Dynamic Compact Thermal Models With Multiple Power Sources: Application to an Ultrathin Chip Stacking Technology

Jordi Palacín, Marc Salleras, Josep Samitier, and Santiago Marco

***Abstract*—Whereas numerical modeling using finite-element methods (FEM) can provide transient temperature distribution in the component with enough accuracy, it is of the most importance the development of compact dynamic thermal models that can be used for electrothermal simulation. While in most cases single power sources are considered, here we focus on the simultaneous presence of multiple sources. The thermal model will be in the form of a thermal impedance matrix containing the thermal impedance transfer functions between two arbitrary ports. Each individual transfer function element $H_{ij}(s)$ is obtained from the analysis of the thermal temperature transient at node "i" after a power step at node "j." Different options for multiexponential transient analysis are detailed and compared. Among the options explored, small thermal models can be obtained by constrained nonlinear least squares (NLSQ) methods if the order is selected properly using validation signals. The methods are applied to the extraction of dynamic compact thermal models for a new ultrathin chip stack technology (UTCS).**

***Index Terms*—Dynamic compact thermal models, modeling, thin electronics.**

I. Introduction

THERMAL effects have an obvious impact in the performance and reliability of electronic systems. There is a continuous trend toward miniaturization and higher degrees of integration. This trend shows in both system-on-chip and system-on-package paradigms, posing new challenges to the electronic design authomation (EDA) industry. It is clear that simulation at the physical level including all geometrical details and physical interactions is not feasible, not even wished. Consequently, simulation is based in compact models at different levels of abstraction. The most well-known compact model is the transistor model. To take into account thermal effects, transistor behavior should be assisted by compact thermal models. At the printed circuit board design level, compact thermal models of packages are essential.

On the other hand, in the past, the analysis of electronic systems was mainly done in isothermal conditions. That is, all devices were kept at the nominal operation temperature. However, dynamic thermal effects have an important role in electronic systems at different time and size scales. At the transistor level, time constants on the order of nanoseconds can lead to strong coupling to system simulation. On the other extreme of the scale, package and system time constant can reach tens of seconds or even minutes. The simulation in such cases is also hindered because of the presence of stiff differential equations.

Dynamic compact thermal models mostly show their usefulness in the design and simulation of systems requiring thermal management (for instance, high-performance computing platforms [1] or power drivers [2]), or just because they operate in a discontinuous manner to save battery. It has also received attention in the area of multichip-modules where power components can be integrated together with control, mixed-signal, or radio frequency (RF) chips. The commutation of the power chip can produce transient temperatures in the other chips, affecting their performance [3], [4].

Microsystems on the other hand are an emerging application where dynamic compact thermal modeling is a must. A number of sensors and actuators rely on thermal operation principles. In some cases, even the operation principle relies on thermal dynamics. Examples are thermal actuators like micropumps [5], active valves [6], shape-memory alloy-based actuators [7], bimetallic-based actuators [8], micropyrotechnic actuators [9], or thermal ink jet heads [10]. For sensors, a typical example is temperature-modulated gas sensors on micromachined hot-plates [11].

Dynamic compact thermal models have been addressed by a number of authors (see some references above and the review by Sabry [12]). In most cases, single power sources are considered, but, here, we focus on the simultaneous presence of multiple sources. The developed methodology will be applied to the extraction of compact thermal models for an ultrathin chip stacking technology (UTCS) with a very large integration density [13], [14]. In the problem at hand, the concurrent presence of several very close power sources and the use of low thermal conductivity materials was considered as the rationally for starting a full set of simulations at the physical level. Main conclusions of this study were presented elsewhere [15].

In this paper, we will focus our attention to the development of compact thermal models from the analysis of the thermal impedance transients obtained from physical simulation (finite-element model). In Section II, we will review the methodology for compact thermal model extraction. In Section III, the UTCS technology will be briefly reviewed. In Section IV, the physical finite-element model (FEM) of the structure will be presented

Manuscript received July 17, 2003; revised September 2, 2004. This work was supported in part by ESPRIT under Project UTCS 24910.

J. Palacín is with the Departament d'Informàtica i Enginyeria Industrial, Universitat de Lleida, 25001 Lleida. Spain, and is also with the Sistemes d'Instrumentació i Comunicacions, Departament d'Electrònica, Universitat de Barcelona, 08028 Barcelona, Spain.

M. Salleras, J. Samitier, and S. Marco are with the Sistemes d'Instrumentació i Comunicacions, Departament d'Electrònica, Universitat de Barcelona, 08028 Barcelona, Spain.

Digital Object Identifier 10.1109/TADVP.2005.850507

together with the main simulation results. In Section V, the analysis of the simulated transients will be presented and several options for model extraction will be compared. Finally, in Section VI, some conclusions will be drawn.

II. Dynamic Compact Thermal Models

In the past, most of the analysis concerning to the thermal behavior of packages were focused on stationary measures of the junction-to-case thermal resistance. For multichip modules (MCM) packages, resistor networks have been used. For the extraction of the compact models, either experimental measurements or physical simulations can be used. While for resistor networks, steady state values are enough [16], for the extraction of dynamic models, we need to have the dynamic evolution of the temperature at different points of the structure.

This can be done experimentally with a fully instrumented package where temperature sensors are strategically located. However, undertaking such an effort presents several difficult issues. Experimental measurements can be seriously polluted with noise and may also suffer from material nonlinearities. Alternatively, it is possible to create a linear numerical model of the electronic system. In such a physical model temperature probes may be located at single nodes of the model or over distributed areas or volumes. Additionally, we have an easy control of the applied power and boundary conditions. Of course, at the end, the physical model has to be always validated through experimental measurements.

The most common case in the literature is the extraction of single-port compact thermal models [17]. By single port we understand those models where the power dissipation and the temperature measurement are carried out at the same point, or in very close proximity. In those cases, the continuum time solution of the heat diffusion equation at this particular point of the structure after a power step is usually known as thermal self-impedance transient. If the point whose temperature is monitored is remotely located from the power source, we use the term thermal trans-impedance transient.

Self-impedance thermal transients can be approximated by a strictly positive multiexponential transient

$$\frac{T(t)-T_i}{P}=\sum_{i=1}^{N}R_i\left(1-e^{-\frac{t}{\tau_i}}\right) \tag{1}$$

where $T(t)$ [°C] is the temperature of interest, T_i [°C] is the initial reference temperature, P [W] is the power dissipation step, R_i [°C/W] is the preexponential coefficient, and τ_i [s] is the time constant. This step response corresponds to a system with the following thermal impedance in the Laplace domain:

$$Z(s)=\sum_{i=1}^{N}\frac{R_i}{1+s\tau_i}. \tag{2}$$

A Foster *RC* network can represent this transfer function easily, where each cell contains a resistor R_i in parallel to a capacitor C_i. Moreover, it can be seen that $\tau_i = R_iC_i$. This model has the advantage that there is a straight correspondence between the fitting parameters and the network elements. However, in the thermal-electrical analogy to represent the thermal behavior of components, such a representation has no physical meaning. For this to occur, all the capacitances have to be grounded. In virtue of this, the Cauer network is usually preferred. In order to transform the Foster network in a Cauer network, the so-called Foster–Cauer transformation has to be applied [17].

