Available online at www.sciencedirect.com
SCIENCE DIRECT®
International Journal of Heat and Mass Transfer 47 (2004) 3707–3720
International Journal of HEAT and MASS TRANSFER
www.elsevier.com/locate/ijhmt

A transient self-adaptive technique for modeling thermal problems with large variations in physical scales

James S. Wilson [a], Peter E. Raad [b,*]

[a] Raytheon Electronic Systems, 13510 North Central Expressway, M/S 400, Dallas, TX 75243, USA
[b] Mechanical Engineering Department, Southern Methodist University, Dallas, TX 75275-0337, USA

Received 17 January 2003; received in revised form 29 March 2004
Available online 8 May 2004

Abstract

The concurrent electro-thermal design of three-dimensional integrated circuits characterized by submicron geometric features requires thermal modeling that can comprehend geometric complexities, multiple materials, temperature-dependent material properties, and multiple spatial and temporal scales. The computational time required for a full-scale transient simulation with traditional discretization schemes far exceeds what is practical for concurrent design practices. A new computational paradigm for a transient, multiple-grid, solution technique has been developed, which adaptively handles the wide ranges of spatial and temporal scales associated with the thermal modeling of high-performance integrated circuits (ICs). As the grid is automatically refined over selected regions of the computational domain, the solution becomes invariant to further reductions in grid spacing and time step size. The use of this self-adaptive approach reduces the computational requirements for transient thermal modeling by over two orders of magnitude, making it possible for the first time to simultaneously perform both the electrical and thermal analysis and design of real ICs.

Keywords: Multi-scale physics; Self-adaptive numerical simulation; Transient computational heat transfer

1. Introduction

Integrated circuits (ICs) are complicated three-dimensional devices constructed of multiple layers of materials whose dimensions vary widely and whose thermal properties can have strong temperature dependence. These layers are fabricated by deposition and selective removal of various materials. The dimensions of an electrically active region (e.g., gate) are typically in fractions of a micrometer while the thickness of circuit topology features (e.g., lines, metalization, pads) can vary from a few tens to a few thousands of angstroms. These features are small compared to the overall dimensions of the IC, which are typically measured in millimeters. The thermal characterization of such devices thus requires that the method used be capable of handling these complexities and temperature dependencies. Furthermore, the simulation tool must also be able to handle the large variations in both spatial and temporal scales normally associated with ICs. Fig. 1 illustrates typical packaged microwave modules that encompass the microwave and digital circuitry required for transmission (color version of figures may be viewed at http://engr.smu.edu/setsl/ijhmt). The top left portion of Fig. 1(a) shows the gallium arsenide (GaAs) and silicon IC devices along with various layers of die attachment and radio frequency (RF) circuitry. The top right portion illustrates a top view of a GaAs monolithic microwave integrated circuit (MMIC), while the bottom right portion details a single field effect transistor (FET)

* Corresponding author. Tel.: +1-214-768-3043; fax: +1-214-768-9900.
E-mail addresses: jsw@raytheon.com (J.S. Wilson), praad@smu.edu (P.E. Raad).

0017-9310/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijheatmasstransfer.2004.03.011

Nomenclature

C_p	heat capacity
C_T	total thermal capacitance of a cell
G	conductance between two adjacent nodes
k	thermal conductivity
L_x, L_y	overall problem dimensions
Q_x, Q_y	heat source dimensions
$\dot{s}$	heat source
S	total heat input within a cell
t	time

Greek symbols

δ_{xx}, δ_{yy}, δ_{zz}	second-order spatial operators
Δt	time increment
Δx, Δy, Δz	spatial increments
ρ	material density

Superscripts

$*, **$	intermediate solution levels
n, $n+1$	current and new time levels, respectively

on the MMIC. The FETs contain several similar channels, each of which contains contacts for the gate, source, and drain. The lower left portion of Fig. 1(a) shows a tunneling electron microscope picture detailing the gate metallization in a cross-sectional view. The spatial scale range in transitioning from modules to detailed gate metal is about five orders of magnitude. This spatial scale range may be even more dramatic when consideration is given to antennas made up of thousands of these modules. Antenna dimensions measured in meters are not uncommon. A similar issue arises in the consideration of the appropriate temporal scales. A module and its associated thermal boundary conditions are typically described in units of minutes or seconds, while temperature changes in the gate region of a GaAs IC must be described in time scales of fractions of a microsecond.

