impact in the transient FTM results, the algorithm based on 2D space convolution plus time superposition needs to be implemented. This requires higher computational time than the basic,

transient FTM for stack configuration based on 3D-convolution [10]. This means that, if for certain specific package structures and cooling solutions, the package thermal impact is low, the 3D-convolution based algorithm could be applied and the correction procedure avoided. For this reason, an a priori estimation of the maximum relative improvement ($impr$) achievable at time t, by applying the transient correction procedure on top of the basic stack FTM, has been derived for uniform power dissipation, continuous in time.

Let's define the $\max(impr)$ at time t_k as

$$\max(impr(t_k)) = \max\left|\frac{\text{err reduction}(t_k)}{\text{exact solution}(t_k)}\right| = \max\left|\frac{(T^k_{stack,unif}-T^k_{exact,unif})-(T^k_{pack,unif}-T^k_{exact,unif})}{T^k_{exact,unif}}\right| \approx \frac{T^k_{stack,unif}-\min(T^k_{pack,unif})}{\min(T^k_{pack,unif})} \quad (12)$$

where $T^k_{exact,unif}$ is the detailed, packaged FEM solution at time step k, when $t = t_k$.

Due to the definition of the BCs, which ensures that, at steady state, $T_{stack,unif} \approx T_{pack,unif}$, the maximum error reduction is achieved in the corners of the die, where the spreading effect is higher. Moreover, being $T^k_{exact,unif}$ unknown, it is approximated by $T^k_{pack,unif}$ and the maximum of the ratio representing $impr$ is achieved for $\min(T^k_{pack,unif})$, the value in the corner. This information can be easily obtained from the coarse FEM packaged results. The minimum temperature data, MIN, needed in the scaling phase (Section 4.D), can be used to this aim. The only difference is that those data are obtained for impulsive power dissipation while, in this case, the error metric is derived for continuous power. A cumulative sum of all these minimum values provides the desired quantity as a function of time:

$$\max\big(impr(t_k)\big) = \frac{T^k_{stack,unif}-\text{cumsum}(MIN)}{\text{cumsum}(MIN)}. \quad (13)$$

Despite the definition of the BCs for the FTM, $T^k_{stack,unif}$ cannot be substituted by the cumulative sum of the maximum temperature values of the coarse FTM. This is because the match between the maximum temperature values in the stack and the package configuration is imposed for steady state. The significant role played by the difference in thermal capacitance between the two configurations needs to be taken into account when considering the relative improvement achievable from modeling a package rather than a stack configuration.

The estimation of the relative error reduction can be easily and quickly computed and it provides useful information about the thermal impact of the package. It allows, therefore, to decide a priori whether the improvement, achievable including the package effect, justifies the higher computational time of the 2D-convolution based methodology. For the two scenarios considered in this paper, for example, $\max\big(impr(1.6)\big) = 0.5$ for the LP configuration while $\max\big(impr(1.6)\big) = 0.02$ for the HP configuration. As a consequence, just the results concerning the LP configuration are shown hereafter.

V. Results

In order to prove the accuracy of the methodology, comparisons have been performed with respect to experimentally validated FEM models [16]. The test case shown in the following refers to a stack of two 8 x 8 mm^2 dies in a face to back configuration. The 200 µm thick top die is connected to the 50 µm thick bottom die through a 13 µm thick layer composed of µbumps and underfill. The stack is mounted on a 330 µm thick substrate and overmolded (cf. LP package configuration in Fig. 1). The final dimensions of the packaged chip are 13.6 x 13.6 x 2.97 mm^3. A resolution of 100 x 100 µm^2 is assumed in space while a time step of 50 ms is considered. The total simulated time is 1.6 s, which is shorter than the time constant of the system. For this reason, also the HSRs are recorded until 1.6 s. The power dissipation is non-uniform in space and non-constant in time.

A. FEM Validation

Fig. 3 (a) shows the maximum temperature increase on the top (red) and the bottom (blue) die as a function of time. Since the PMs vary with time, the location of the maximum temperature is not fixed. Full lines represent the results obtained by the corrected FTM, dashed lines by the basic, uncorrected transient FTM and circles the ones from the FEM, with respect to which the FTM is validated. As it is visible, a significant improvement is achieved by applying the correction procedure.