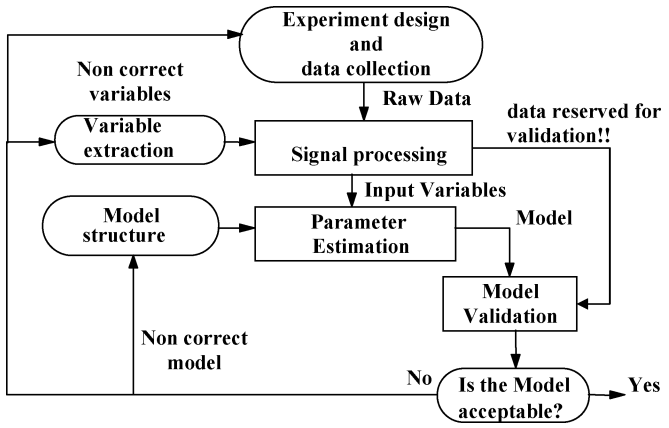

Fig. 1. General modeling procedure.

The analysis of the thermal impedance transient encompasses several steps that are shared to any dynamic modeling procedure. Such steps and the loop to follow are summarized in Fig. 1. The main steps are as follows:

1) experiment design and data collection;
2) signal processing, for instance, filtering;
3) model structure or model order selection, in this case, how many exponentials do we want to include in our model;
4) parameter estimation, usually by a least squares procedure;
5) model validation.

Concerning this general modeling procedure, several remarks are in order.

1) Multiexponential analysis is a classic problem in signal processing. Because exponential decays are not orthogonal, the process is extremely sensitive to noise and truncation. In fact, in the signal processing parlance, we would say that the fitting of a multiexponential transient is an ill-posed or ill-conditioned problem. This means that small differences in the signal under analysis can lead to strong deviations in the estimated parameters. This factor combined with the ubiquitous nature of multiexponential decays has produced an extremely rich literature [18].
2) Multiexponential fitting is a nonlinear minimization problem and is consequently prone to become trapped in a local minimum.
3) The errors of the fitting procedure always decrease by adding exponential terms. However, this can lead to overfitting and to gross inaccuracies in the estimated parameters. A parsimony principle ("Occam's razor") has to be applied either using cross-validation or theoretically based penalty functions as the Rissanen Minimum Description Length (MDL) or the Akaike Information Criterion (AIC). Unfortunately, small model orders can also give very biased estimates of the parameters because of model inadequacy [19].

Because of the difficulties regarding the analysis of multiexponential transients, a number of methods have been applied for the analysis. We may distinguish between nonparametric methods and parametric methods. Nonparametric methods do not assume a particular model order and deliver a continuous time-constant spectrum, and on many occasions, they are based in the deconvolution procedure first proposed by Gardner [20]. These methods have been advocated by Szekely [17] and the authors [21]–[23].

It is beyond the scope of this paper to review these techniques, but a thorough description can be found in [21], [23]. In this paper, we have used multiexponential transient spectroscopy (METS) followed by Jansson deconvolution (see [23] for details). In any case, it is worthwhile to remind that these techniques produce a continuous time constant spectrum description of the step response

$$\frac{T(t)-T_i}{P}=\int_0^\infty G(\tau)\left(1-e^{-\frac{t}{\tau}}\right)d\tau \tag{3}$$

where $G(\tau)$ $[\mathrm{K}/(\mathrm{W}\cdot\mathrm{s})]$ is the time constant spectrum of the thermal impedance transient. In the discrete case, the time constant spectrum is given by

$$G(\tau)=\sum_{i=1}^{N} R_i\delta(\tau-\tau_i). \tag{4}$$

Nonparametric or deconvolution techniques have several advantages: they can deal with discrete and continuous time constant distributions, and they can give indications about the model order, presence of negative amplitude terms, and an initial approximation to the value and position of the time constants. However, they suffer from a fundamental problem, and this is that being a nonparametric tool they do not provide the Foster parameters needed to build the thermal *RC* network. Moreover these deconvolution methods are not formulated to minimize any loss function (e.g., least squares).

A close approximation to this approach, although based on linear least squares estimators, is the so-called exponential series method (ESM) [24]. In this method, a high density of fixed probe exponential functions covering a wide range of time constants are used to approximate the thermal impedance transient. Because the time constants are now fixed, this approach converts the general nonlinear least squares (NLSQ) problem of fitting multiexponentials to a linear one, avoiding the problem of local minima. However, this method suffers from collinearity of the basis functions. For the better performance, it is recommended to have the probe functions equally distributed in the logarithm of time, as well as the sampling points. In this way, equal importance is given to all time decades; otherwise the fitting procedure is focused on the good estimation of only the slowest part of the transient. However, when very high density probes are used, this method cannot be considered as a parametric approximation, and due to the nonorthogonal character of the exponentials, the presence of a certain exponential term tends to spread out in neighboring time constants providing a broader peak, instead of a single exponential. It can be argued that a high number of fitting parameters can provoke overfitting. However, the optimum density of exponential functions can be determined from validation results after a scan in the number of exponentials per decade. For self-impedance transients, it is known that all the exponential terms have positive amplitudes. For this reason, it is convenient to use a nonnegative least squares algorithm [25].

For a real parametric representation, nonlinear least squares fitting in the time domain is the straightforward option. As previously mentioned, this method is quite prone to errors so some basic precautions are needed. First and most important when dealing with experimental transient is digital filtering followed by subsampling to have points log spaced. A suboptimal method with a moving average filter of adaptive length has been described by the authors and shown to provide better results than other heuristic methods [26]. The second recommended option is to use an adequate initialization procedure taking into account the results of the nonparametric deconvolution process. Random initialization can provide very variable and inconsistent results. For this, a least squares problem is also recommended to make the variable transformation $z=ln(t)$ to avoid the presence of negative time constants and to constraint the solutions to positive amplitudes.

However, for MCM packages, several chips can contribute to the total power dissipation, so it is not enough to consider the single-port model for every isolated chip. It would be very convenient to have a single dynamic compact model that would represent the whole package. The extraction of dynamic thermal networks considering several power sources has been addressed only recently [27].

As in the previous case, the problem can be split in model structure selection and parameter estimation. Two main approaches can be distinguished: Christiaens *et al.* [28] start by a heuristic proposal of a single *RC* multiport network. The topology and complexity of this network is selected by the engineer from his knowledge of the internal structure of the package. This method cannot be considered as optimal since it largely depends on the physical understanding of the heat pathflow within the package. So the selection of the optimal topology remains an open question. Once the *RC* network topology has been selected, the next step is parameter estimation by least squares. For this purpose, the authors transform the data, from the step response to the impulse response and then to the frequency domain, where the behavior of the system can be easily described in terms of the conductance and capacitance matrices. However, this transformation to the frequency domain requires the use of the fast Fourier transform in the time scale using a constant sampling period. Since the thermal transients may span several time decades, the total amount of points can be very large leading to large computational costs.