Microwave devices are often operated in a pulsed mode, where the device alternates between on- and off-states. The duration of the on-state typically ranges from approximately 1 μs to 1 ms. The small physical dimensions of the device channel (or junction) allow its temperature to vary significantly during this time period. Circuit designers need temperature predictions for both steady-state and transient conditions because many of the design parameters are sensitive to temperature. For example, device efficiency, gain, and power output for an amplifier all decrease with temperature [1]. The need for temperature predictions has resulted in the development of a number of methods, some of which are described in the next section on prior work. The development trend appears to have been driven by solution methodologies rather than by the physics of the real problem, as it should be. The distinguishing feature of the method described in this paper is that, unlike the approximate methods in previous works, it was developed with the *real problem* in mind.

Fig. 2 illustrates the tradeoffs associated with the use of numerical methods for design efforts. The y-axis represents the time requirement for model generation and the solution of the resulting model. The x-axis represents increasing complexity. Closed form solutions are very quick to set up and solve, but the answers are of limited use. Adding more details to the problem provides answers that are now useful for design studies, but at a high cost of user effort and computer time. The objective of the development effort described in this paper was to move both the model generation time and the solution time into the shaded region in the lower right-hand corner of Fig. 2. This was accomplished by designing a solution method that addresses the full problem as opposed to developing a method and then finding what class of problems it could solve.

2. Prior work

The need for temperature predictions by IC designers has resulted in several estimation methods ranging from analytical approximations to full three-dimensional numerical simulations. Analytical approximations [2,3] are restricted to simple geometries, but offer moderately fast simulation times. The transient capability of such methods is also extremely limited [4]. An additional limitation is the requirement for constant thermal properties (except for applying a Kirchhoff transform with constant temperature base of the semiconductor). Solutions to more realistic problems may be obtained by numerical modeling, which is supported by several methodologies. These include finite difference, finite element, and boundary element methods, which are available in the form of commercial software packages [5,6]. However, the computational time requirements of detailed simulations prohibit concurrent engineering design. These packages also must rely heavily on the experience of the analyst since three-dimensional grid convergence studies are both difficult to perform and totally unreasonable from the standpoint of solution time. Also, the addition of temperature-dependent properties requires the use of an iterative procedure. Dawson [7] gives a more extensive list of approximation methods for steady-state problems, but does not list any