Fig. 3 (b) shows the percentage error, defined as

$$\%err(M,t) = \frac{|T_{FEM}(M,t)-T_{FTM}(M,t)|}{T_{FEM}(M,t)} \cdot 100 \quad (14)$$

where M is the location of maximum temperature at time t, T_{FEM} are the temperature increases obtained by FEM and T_{FTM} the ones obtained by FTM. T_{FTM} can refer to results obtained by different FTM approaches, according to what is specified in the legend. For readability purpose, just the errors on the bottom die are shown. The full and the dashed blue lines refer, respectively, to the errors with and without the package correction application. For this test case, the selection of the computationally more expensive algorithm allows to keep the error at maximum temperature always below 5%. The choice of the faster FTM option without package correction would result in an error up to 35% at the end of the simulation. It is worth to note that the error referring to the basic FTM methodology has a tendency to grow with time. This is due to the increasing impact of the package during chip activity. Immediately after power dissipation, indeed, just the die stack and a small part of the package affect the temperature rise. As time passes, more package volume is involved in the heat storage and transport and, therefore, its impact on the final result increases. The fast increase of the error related to $T_{FTM,basic}$ in the final stage of the simulation is due to the fact that no power is dissipated after 1.4 s. This means that the system is cooling down at this stage and just old power inputs, for which the package impact is higher, are present. Including the package correction in the algorithm accounts for all these aspects.

In Fig. 3 (c) the temperatures obtained in a fixed location by

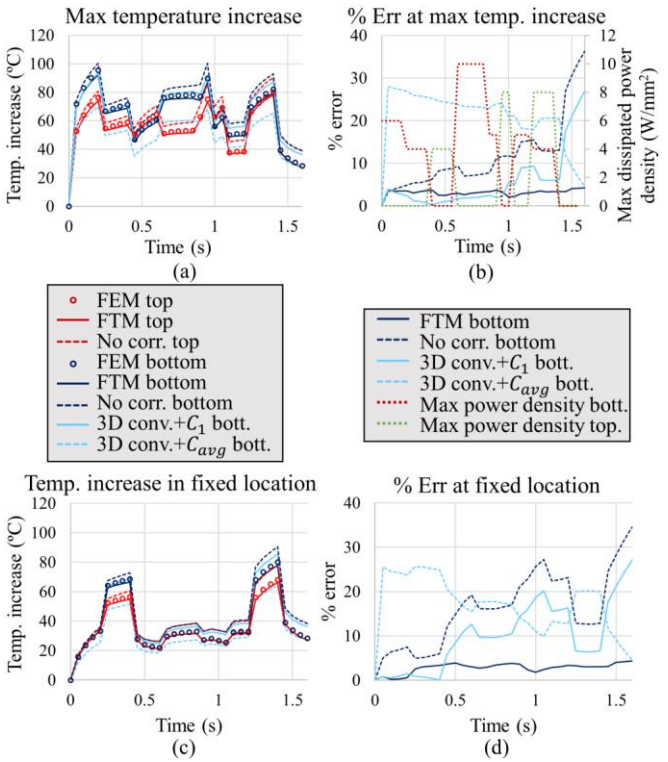

Fig. 3. Results obtained for a LP package configuration and time varying power maps. Different curves refer to results obtained with different methodologies and/or to different dies according to the legend. (a) Maximum temperature increase. (b) % Error in the location of the maximum temperature. On the right vertical axis, the maximum dissipated power density is reported. The dotted curves refer to this axis. (c) Maximum temperature increase in a fixed location. (d) % Error in the same fixed location as in Fig. 5 (c).

different methods are reported, with the same legend as in Fig. 3 (a), as a function of time. The graph demonstrates that the application of the package correction significantly improves accuracy everywhere, not just in the locations of maximum temperature. Fig. 3 (d) is the analogous as Fig. 3 (b) referring to the fixed point results in Fig. 3 (c). Similar remarks as for Fig. 3 (b) can be made here. The cooling down experienced by the system in this location, at 0.4 s and 1.4 s, confirms, once more, the importance of implementing the package correction to achieve high accuracy everywhere. The error related to $T_{FTM,basic}$ shows, indeed, a fast increase around these points in time, which doesn't appear when the correction is applied.

Concerning the computational time, a two orders of magnitude speed-up has been achieved when comparing the FTM, implemented in Matlab, and the FEM, implemented by using the commercial software MSC.Marc (~7min vs ~20hr, 16 x Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz. In both cases, the parallel license is not available).

B. Alternative package correction approaches

The error metric proposed in Section 4.E gives an indication of the improvement, with respect to the basic stack FTM, achievable when the package thermal impact is included via the algorithm illustrated in this paper. However, there could be other options in between the basic, stack FTM algorithm and the packaged one. Methodologies in which the correction is applied a posteriori, on the final temperature profile at time t, might, for example, be considered. In these cases, the subdivision of the power dissipation into its constituent impulsive components is unnecessary and 3D-convolution can be applied, with the advantage of shorter computational time. While for the correction methodology presented before, multiple, time dependent, correction profiles are used, in these approaches just one is selected. In this paper, two options are proposed for the single correction profile: (1) an average, C_{avg}, of the different C_k and (2) C_1, which is obtained at the end of the power impulse. This second choice is driven by the fact that, as already stated in Section 4.D, the latest dissipated power impulse has, in most cases, the biggest impact on the final temperature profiles.