The second proposal by Szekely *et al.* consists on the formulation of a thermal admittance matrix between the different ports [17]

$$\begin{pmatrix} P_1(s) \\ \vdots \\ P_n(s) \end{pmatrix} = \begin{pmatrix} Y_{11}(s) & \cdots & Y_{1n}(s) \\ \vdots & \ddots & \vdots \\ Y_{n1}(s) & \cdots & Y_{nn}(s) \end{pmatrix} \begin{pmatrix} T_1(s) \\ \vdots \\ T_n(s) \end{pmatrix} \tag{5}$$

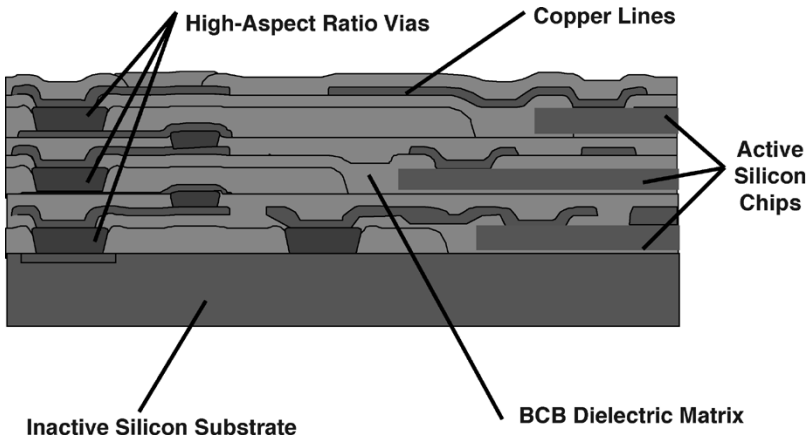

Fig. 2. Basic UTCS structure.

where P_i is the dissipated power at the input port (i), and T_i is the temperature at the same port. Every element in the matrix leads to an *RC* network and these networks appear coupled by voltage-controlled current sources (VCCS). The main drawback of this method is that the analysis of every element in the admittance matrix is made independently, leading to an unnecessary increase in the number of parameters to estimate. On the other hand, as far as we know, no restrictions are included to assure the physical interpretability of the generated models. In consequence, floating capacitances may appear to represent the behavior of the elements outside the diagonal.

Here, we will explore a hybrid approach between both proposals. In fact, our method has been inspired by the work of Prof. Szekely, but some modifications have been included in the modeling procedure.

First, noticeable difference is that we use the impedance matrix instead of the admittance matrix. In this case, and by virtue of the superposition principle, the temperature increase at a certain point will receive as many contributions as power sources in the model

$$\begin{pmatrix} T_1(s) \\ \vdots \\ T_n(s) \end{pmatrix} = \begin{pmatrix} Z_{11}(s) & \cdots & Z_{1n}(s) \\ \vdots & \ddots & \vdots \\ Z_{n1}(s) & \cdots & Z_{nn}(s) \end{pmatrix} \begin{pmatrix} P_1(s) \\ \vdots \\ P_n(s) \end{pmatrix}. \quad (6)$$

For the analysis of every element in the matrix we need the temperature transient at the different points of interest: basically, the thermal ports, after a power step. It is important to remark that this impedance matrix constitutes the compact thermal model of the component, and it permits to predict the temperature time evolution at the thermal ports under arbitrary power loads. In particular, in the results section, it will be used to predict temperatures for random power signals.

Moreover, if the thermal impedance matrix has to represent a unique *RC* network in the sense proposed by Christiaens, it is clear that some dependencies between the different elements of the matrix must appear. For our discussion, the most relevant point is that the different elements have to share the same set of poles. In other words, the temperature transients at the different ports have a common set of time constants. This restriction leads to a considerable reduction in the number of parameters to fit. For instance for a m thermal port system fitted with an n order model, we pass from $m(m+1)n$ in case no restriction is applied, to $n \cdot (1 + m(m+1)/2)$ parameters. To put some numbers, for a three-port system approximated with three exponentials, we go from 36 free parameters to 21 parameters. Finally, we have to take into account the different nature of the single-port transients, or self-impedance transients (power source and temperature probe at the same point) and the transients where the temperature probe is located remote to the power source: trans-impedance transients. While in the first set all the exponential amplitudes have to be positive, in the second set negative amplitudes must appear. It is important to incorporate such a constraint in the fitting procedure. We would refer to the option of sharing the poles as constrained least squares.

The application of these techniques to dynamic compact thermal models will be presented further after applied to the modeling of the UTCS structure. Before entering in more details, we would like to remark that in this particular case, we are more interested in the modeling of the first-level packaging, that is, the thermal interaction between the chips in the stack, rather than the thermal transfer to the ambient. This point will be clear in the description of the FEM model. In this paper, we do not address the methodology to obtain boundary-independent compact thermal models.

III. Ultrathin Chip Stacking Technology

Recent developments of packaging technologies in three-dimensional (3-D) stacking and ultrathin electronics have provided an opportunity for important savings in mass, volume, and power consumption. In the last years, thickness of packages has greatly diminished. The European ESPRIT project Ultra Thin Chip Stacking (UTCS) 24910 combines state-of-the-art techniques concerning wafer and chip thinning, transport, and attachment technologies, together with planarization techniques and high-density interconnects to provide unprecedented increases in integration densities.

The basic structure can be seen in the Fig. 2. Commercial chips of different technologies are thinned down to 15 μm, transferred to a host silicon substrate, and connected to each other using planarization and interconnection techniques already developed for MCM-D technologies [29]. Note that this

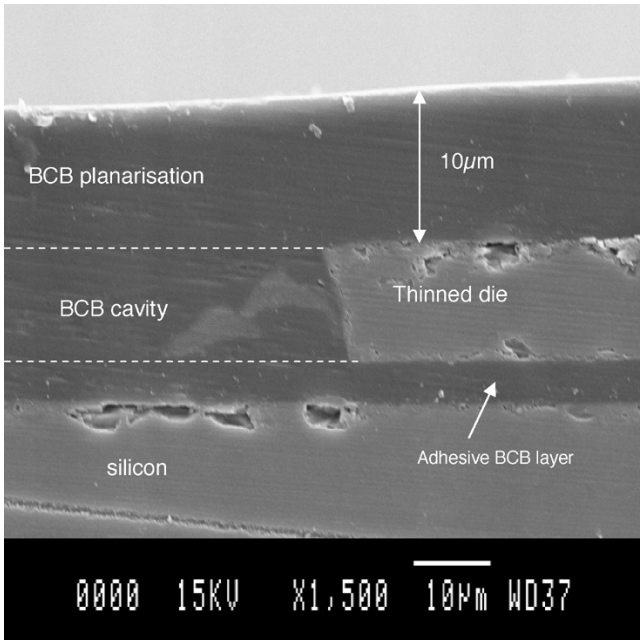

Fig. 3. UTCS cross section.

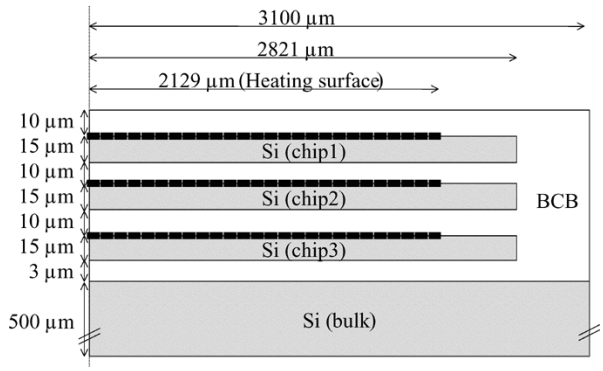

Fig. 4. Model scheme for physical simulation.