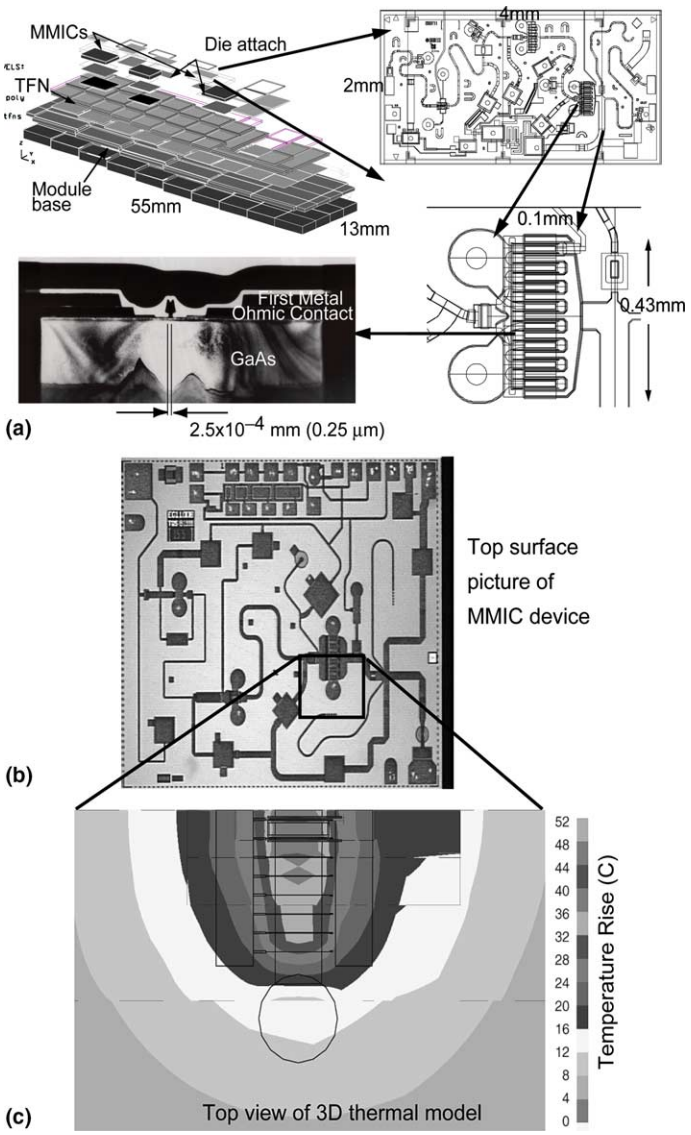


Fig. 1. Typical MMICs: (a) range of spatial scales, (b) top view, (c) thermal profile (see color version at http://engr.smu.edu/setsl/ijhmt).

transient solution methods. Because of the excessive time required for modeling transient three-dimensional problems, the majority of the applicable published literature has been limited to either two-dimensional problems with complex geometry, or simplified three-dimensional problems. In addition, the published transient analyses have been limited to problems in which the thermal material properties are constant.

Problems such as the ones mentioned above include geometrical complexity and significant variations in spatial scales. The spatial complexity can be addressed by the use of unstructured grids, which allow for higher mesh resolution in regions where the geometry is complex as well as in regions where strong solution gradients are anticipated. The main disadvantage to unstructured grids, however, is the significant computational effort required to solve the resulting matrix, especially in transient simulations.

When considering computational efficiency, the preferable approach is to solve the problem with a manageable grid density and then ascertain the portions of the domain requiring additional refinement. This approach was successfully introduced by Berger [8] and Berger and Oliger [9] for the solution of hyperbolic problems, and later extended to problems in fluid dynamics by van der Wijngaart [10]. Extensions to these

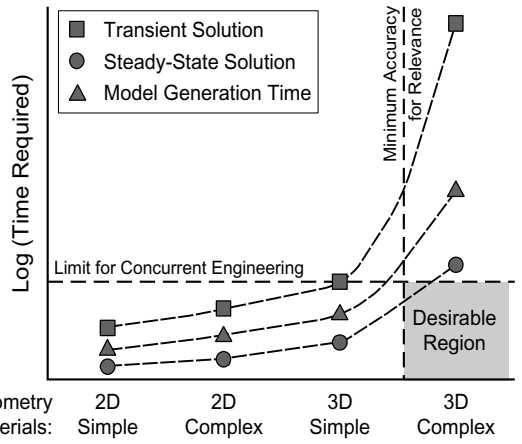


Fig. 2. Relevant simulations must comprehend temperature-dependent properties and complex geometries.

earlier works have continued to appear in the literature, but have been limited to either two-dimensional domains [11] or have limited the adaptive meshing to space only [12]. More recent representative works include those of Cook [13] who demonstrated that the mesh adaptation errors associated with the boundaries are small and localized. Roma et al. [14] applied the adaptive implementation to immersed boundaries, but did not adaptively refine the time scales. Powell et al. [15] demonstrated the adaptive approach on magnetohydrodynamics problems with a two-step scheme in time and used fixed threshold error criteria to guide their grid adaptation.

Extending the self-adaptive concepts to the solution of transient three-dimensional problems whose spatial and temporal scales vary over *many orders of magnitude* is the subject of this work. The novel technique presented herein successfully reduces the computational time requirements to the point that concurrent electro-thermal design becomes feasible for the first time.