Results obtained with these algorithms for the bottom die are shown in Fig. 3 with light blue curves. Full lines represent results obtained by applying C_1 to the 3D-convolution final results ($T_{FTM,basic}$), while dashed lines the results obtained by applying C_{avg}. The figures prove that the approach based on the application of an average correction after 3D-convolution does not provide any improvement in accuracy with respect to the basic FTM. This is because, even if the *shape* of the correction profiles is almost always the same (Section 4.D), these profiles are scaled between different extreme values and, thus, the applied corrections significantly differ from time to time.

The results referring to the application of C_1 after 3D-convolution in Fig. 3 (a) and (b) show, instead, good accuracy, mainly in the heating up phase. During the cooling down phase (after 1.4 s), however, the accuracy is drastically reduced. This is because, since C_1 is computed at the end of the power pulse, it is not representative for the cooling down phase. Other correction factors should be used in this situation, since no power is being dissipated at *present*. It has to be noted that the plotted curves refer to the maximum temperature, whose location varies during chip activity and follows the power dissipation position.

If the location is fixed, as in Fig. 3 (c) and (d), the accuracy of the C_1 correction approach deteriorates. This is, once more, due to the fact that, with respect to the previous situation, cooling-down phases have higher importance in a fixed location. In this case the application of C_1 after 3D-convolution is better than no correction at all but it is much worse than the fully corrected method based on 2D-convolution.

This means that, if the interest is just in accurately predicting peak temperatures during chip activity, then the application of C_1 after 3D-convolution provides sufficiently accurate results in shorter computational time. On the other hand, if more importance is given to the whole temperature profile, for which this high resolved FTM has been developed, the computationally more expensive methodology based on 2D-convolution plus time dependent package corrections performs much better. However, in case of time constraints or reduced package impact, the algorithm based on 3D-convolution followed by C_1 correction represents a good alternative.

VI. CONCLUSIONS

In this paper a transient FTM methodology for 3D, packaged, ICs has been presented. It can be considered as a multi-scale

strategy whose core is constituted by a convolution based algorithm that allows computing the temperature increase, due to a generic, time varying, power map in a stack configuration. The package spreading and capacitive effect is included via correction profiles. They are based on the ratio between the time dependent thermal responses of the package and of the stack configurations to uniform, impulsive, power dissipation. In order to apply these corrections, 2D-convolution with subsequent time superposition needs to be implemented, instead of the less computationally expensive 3D-convolution, to obtain the basic, time dependent, FTM temperature profiles. Nevertheless, a two orders of magnitude speed-up in computational time is achieved with respect to FEM. Moreover, by applying this correction strategy, the maximum error on the peak temperature is reduced from 35% to 5%.

An error metric is also provided to allow the user to decide if, for a specific situation, the relative improvement, achievable including the package impact, is worth the higher computational time. Other strategies, which aim to implement the package correction while keeping a 3D-convolution based algorithm, are also presented and compared with the basic FTM and the 2D-convolution based corrected algorithm. From the comparisons it results that this last option provides a significantly higher accuracy all over the die.

REFERENCES

[1] A. K. Coskun, T. S. Rosing, K. A. Whisnant, and K. C. Gross, "Static and dynamic temperature-aware scheduling for multiprocessor SoCs," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst*, vol. 16, no. 9, pp. 1127-1140, 2008.

[2] Z. Liu, S. X. D. Tan, H. Wang, Y. Hua, and A. Gupta, "Compact thermal modeling for packaged microprocessor design with practical power maps*," Integration, the VLSI Journal*, vol. 47, no. 1, pp. 71-85, 2014.

[3] Y. C. Gerstenmaier, and G. Wachutka, "Time dependent temperature fields calculated using eigenfunctions and eigenvalues of the heat conduction equation," *Microelectronics journal*, vol. 32, no. 10, pp. 801-808, 2001.

[4] F. Christiaens, B. Vandevelde, E. Beyne, R. Mertens, and J. Berghmans, "A generic methodology for deriving compact dynamic thermal models, applied to the PSGA package," *IEEE Trans. Compon., Packag., Manuf. Technol. A*, vol 21, no. 4, pp. 565-576, 1998.

[5] A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza, "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling," in *Proc. ICCAD*, 2010, pp. 463-470.