TABLE I
MATERIAL PROPERTIES

Material	Density [Kg/m^3]	Specific Heat [J/Kg.K]	Thermal conductivity [W/m.K]
Si	2330	836	117.5
BCB	1051	1176	0.18

technology uses extremely thin chips, when compared with other proposals in thin electronics that use $\sim$ 50-μm-thick chips [30]. The developed technology permits the successive stacking of thin chips on the same host. Vertical interconnection among different levels is achieved through high aspect ratio vertical interconnects (HARVI) based on the deposition of thick copper studs. Residual thermomechanical stresses in these structures due to the fabrication process were analyzed within the project [31]. Planarization and electrical insulation is accomplished by using BenzoCycloButene (BCB). Fig. 3. shows a scanning electron micrograph corresponding to the integration of the first-level UTCS structure. Note the BCB adhesive layer that isolates the thin chip from the silicon substrate. The process flow for the fabrication of a 3-D UTCS stacking is an extension of the MCM-D technology of IMEC (Leuven, Belgium). The main features are Ti/Cu lines for interconnections at the same level, Cu studs (HARVIs) for interconnections at different levels, and the use of Cyclotene 4020-40 and 4020-46 for planarization and cavity formation. BCB is chosen because its good planarization properties and because its low temperature processing. A detailed description of the technology can be found in [13], [32].

For this technology, the vertical integration principle itself is an obvious factor of heat concentration, and on the other hand, the use of BCB having a poor thermal conductivity, as adhesive and planarization material, can lead to increased thermal resistance.

While the stack can be considered as the first-level packaging, no predefined option for the second level package has been selected. Vertical heat flow is the expected flow-path in most packaging cases. For such thin-chip, an eventual flip-chip would be carried out with solder balls over the substrate but not on top of the stack. In this case, again the main flow-path will be across the thin-chips toward the silicon substrate.

IV. PHYSICAL MODELING

A scheme of the model is presented in Fig. 4. Three identical thin chips are integrated in the BCB matrix over the silicon host. To retain the basic thermal behavior of the structure with a minimum computational cost, an axisymmetrical model has been built in ANSYS 7.0. The bottom silicon substrate will appear as a heat sink, and the bottom surface will be considered isothermal. This will force most of the heat flow in the vertical direction. In other words, the thermal resistance from the chip to the case is not considered, only the thermal resistance induced by the thin-chip stack. The critical dimensions of the model are shown in the scheme. Each chip has an area of 0.25 cm^2. On the top of the thin chip, power is dissipated uniformly over an area of 0.14 cm^2. The host silicon chip has an area of 0.30 cm^2 and a thickness of 500 μm. The material properties used in the simulation have been considered independent of the temperature and are summarized in Table I.

The meshed model contains 4196 nodes to a total of 1746 elements. Most of the elements were quadratic quadrilaterals, except at the borders of the model where there is almost no variation of the temperature. Quadratic triangles were used for the transition from highly dense meshed areas to coarse meshed areas. The bottom surface is kept isothermal, while in the top and lateral surfaces, natural convection is applied with a heat exchange film coefficient of 2 W/m$^2 \cdot$ K. With these conditions, the preferred heat flow path is perpendicular to the chips.

To understand the thermal behavior of the structure, we present first the steady state results. As an example, the final temperature distribution when 1-W power is dissipated at the bottom chip is shown in Fig. 5. It can be observed that the maximum temperature gradient appears at the BCB region surrounding the heated chip (bottom chip) and that the temperature is not uniform along the thin chips. This can be more

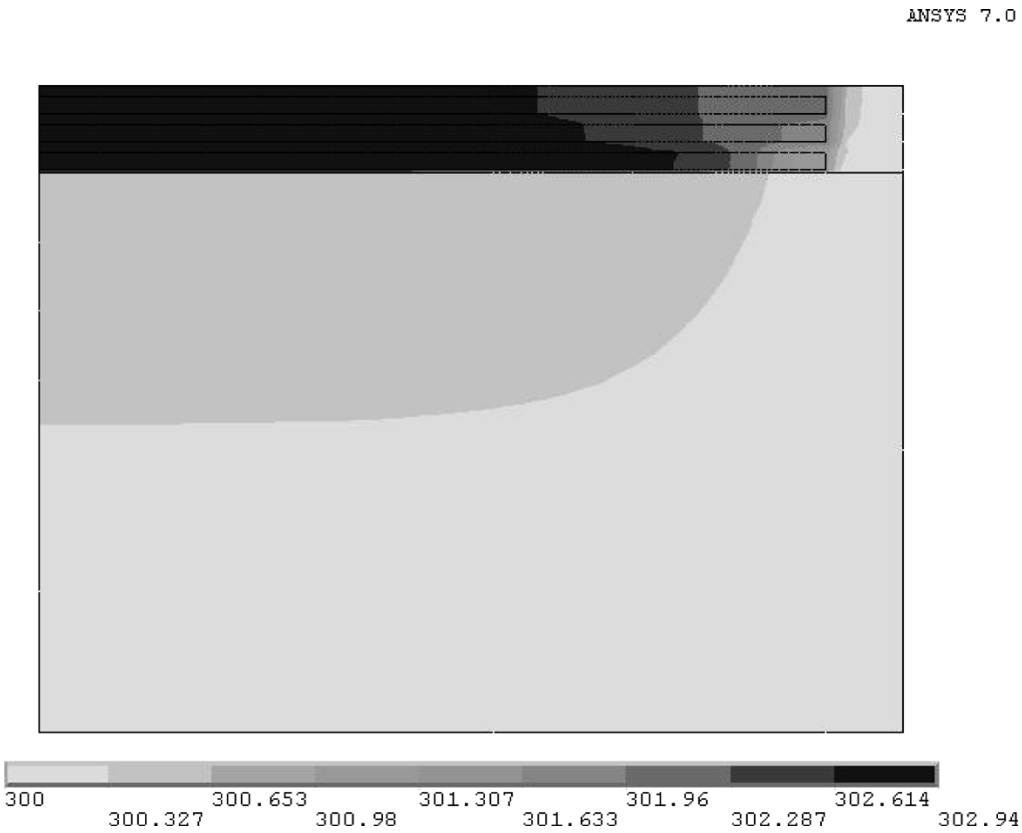

Fig. 5. Temperature distribution (K) for 1-W power dissipation on the bottom chip.

clearly appreciated at the top chip. Moreover, and since most of the heat flows toward the silicon substrate, the temperature gradient perpendicular to the chip toward the top surface is very small. The mean thermal resistance is: 18 °C/W for the top chip (chip 1), 11 °C/W for the intermediate chip (chip 2), and 3 °C/W for the bottom chip (chip 3). These values are not very high despite the low conductivity of the BCB due to the large area of the heat flow.

In the following discussion, we will consider that the junction of interest is located at the center of each chip. Power step temperature transients can be observed in Fig. 6, where $T_{xy}(t)$ identifies the transient temperature measured at chip (y) center, when applying power to chip (x): $T_{11}(t)$, $T_{22}(t)$ and $T_{33}(t)$ are self-impedance transients, and $T_{12}(t)$, $T_{13}(t)$, and $T_{23}(t)$ are the trans-impedance transients.

The starting point for the development of thermal models for this system are the temperature transients after a step power in every chip. It seems this would lead to at least nine temperature transients, but, in fact, only six of them are independent. In fact, we may arrange the transients in a matrix fashion as follows:

$$\begin{pmatrix} T_{11}(t) & T_{12}(t) & T_{13}(t) \\ T_{21}(t) & T_{22}(t) & T_{23}(t) \\ T_{31}(t) & T_{32}(t) & T_{33}(t) \end{pmatrix}. \tag{7}$$

This matrix is symmetric. These transients have been obtained directly from the FEM model with a log spaced time vector at 20 points/decade.