3. Motivating example problem

Consider the top surface view of a GaAs power amplifier MMIC illustrated in Fig. 1(b). This device contains two FETs. Each FET contains an array of gate fingers with the sum of the gate fingers comprising the FET junction. Source and drain connections are located on either side of the gate fingers. The gate connection is made with a metal trace directly above the active region and silicon nitride passivation exists in the region between the gate and source (or drain) pads. The device is mounted to a copper–molybdenum heat sink with gold–tin (80/20) solder. Heat generation occurs in the semiconductor layer of the GaAs and must dissipate by conduction down through the GaAs, die attach, and heat sink. The surface metal and passivation layers provide some heat spreading and must be included for an accurate model. The physical geometry is inherently three-dimensional with four to five orders of magnitude changes in spatial scales and the added complexity of temperature-dependent thermal conductivity.

The temperature field for one of the FETs was obtained by the use of an electrical network analogy solution method [16] that involved a biased meshing scheme to concentrate the mesh density around one of the device fingers anticipated to be the hottest during operation. Initial model sizes were in the range of 80,000 nodes and over 140,000 resistors, but after expending considerable effort to minimize the number of nodes without changing the results (i.e., testing for grid convergence), the model comprised about 32,000 nodes and 87,000 resistors. The grid generation time required about two days of effort on the part of an experienced analyst, which in and of itself impedes prototyping.

Fig. 1(c) shows a contour plot of the predicted steady-state temperature field on the top surface. While a steady-state solution required 32 min of CPU time (Sun 275 MHz Sparc Ultra), transient solutions require excessive computational time. The small grid spacing required to resolve the FET channels results in an excessively small time step if an explicit time-marching method is to be used. Implicit methods relax the stability requirement, but the nonlinear thermal conductivity requires continual recalculation of the resistor matrix. An implicit transient solution was obtained for one period of a pulsed operating scenario with a variable time step size selection routine. However, the time required to obtain this solution was well over 8400 min of CPU time (Sun 275 MHz Sparc Ultra). Normally, a simulation spanning several pulses is required in order to ensure that a representative solution has been obtained. In addition, a grid convergence in time is also necessary to ensure temporal accuracy. As a result, even though the network analogy method is efficient compared to FEA tools, the solution time requirements prohibit any interactive design process.

4. Motivation for a new paradigm

A new solution technique has been developed with the primary goal of decreasing the solution time as well as alleviating the effort required of an analyst in both model generation and grid convergence decisions. The new solution approach has two distinguishing features that make it novel. The first feature is the ability of the method to handle problem geometries that are characterized by multiple materials of arbitrary sizes and locations. Any attempt to try to resolve the smallest geometric features with a single computational mesh will

result in a mesh density that is prohibitively large, especially in three spatial dimensions. An additional drawback is that the geometric complexity, and *not* the temperature gradients, would dictate the mesh density.

The second distinguishing feature is that the new solution technique uses multiple grids in time, which allows the use of different time step criteria for each subregion. The time step selection is chosen to be commensurate with the local physics, and in doing so, avoids expending computational effort on trying to resolve in time what cannot yet be resolved in space. Indeed, solutions in coarser regions serve only to establish time-dependent boundary conditions for the (more refined) subregions. Since the spatial resolution is low in coarser regions, there is no advantage to the use of high temporal resolutions there. Consequently, larger time steps can be used in the coarser domains, resulting in significant savings in computational effort. The computational time for transient problems is further reduced by the use of a locally one-dimensional (or factored-implicit) scheme. The solution method for rapidly solving the transient portion of the problem is the subject of a United States Patent [17].