[6] B. Barabadi, Y. K. Joshi, and S. Kumar, "Rapid multi-scale transient thermal modeling of packaged microprocessors using hybrid approach," in *Proc. EPTC*, 2012, pp. 157-164.

[7] F. Beneventi, A. Bartolini, A. Tilli, and L. Benini, "An effective gray-box identification procedure for multicore thermal modelling," *IEEE Trans. Comput.*, vol. 63, no. 5, pp. 1097-1110, 2014

[8] J. H. Park, A. Shakouri, and S. Kang, (2008, March). "Fast evaluation method for transient hot spots in VLSI ICs in packages," in *Proc. ISQED*, 2008, pp. 600-603.

[9] Y. C. Gerstenmaier, and G. K. Wachutka, "Efficient calculation of transient temperature fields responding to fast changing heat sources over long duration in power electronic systems," *IEEE Trans. Compon. Packag. Technol.*, vol. 27, no. 1, pp.104-111, 2004.

[10] F. L. T. Maggioni, H. Oprins, D. Milojevic, E. Beyne, I. De Wolf, and M. Baelmans, "3D-Convolution Based Fast Transient Thermal Model for 3D Integrated Circuits: Methodology and Applications," in *Proc. SEMI-THERM*, 2015, pp. 107-112.

[11] F. L. T. Maggioni, H. Oprins, E. Beyne, I. De Wolf, and M. Baelmans, "Fast convolution based thermal model for 3D-ICs: Methodology, accuracy analysis and package impact," *Microelectronics Journal*, vol. 45, issue 12, pp. 1746-1752, 2014.

[12] V. M. Hériz, J. H. Park, T. Kemper, S. M. Kang, and A. Shakouri, "Method of images for the fast calculation of temperature distributions in packaged VLSI chips," in *Proc. THERMINIC*, 2007, pp. 18-25.

[13] S. H. Pan, N. Chang, and T. Hitomi, "3D-IC dynamic thermal analysis with hierarchical and configurable chip thermal model," in *Proc. 3DIC*, 2013, pp. 1-8.

[14] J. Yu, (1971). "Application of conformal mapping and variational method to the study of heat conduction in polygonal plates with temperature/dependent conductivity," *Int. J. Heat Mass Tran.*, vol. 14, no. 1, pp. 49-56, 1971.

[15] G. N. Ellison, "Maximum thermal spreading resistance for rectangular sources and plates with nonunity aspect ratios," *IEEE Trans. Compon. Packag. Technol*., vol. 26, no. 2, pp. 439-454, 2003.

[16] H. Oprins, V. Cherman, G. Van der Plas, F. Maggioni, J. De Vos, T. Wang, R. Daily, and E. Beyne "Experimental thermal characterization and thermal model validation of 3D packages using a programmable thermal test chip," in *Proc. ECTC, 2015.*

Federica L. T. Maggioni received her B.Sc. and M.Sc. in Applied Mathematics from Università degli studi di Milano, Milano, Italy in 2009 and 2011 respectively. She is currently pursuing her Ph.D. degree in Mechanical Engineering at KU Leuven, Leuven, Belgium and in collaboration with imec, Leuven, Belgium. Her current research topic is the development of computationally fast modeling techniques for the thermal analysis of 3D-ICs.

Herman Oprins received the M.Sc and Ph.D. in Mechanical Engineering from KU Leuven. He joined imec in 2003, pursuing the Ph.D. with KU Leuven and working on modeling and experimental projects on thermal management of electronic packages. Since 2009 he is working as senior research engineer in imec, where he is involved in the thermal experimental characterization, thermal modeling and management of 3D-ICs, electronic packages, GaN transistors, photovoltaic modules and microfluidics.

Eric Beyne (M'83) received the master's degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1983 and 1990, respectively. He has been with imec, Leuven, since 1986, where he has been involved in advanced packaging and interconnect technologies. He is currently a fellow with imec, where he is the Program Director of the 3-D System Integration Program.

Ingrid De Wolf received the Ph.D. in Physics from the KU Leuven. In 1989 she joined imec, where she worked in the field of microelectronics reliability. From 1999 to 2014, she headed the group REMO, focused on reliability, test and modelling. She authored or co-authored several book chapters and more than 350 publications. She is chief scientist in imec, IEEE senior member and professor at the Dep. of Materials Engineering of the KU Leuven.

Martine Baelmans is professor at the Dep. of Mechanical Engineering at KU Leuven, where she leads the group on thermal-fluid engineering since 1996. She graduated from KU Leuven and obtained her Ph.D. in Engineering in 1993. She has authored or co-authored more than 250 papers in applications on fluid mechanics and heat transfer. Research applications range from thermal management in power electronics, power transformers and energy systems over liquid and two-phase microelectronics cooling.