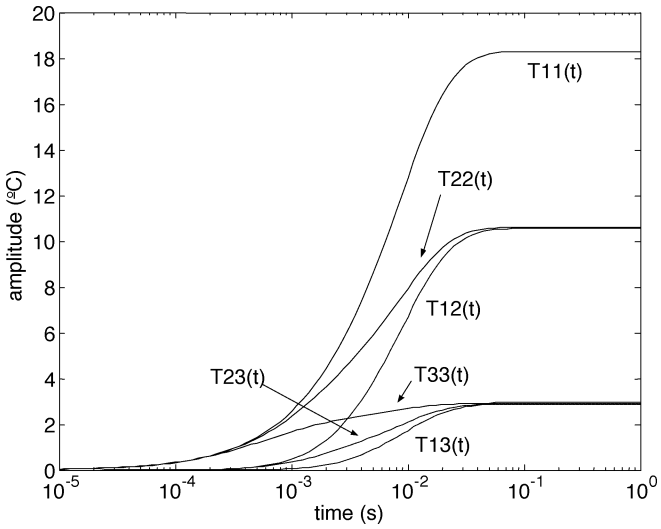

Fig. 6. Diagonal temperature transients (solid) and off-diagonal temperature transients (dotted). T_{ij} stands for measuring temperature at chip i and a power step applied at chip j.

Concerning the time evolution of the temperature, several facts can be noticed. On the one hand, the slowest temperature rise corresponds to the top chip due to the increased thermal resistance to the host silicon. Another point to notice is that every source chip acts as a heat spreader and in fact achieves almost uniform temperature before the other chips begin to heat up. This can be understood again on the basis of the big differences in thermal conductivity between Si and BCB.

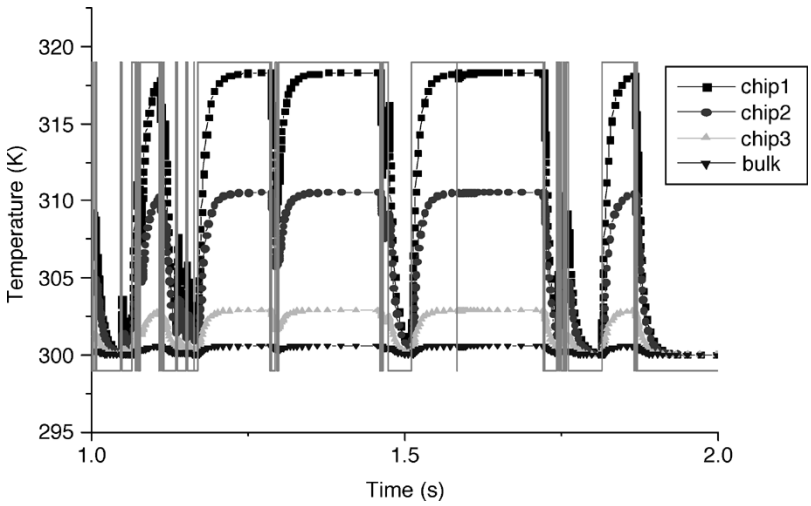

Fig. 7. Temperature waveform for model validation after a random power signal applied to chip 1. Solid line shows the power excitation switching between 0 and 1 W.

Finally, it can be observed that when powering chip 2, chip 1 reaches the same final temperature, and when powering chip 3, chips 1 and 2 also reach the same temperature. This is due to the preferred heat flow toward the silicon substrate that acts as heat sink.

V. Dynamic Thermal Compact Modeling

As stated in Section II, the multiexponential analysis of each temperature transient permits to obtain the elements of the thermal impedance matrix according (1), (2), and (6).

The analysis of each transient can be accomplished using different methods. To compare their relative merits, validation with an independent power waveform has been carried out. The validation signal is a random binary power signal (see Fig. 7). The power signal is applied to each chip separately, and the evolution of temperatures at the three chips is computed from a transient simulation of the physical model obtaining nine validation signals. For all the methods, we have computed the fitting root mean square (rms) error for the step response (or estimation error), but also the rms error in the prediction of temperatures with the validation power waveform.

First analyses of the transients have been performed using a deconvolution procedure (see Section II). In the elements of the diagonal or self-impedance transients, the nonparametric deconvolution technique provides a continuous spectrum of positive amplitudes, while for the off-diagonal elements, we obtain positive and negative amplitudes. The results can be observed in Fig. 8(a) and 8(b) (dotted line). This analysis reveals the presence of exponential terms in the range between 10^{-4} to 10^{-2} s. The rms errors for validation and estimation are listed in Table II.

A thermal model can be obtained by sampling the Jansson time constant spectrum. However, this cannot be considered a compact model, since the model order equals the number of samples of the spectrum.

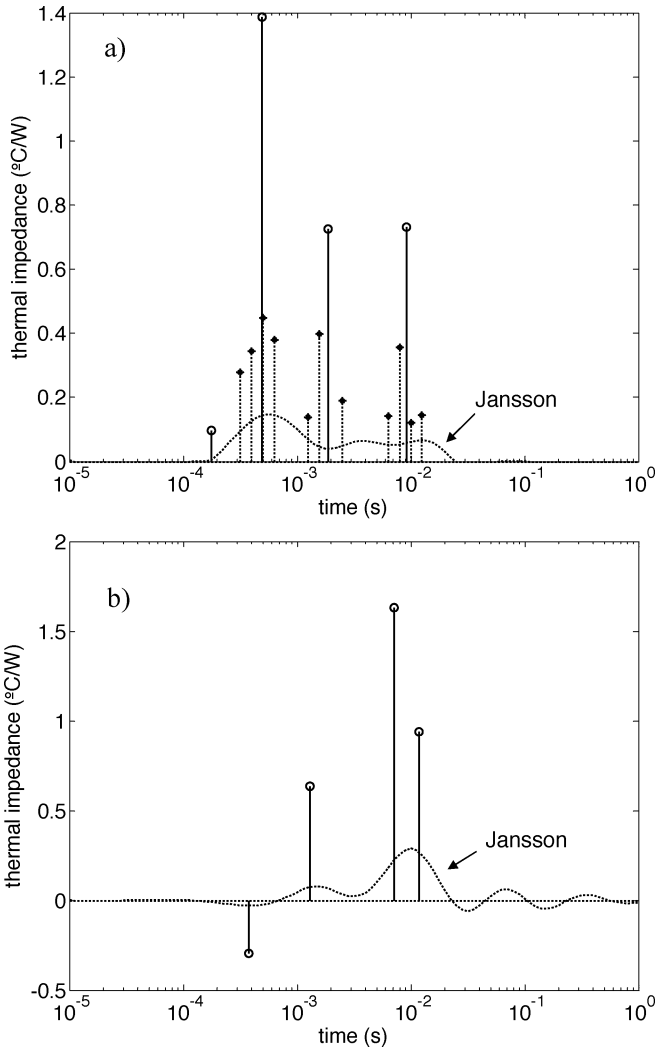

Fig. 8. (a) Self-impedance time spectrum for chip 3 (bottom chip): truncated Jansson deconvolution (dotted line), nonnegative ESM (dotted line discrete time-spectrum), and NLSQ free (solid line discrete time-spectrum). (b) Trans-impedance for chip 2 (middle chip) to chip 3 (lower chip): Jansson deconvolution (dotted line) and NLSQ free (solid line discrete time-spectrum).