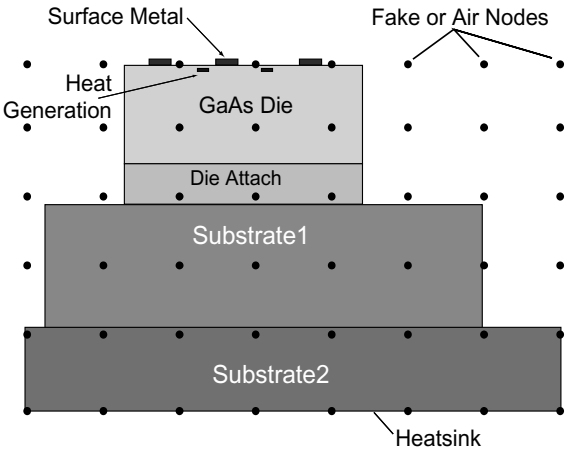
5. Governing equations, typical geometries, and boundary conditions

The Fourier heat equation that describes the transient thermal behavior of microwave integrated circuits is

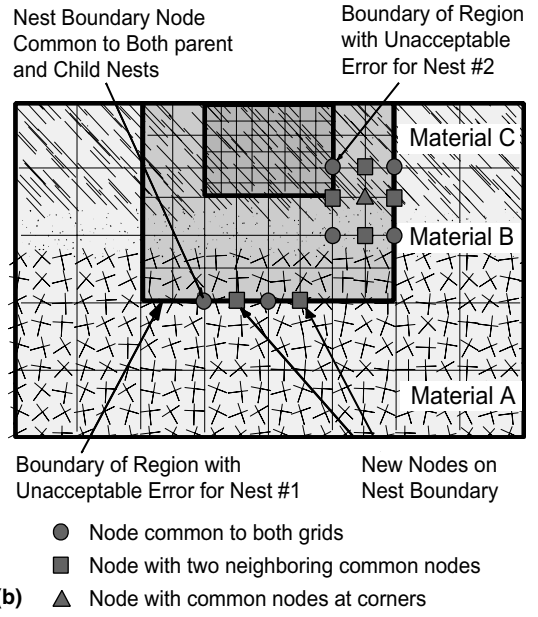
$$\nabla(k\nabla T) + \dot{s} = \rho C_{\mathrm{p}} \frac{\partial T}{\partial t} \tag{1}$$

The material properties and the heat source ($\dot{s}$) are functions of both space and temperature. The source term is also a function of time. While the heat equation is very well known, the new concepts introduced in this work relate to the application of this equation to problems with large variations in spatial and temporal scales, multiple materials, and temperature-dependent properties. A full transient simulation of a realistic IC is impractical with conventional techniques.

A schematic representation of the geometry of a typical cross-section through an IC assembly is shown in Fig. 3(a) (the uniformly spaced nodes shown will be addressed below). As depicted, the second substrate is mounted to the heat sink, and the problem may be described as a series of stacked materials. A typical boundary condition for this type of problem is a known isothermal surface at some level below the GaAs (i.e., at the heat sink). Because of the small size of the heat sources, the temperature field near the top surface exhibits high gradients. An adiabatic boundary condition is specified for all other exposed surfaces since for these types of problems, radiation and convection effects are negligible when compared with conduction.



(a) Arbitrary Crude Mesh: Node Points Indicated by Dots



(b)

Fig. 3. Multiple materials and nests: (a) geometry and initial mesh, (b) nest boundary nodes over three-material problem.

6. Methodology

The modeling approach developed in this work is based on a finite volume approach where the temperature field is obtained at discrete locations. Approximations representing the spatial derivatives may be derived from either a control volume (CV) or a Taylor-series expansion approach. Both approaches have advantages and disadvantages with respect to the derivation and solution of the discretized equations.

A discretization resulting from a Taylor-series expansion (i.e., finite difference) lends itself naturally to

the use of computationally efficient, factored solvers in multi-dimensional problems. This feature is highly desirable for computational efficiency, especially in transient problems where the system of equations must be solved repeatedly to advance the solution. However, the Taylor-series approach is not well suited for handling multiple materials. Also, nodes located on or near physical boundaries require special treatment.

The CV approach, on the other hand, is more closely tied to the physics of these types of problems. As such, the treatment of multiple materials between computational nodes can be handled in a natural and straightforward fashion. The CV approach also provides more flexibility when dealing with computational nodes that are not aligned with physical boundaries. However, the correct application of a factored (or locally one-dimensional) approach in the context of a control volume discretization is not obvious. Consequently, the implementation of a factored-implicit approach along with a control volume discretization was one of the important accomplishments of the current development. The solution methodology involves two steps, namely the generation of multiple domains (and their associated grids), and the solution of the physics in each resulting grid.

6.1. Subdividing the computational domain

Since the physical dimensions of the various materials used in high-performance analog electronic devices vary greatly (μs to cm), a uniform mesh that resolves all of the details in three dimensions would result in an excessive number of nodes. A common method for dealing with scale variations is to skew (or bias) the mesh in order to concentrate more nodes in areas where a higher resolution level is required. The shortcoming of using a biased-mesh approach to resolve the geometry is that the problem geometry, and not the temperature gradients (or physics), will end up dictating the meshing.

The meshing strategy used in the development of the current technique was designed to ensure that the method is (i) automatic and adaptive, (ii) independent of user expertise, and (iii) independent of materials, geometry features, and potential locations of sources.

An additional objective in designing a meshing approach was to eliminate as much "engineering judgment" as possible in deciding how much or how little of the problem geometry should be considered. The described approach allows the user to include the difficult problem geometry associated with the top surface features of an IC over an area larger than would be anticipated to influence the zone(s) of interest, and then to allow the embedded error prediction technique to determine which regions need to be further refined. The strength (and novelty) of the current method is that it uses effective thermal properties that are consistent with the local grid spacing at the particular grid level being used.

An illustration of an arbitrary (and crude) mesh is shown in Fig. 3(a). It is noted that the nodes are not aligned with the physical geometry. The fake (or air) node terminology refers to nodes that are either in a location occupied by air or outside the physical domain, as will be described in more detail later. Upon obtaining a solution to the arbitrary mesh and making an estimate of the error in the solution, the predicted error may be used to refine the mesh in regions where both a detailed solution is desired and the predicted error is too large. Given the need to quickly define regions of higher-than-acceptable error and interpolate values for new points, a uniform mesh is highly desirable and was therefore selected. A mesh becomes a parent mesh as soon as part of its volume is flagged as needing further refinement. The areas requiring refinement become child meshes. A child mesh will have in common with its parent mesh at least every other point in each of the three dimensions.

6.1.1. Error estimate

The meshing strategy begins by placing nodes at the global extents of the problem geometry. A fundamental aspect of starting with an arbitrary number of nodes to model a problem is the use of effective (or smeared) properties between control volumes. The nesting approach uses an estimate of the error in the steady-state solution to define one or more regions that require additional refinement. The knowledge that absolute errors in temperature will be located within the source region(s) where the peak temperatures occur provides guidance in establishing boundaries for child grids. The newly created child regions requiring further refinement are then solved and themselves searched for internal regions of unacceptable error levels. This zooming process continues until a convergence criterion is met. The convergence criterion involves a check on the error estimate as well as a check that the grid size is actually small enough to resolve the source (or sources). For convenience and computational efficiency in solving the transient portion of the problem with a factored-implicit scheme, the child grid regions are selected to be rectangular (or parallelepiped in three dimensions). Starting with an arbitrary odd number of nodes for an initial mesh, a coarser grid made up of every other node is created for the purpose of comparing the two grids. A solution is obtained on each of the two grids, making it possible to estimate the solution error simply by comparing the two solutions at the common nodes between the fine and coarse grids. It is tempting to try to improve the solution by adding the error computed from a Richardson extrapolation to the finer solution, but this approach is beneficial only for the first nest where the boundary nodes do not contain errors.