TABLE II
RMS OF RESIDUALS: AVERAGE VALUES FOR DIAGONAL AND OFF-DIAGONAL TRANSIENTS

Method	Diagonal terms RMS error		Off-diagonal terms RMS error	
	Estimation	Validation	Estimation	Validation
Truncated Jansson.	0.26 ºC	0.22 ºC	-	-
Jansson Deconvolution	-	-	0.13 ºC	0.12 ºC
Non Negative ESM	0.00034 ºC	0.068 ºC	-	-
ESM	-	-	0.00097ºC	0.044ºC
NLSQ free	0.0029 ºC	0.069 ºC	0.0011 ºC	0.044 ºC
NLSQ constrained	0.026 ºC	0.099 ºC	0.021 ºC	0.061 ºC

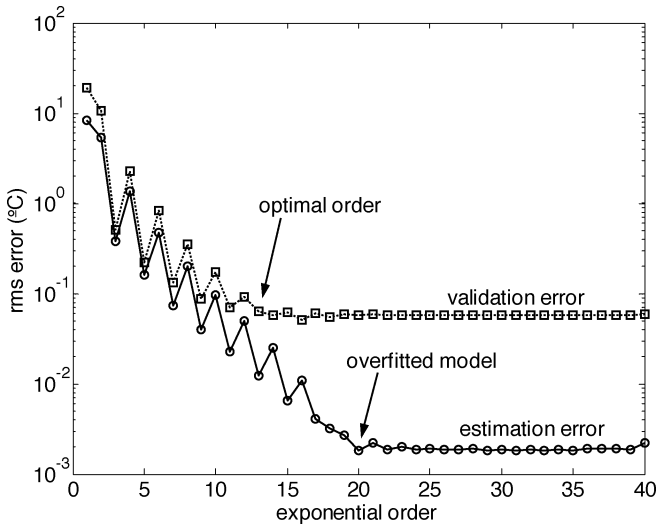

Fig. 9. Evolution of the reconstruction error obtained with ESM for the diagonal terms.

A second approach to the analysis of the transient thermal impedances is to use a semiparametric approach known as ESM already described in Section II. This analysis has been performed using ten exponential terms per decade although even higher densities are possible. As an example, Fig. 8(a) shows the resultant time constant spectrum for a self-impedance transient: $T_{33}(t)$. The residuals are very small and the algorithm automatically sets to zero most of the potential contributions. Only about 11 exponential terms are selected by the algorithm as relevant (amplitudes different from zero).

In Fig. 9, the evolution of the error in estimation and validation is plotted against the total number of exponential terms. Several conclusions can be drawn. In both cases, the error decreases as exponential terms are added to the model. At a certain point, the error saturates. Note that the validation error is much larger than the estimation error. In other words, estimation errors provide overoptimistic results concerning the predictive performance of the model. In addition, the total number of exponentials needed to fit the transients can be much lower than the suggested by the estimation results. The error decrease saturates before in validation than in estimation. In any case, it seems that, concerning predictive accuracy, overfitting is not an issue since

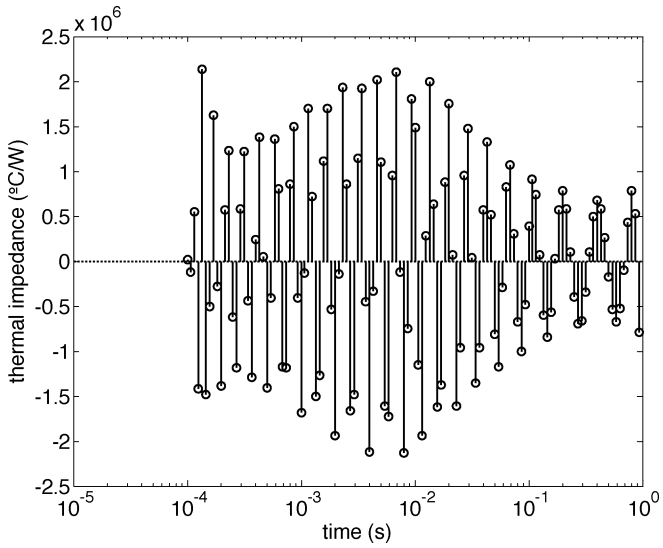

Fig. 10. ESM trans-impedance time-spectrum for $T_{12}(t)$ analysis.

an increase in the validation error is not observed even with high densities of exponential terms.

If we analyze the results for the off-diagonal terms (trans-impedance terms), the ESM provides very low reconstruction errors but at the expense of nonrealistic time constant spectrums with many nonzero terms (see Fig. 10) producing a large thermal model. Due to the nonorthogonality of the signals, the algorithm tends to provide close exponential terms of almost the same amplitude but opposite sign. This result has no physical sense, and it is a consequence of the excessive model order. It can also be observed in Table II that the validation error is much higher than the estimation error (a factor 200 for self-impedance transients and a factor 45 for transimpedance transients). However, in this particular case, the errors are still contained because the estimation errors are very low. This can be explained because the initial transients originated from physical simulation and, in consequence, they only contain numerical errors. From our point of view, this huge increase of errors between validation and estimation can be a determinant factor in the case of empirical transients where the noise levels can be higher. In summary, ESM provides good results for self-impedance transients using the nonnegative least squares algorithm. Its use for trans-impedance transients is not recommended.

The full parametric solution (free nonlinear least squares) to this problem requires that the user first selects the order of the model for each transient. Information from deconvolution may be useful at this step, because we may have an idea about where (in the time scale) the amplitudes concentrate. Otherwise, a scan in the order of the model is performed, and the selected model order corresponds to the knee in the mean square error versus order plot. We have to remark that the least squares problem becomes fully nonlinear with the problems already mentioned in Section II.

At this point, two options arise. First is to use fully free models; that is, there is no restriction on the time constants that appear in the model. All the transients in the matrix are considered as independent signals. The second option is to