6.1.2. Define nesting templates

The error estimate is then used to define one or more smaller regions, which, for convenience, are hereafter referred to as *child regions*. Grids used to discretize these child regions will be referred to as *child grids* (or *child meshes*). The consecutively refined grid levels will be referred to as *nests*. The grid spacing for each child mesh is decreased and solved with fixed boundary values around its outer surface where the boundary values are inherited from the parent grid. After solving over the child grid, a comparison is made between the common points of the parent and child grids and additional child grids are created as necessary until the solution on the most recent (i.e., finest) child grid satisfies the user-prescribed error criteria.

It is advantageous to define which locations are of interest to the user. A typical high-performance integrated circuit may have many heat sources, but the user may be interested in highly accurate results (e.g., temperature) at only one or two sources. For this reason, only the locations of interest need to be resolved. A parent nest may contain several regions that need further refinement, but as long as child-nest boundaries with error predictions satisfying an established criterion may be constructed around a location of interest, the error prediction at a point outside of the child nest is deemed not important. Applying this principle to problems with multiple sources yields independently solvable problems when two or more distinct child nests are created. This is also a highly attractive feature when considering a parallel implementation of the described methodology.

On each grid level, the nodes that define the boundary for child nests form a *nest template*. To solve the transient problem, the initial grid is solved in time and the boundary values of the nest templates are saved as a function of time. Then, using the boundary values saved from the transient solution on its parent grid, a child nest is solved in time. A principal and significant advantage to this approach is that the denser grid spacing is confined to the child nests, with each child nest refining a smaller region of the problem than the parent nest. A grid refinement strategy of increasing the resolution by an even factor of 2 was chosen.

The overall solution for a particular problem is a special union of all solutions obtained on the nested grids, where the solution from each child grid overwrites the corresponding solution on its parent's grid. It is important to note that once the overall solution is achieved, the solutions on portions of the parent grids that are not overwritten already satisfy the desired error criterion.

6.1.3. Interpolating values for new boundary nodes

The concept of dividing the problem domain into nested levels is illustrated in Fig. 3(b). An important feature of the new method is that the grid lines need not necessarily line up with the material interfaces. Some of the nodes on a nest boundary are common to both the parent and child grids. For these points, the child-node inherits the solution value of the parent-node. The nodes that are new, as a result of refining the mesh, require some form of interpolation. In keeping with the spirit of simplicity, linear interpolation (weighted by thermal conductivity) of the steady-state values of new nodes was investigated first. Higher-order interpolation was also used to determine whether the overall solution time could be reduced. The solution time would only be reduced if the high-order interpolation allowed the creation of smaller child nests.

In the case of three-dimensional nests, nest boundaries become surfaces. Values for some newly created nodes must be interpolated based on information from neighboring nodes as illustrated in Fig. 3(b). Values for nodes that have common nodes on either side (represented by squares) are interpolated based on a weighted-average of the thermal conductance values of the common nodes (represented by circles). Node values in the center of an area (represented by triangles) formed by common nodes are interpolated based on a weighted-average of the thermal conductance values of the four common nodes. As the geometric features are resolved, new nodes will eventually need to be created between a common node and an air-node. This is especially true when resolving the top surface metalization layer(s). The temperature value of the air-node will not influence the interpolation since the interpolation scheme is weighted by the magnitude of the local thermal conductance and the thermal conductance for an air-node is several orders of magnitude smaller than that for a solid material.

7. Solution methodology

7.1. Factored implicit scheme

The use of a locally one-dimensional (or factored-implicit) scheme is computationally advantageous for the solution of transient partial differential equations [18]. However, two important points must be addressed to correctly apply the implicit scheme with a control volume analysis based on an arbitrary mesh. The first point is that the factored scheme is based on approximating derivatives solely on the basis of spatial coordinates. But, the existence of multiple materials requires the inclusion of material properties in the approximation. Hence, the factorization operators need to be described in terms of heat fluxes rather than temperature gradients. The second point is related to the fact that regions of high errors coincide with regions of high spatial gradients in the solution. The use of factorization, however, introduces a factorization error, which

has local maxima near material interfaces. These regions do not coincide with regions of high spatial gradients, but rather with regions in which the factorization spatial operators do not commute. The additional factorization error then limits the allowable size of the time step.