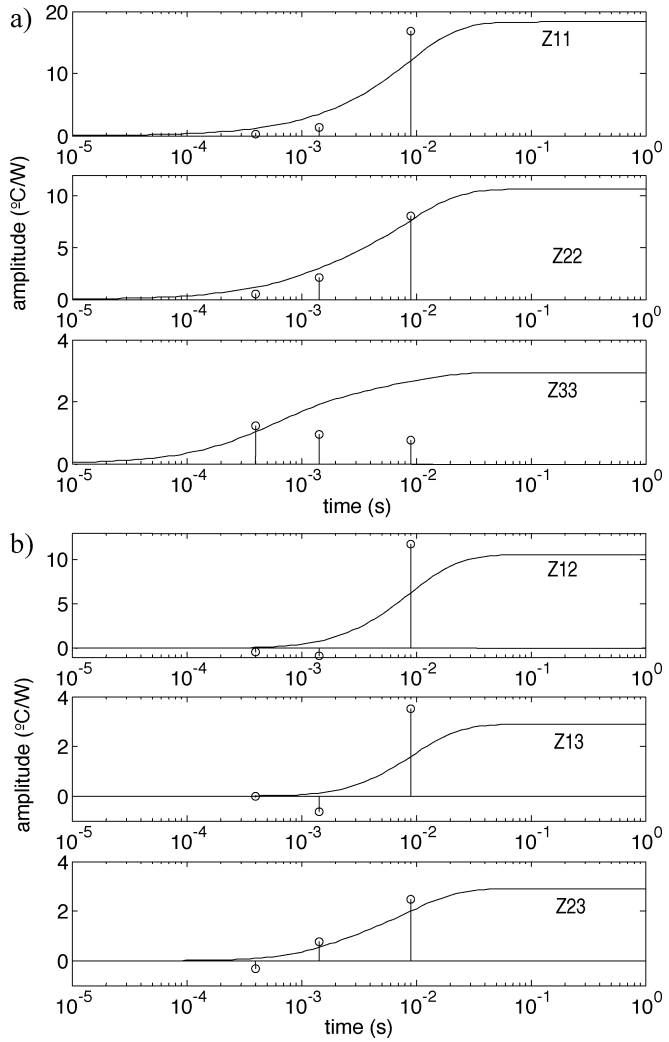

Fig. 11. (a) Discrete time-spectrum from constrained NLSQ for transients $T_{11}(t)$, $T_{22}(t)$, and $T_{33}(t)$. (b) Discrete time-spectrum from constrained NLSQ for transients $T_{12}(t)$, $T_{13}(t)$, and $T_{23}(t)$.

restrict the feasible time constants to a common set. This requirement comes from the representation of the system by a single, although unknown, *RC* network. We have followed both approaches and compared the solution.

In the first option, the best results are obtained using four exponential terms per transient to a total of 48 free parameters. Some examples of the obtained time spectrums can be observed in Fig. 8(a) and 8(b). In the second option, best results are obtained for three shared exponential terms to a total of 21 free parameters [Fig. 11(a) and 11(b)]. The mean square error in estimation is smaller in the first case than in the second. This reduced error variance is a direct consequence of the presence of a larger number of parameters. As an illustration, the final thermal impedance matrix for this case is listed in, as shown in (8) at the bottom of page.

Table II summarizes the reconstruction errors for the different transients. As it can be observed, the worst errors in validation and in estimation transients come from the Jansson deconvolution. This is something that could be expected in the sense that Jansson deconvolution is a nonparametric approach to the time constant spectrum estimation that does not try to minimize mean square errors.

On the other hand, ESM and parametric approximation show good agreement with deconvolution, but the fitting residuals are much better. From the comparison, it can be observed that nonparametric analysis (METS +Jansson) tend to spread the discrete exponential terms to peaks of finite width. In other words, nonparametric techniques show, in general, finite resolution power (see [23] for more details).

As it can be observed, the best results in estimation are obtained by ESM. The results of free nonlinear least squares also give good results. Diagonal transients present a slightly higher error than off-diagonal, due to the constraint on permitting only positive amplitudes. The smallest model comes from the proposed constrained NLSQ method that presents validation errors in the range of tens of mK, only slightly higher than those provided by the free NLSQ.

VI. Conclusion

Dynamic compact thermal models have been obtained for an ultrathin chip stacking technology where several chips can dissipate heat simultaneously. A multiport dynamic model has been obtained and the model has been implemented as a thermal impedance matrix. Different techniques for multiexponential signal analysis have been reviewed and compared. Jansson deconvolution provides a nonparametric time constant spectrum, but the prediction accuracy is limited. The semiparametric exponential series method provides the lowest errors in the reconstruction of the transients. However, it produces noninterpretable time constant spectrums when analyzing the transimpedance transients.

Concerning parametric approximations, the simplest model (in terms of free parameters) comes from the proposed constrained nonlinear least squares that also show good prediction performance. Free nonlinear least squares models can provide slightly better results if the model order is selected properly.

Although the final comparison among the different models can be based on the prediction accuracy in validation with arbitrary power signals, it is also important to consider the physical interpretability of the obtained time constant distribution.

$$\frac{1}{5.19\cdot 10^{-9}s^3+1.71\cdot 10^{-5}s^2+0.01s+1}\times\begin{bmatrix} 1.68\cdot 10^{-5}s^2+4.49\cdot 10^{-2}s+18.28 & -9.06\cdot 10^{-7}s^2+9.97\cdot 10^{-3}s+10.56 & -1.64\cdot 10^{-7}s^2+5.73\cdot 10^{-4}s+2.88 \\ -9.06\cdot 10^{-7}s^2+9.97\cdot 10^{-3}s+10.56 & 1.84\cdot 10^{-5}s^2+3.94\cdot 10^{-2}s+10.61 & 8.18\cdot 10^{-8}s^2+8.38\cdot 10^{-3}s+2.90 \\ -1.64\cdot 10^{-7}s^2+5.73\cdot 10^{-4}s+2.88 & 8.18\cdot 10^{-8}s^2+8.38\cdot 10^{-3}s+2.90 & 1.96\cdot 10^{-5}s^2+2.31\cdot 10^{-2}s+2.94 \end{bmatrix} \quad (8)$$

REFERENCES

[1] K. Skadron, M. R. Stan, W. Huang, Z. Lu, K. Sankaranarayanan, and J. Lach, "A computer-architecture approach to thermal management in computer systems: opportunities and challenges," in *Proc. EUROSIME*, 2004, pp. 415–422.

[2] C. S. Yun, P. Maloberti, M. Ciappa, and W. Fichtner, "Thermal component model for electrothermal analysis of IGBT module systems," *IEEE Trans. Adv. Packag.*, vol. 24, no. 3, pp. 401–406, Aug. 2001.

[3] T. Hauck and C. Bohm, "Thermal RC-network approach to analyze multichip power packages," in *Proc. 16th Annu. IEEE Semiconductor Thermal Measurement Management Symp., SEMI-THERM XVI*, 2000, pp. 227–234.

[4] M. Ishizuka and Y. Fukuoka, "Application of the thermal network method to the transient thermal analysis of multichip modules," in *Proc. 2nd IEMT/IMC Symp.*, 1998, pp. 161–166.

[5] M. Carmona, S. Marco, J. Samitier, M. C. Acero, J. A. Plaza, and J. Esteve, "Modeling the thermal actuation in a thermo-pneumatic micropump," *J. Electron. Packag.*, vol. 125, pp. 527–530, 2003.

[6] P. W. Barth, "Silicon microvalves for gas flow control," in *Proc. Transducers*, Stockholm, Sweden, Jun. 1995, pp. 276–279.

[7] P. Krulevitch, A. P. Lee, P. B. Ramsey, J. C. Trevino, J. Hamiltron, and M. A. Northrup, "Thin film shape memory alloy microactuators," *IEEE-ASME J. Microelectromech. Syst.*, vol. 5, no. 4, pp. 270–282, Dec. 1996.

[8] Q. Zou, U. Sridhar, and R. Lin, "A study on micromachined bimetallic actuation," *Sens. Actuators A*, vol. 78, pp. 212–219, 1999.

[9] J. Palacin, M. Salleras, M. Puig, J. Samitier, and S. Marco, "Evolutionary algorithms for compact thermal modeling of microsystems: application to a micro-pyrotechnic actuator," *J. Micromech. Microeng.*, vol. 14, pp. 1074–1082, 2004.

[10] C. Rembe, S. aus der Wiesche, and E. P. Hofer, "Thermal ink jet dynamics: modeling, simulation and testing," *Microelectron. Reliab.*, vol. 40, pp. 525–532, 2000.

[11] J. Puigcorbé, A. Vila, J. Cerda, A. Cirera, I. Gracia, C. Cane, and J. R. Morante, "Thermo-mechanical analysis of micro-drop coated gas sensor," *Sens. Actuators A*, vol. 97–98, pp. 379–385, 2002.

[12] M. N. Sabry, "Dynamic compact thermal models: an overview of current and potential advances," in *Int. Workshop Thermal Investigations ICs Systems*, Madrid, Spain, Oct. 1–4, 2002, p. 229.

[13] S. Pinel, J. Tasselli, F. Lepinois, A. Marty, J. P. Bailbé , E. Beyne, R. Van Hoof, O. Vendier, M. Huan, S. Marco, and J. R. Morante, "Ultra thin chip vertical integration technique," in *Proc. 13th Eur. IMAPS Conf.*, Strasbourg, France, 2001, pp. 299–302.

[14] O. Vendier, M. Huan, C. Drevofi, J. L. Cazaux, E. Beyne, R. Van Hoof, A. Marty, S. Pinel, J. Tasselli, S. Marco, and J. R. Morante, "Ultra thin electronics for space applications," in *Proc. 51st Electronic Components Technology Conf.*, 2001, pp. 767–771.

[15] S. Pinel, J. Tasselli, J. P. Bailbe, A. Marty, O. Vendier, M. Huan, S. Marco, J. R. Morante, E. Beyne, and R. Van Hoof, "Thermal management in a new ultrathin chip stack technology," in *Proc. EuroSimE*, Paris, France, Apr. 2001, pp. 269–278.

[16] A. Bar-Cohen and W. B. Krueger, "Thermal characterization of chip packages-evolutionary development of compact models," *IEEE Trans. Compon. Packag. Manuf. Technol. A*, vol. 20, no. 4, pp. 399–410, Dec. 1997.

[17] V. Szekely, "Identification of *RC* networks by deconvolution: chances and limits," *IEEE Trans. Circuits Syst I*, vol. 45, no. 3, pp. 244–258, Mar. 1998.

[18] A. Istratov and O. Vyvenko, "Exponential analysis in physical phenomena," *Rev. Sci. Instrum.*, vol. 70, no. 2, pp. 1233–1257, 1999.

[19] L. Ljung, *System Identification-Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[20] D. G. Gardner, J. C. Gardner, G. Laush, and W. W. Meinke, "Method for the analysis of multiexponential decay curves," *J. Chem. Phys.*, vol. 31, no. 4, pp. 978–986, 1959.

[21] S. Marco, J. Samitier, and J. R. Morante, "A novel time-domain method to analyze multicomponent exponential transients," *Meas. Sci. Technol.*, vol. 6, no. 2, pp. 135–142, 1995.

[22] M. Carmona, S. Marco, J. Palacin, and J. Samitier, "A time-domain method for the analysis of thermal impedance response preserving the convolution form," *IEEE Trans. Compon. Packag. Technol.*, vol. 22, no. 2, pp. 238–244, Jun. 1999.

[23] S. Marco, J. Palacín, and J. Samitier, "Improved multiexponential transient spectroscopy by iterative deconvolution," *IEEE Trans. Instrum. Meas.*, vol. 50, no. 3, pp. 774–780, Jun. 2001.

[24] A. Siemiarczuk, B. D. Wagner, and W. R. Ware, "Comparison of the maximum entropy and the exponential series method for the recovery of distributions of lifetimes from fluorescence lifetime data," *J. Phys. Chem.*, vol. 94, pp. 1661–1666, 1990.

[25] L. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. Englewood Cliffs, NJ: Prentice-Hall, 1974, pp. 160–161.

[26] J. Palacin, S. Marco, and J. Samitier, "Suboptimal filtering and nonlinear time scale transformation for the analysis of multiexponential decays," *IEEE Trans. Instrum. Meas*, vol. 50, no. 1, pp. 135–140, Feb. 2001.

[27] M. Rencz and V. Székely, "Dynamic thermal multiport modeling of IC packages," *IEEE Trans. Compon. Packag. Manuf. Technol. A*, vol. 24, no. 4, pp. 596–604, Dec. 2001.

[28] F. Christiaens, B. Vandevlede, 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. Manufact. Technol. A.*, vol. 21, no. 4, pp. 565–576, Dec. 1998.

[29] S. Pinel, J. Tasselli, J. P. Bailbé, A. Marty, P. Puech, and D. Estève, "Mechanical lapping, handling and transfer of ultrathin wafers," *J. Micromech. Microeng.*, vol. 8, pp. 338–342, 1998.

[30] E. Jung, A. Eumann, D. Wojakowski, A. Ostmann, C. Landesberger, R. Aschenbrenner, and H. Reichl, "Ultra thin chips for miniaturized products," in *Proc. 52nd Electronic Components Technology Conf.*, 2002, pp. 1110–1113.

[31] S. Leseduarte, S. Marco, E. Beyne, R. Van-Hoof, A. Marty, S. Pinel, O. Vendier, and A. Coello-Vera, "Residual thermo-mechanical stresses in thinned-chip assemblies," *IEEE Trans. Compon. Packag. Manufact. Technol. A*, vol. 23, no. 4, pp. 673–679, Dec. 2001.

[32] E. Beyne, S. Pinel, O. Vendier, A. Coello-Vera, and J. Tasselli, "Method of Transferring Ultrathin Substrates and Application of the Method to the Manufacture of a Multilayer Thin Film Device," Eur. Patent EP1041620, Oct. 4, 2000.

Jordi Palacín received the B.S. degree in electronics from the Universitat de Barcelona, Barcelona, Spain, in 1997.

He has been an Associate Professor in the Departament d'Informatica i Enginyeria Industrial, Universitat de Lleida, Lleida, Spain, since 1997. His research interests include signal processing, sensors, and robotics.

Marc Salleras received the physics and the electronic engineering degrees from the University of Barcelona, Barcelona, Spain, in 2000 and 2003, respectively, and the M.S. degree in numerical methods in engineering from the Polytechnic University of Barcelona in 2003. He is currently pursuing the Ph.D. degree at the University of Barcelona in the Electronics Department. His research interests include thermal simulation and modeling.

Josep Samitier is a Full Professor in the Electronics Department, University of Barcelona, Barcelona, Spain, and he has been Director of the research group "Instrumentation and Communication Systems" (SIC) since February 1995. Since March 2001, he has been a Director of the Electronics Department and Deputy Head of the Barcelona Science Park, University of Barcelona. He has participated in several European projects concerning microsystems and nanotechnology.

Santiago Marco received the degree in physics in 1988 and the Ph.D. (honor award) degree from the Universitat de Barcelona, Barcelona, Spain, in 1998 and 1993, respectively.

He has been an Associate Professor in the Departament d'Electronica, Universitat de Barcelona, since 1995. From 1990 to 1993, he was regular visitor of the Centro Nacional de Microelectrònica, Bellaterra, Spain. In 1994, he was a Postdoctoral Researcher in the Department of Electronic Engineering, Universita di Roma "Tor Vergata," working in data processing for chemical sensors. He has published about 50 papers in scientific journals and books, as well as more than 100 conference papers. His current research interests are in chemical instrumentation based on intelligent signal processing and microsystem modeling.