7.2. Formulation of the factored-implicit equations

A central difference approximation of the second derivative in a spatial direction appears as

$$\delta_{xx}\Delta T_i^n = \frac{1}{\Delta x}\left(\frac{k_{i-1/2}(\Delta T_{i-1/2}^n - \Delta T_i^n)}{\Delta x} + \frac{k_{i+1/2}(\Delta T_{i+1/2}^n - \Delta T_i^n)}{\Delta x}\right) \tag{2}$$

where $\Delta T_i^n \equiv T_i^{n+1} - T_i^n$. The approximation of the heat equation with a source $\dot{s}$ appears as

$$\left[1 - \frac{\Theta\Delta t}{\rho C_p}(\delta_{xx}+\delta_{yy}+\delta_{zz})^{n+1}\right]\Delta T^n = \frac{\Delta t}{\rho C_p}\left[(1-\Theta)(\delta_{xx}+\delta_{yy}+\delta_{zz})^n + \Theta(\delta_{xx}+\delta_{yy}+\delta_{zz})^{n+1}\right]T^n + \frac{\Delta t}{\rho C_p}\left[(1-\Theta)\dot{s}^n + \Theta\dot{s}^{n+1}\right] \tag{3}$$

where Θ specifies the degree of implicitness (e.g., $\Theta = 0.5$ for trapezoidal and $\Theta = 1$ for pure implicit). All terms on the right-hand side are known except for the spatial operators at the $n+1$ time level. If the thermal conductivity is constant, then the spatial operators become independent of time and the right-hand side becomes known and iteration across a time step is not required.

To further simplify the equation and cast it in a form more readily useable for multiple materials and discrete heat locations, Eq. (3) is multiplied and divided by the control volume of a cell ($\Delta x\Delta y\Delta z$) so that properties are not on a per unit volume basis. As a result, new spatial operators L_x, L_y, and L_z are introduced such that

$$L_x\Delta T_i = G_{i-1/2}(\Delta T_{i-1} - \Delta T_i) + G_{i+1/2}(\Delta T_{i+1} - \Delta T_i) \tag{4}$$

Operators L_y and L_z are defined similarly. Using the new operators in Eq. (3) leads to the following form:

$$\left[1 - \frac{\Theta\Delta t}{C_T}(L_x+L_y+L_z)^{n+1}\right]\Delta T^n = \frac{\Delta t}{C_T}\left[(1-\Theta)(L_x+L_y+L_z)^n + \Theta(L_x+L_y+L_z)^{n+1}\right]T^n + \frac{\Delta t}{C_T}(1-\Theta)S^n + \Theta S^{n+1} \tag{5}$$

Eq. (5) may be factored into three one-dimensional equations as follows:

$$\left(1 - \frac{\Theta\Delta t}{C_T}L_x\right)\Delta T^* = \frac{\Theta\Delta t}{C_T}\Phi \tag{6a}$$

$$\left(1 - \frac{\Theta\Delta t}{C_T}L_y\right)\Delta T^{**} = \Delta T^* \tag{6b}$$

$$\left(1 - \frac{\Theta\Delta t}{C_T}L_z\right)\Delta T^n = \Delta T^{**} \tag{6c}$$

The factorization step introduces errors involving the products of the three operators. The order of these errors is the same as the temporal truncation error, which has already been neglected. The prime advantage of factorization is that the solution of the banded matrix representing the three-dimensional problem reduces to the highly efficient solution of three tridiagonal systems.

To avoid non-symmetry due to different nodal capacitance values (C_T), the solution of each of the above equations begins with multiplying each side of the equality by $C_T/(\Theta\cdot\Delta t)$. By using the air-nodes to model adiabatic boundaries, each surface actually has a fixed (or Dirichlet) boundary condition. As the nesting progresses, each nest will have surfaces with known fixed values at the n and $n+1$ time levels. Fixed temperature boundary conditions at the intermediate (*) and (**) solution levels are specified by use of the three one-dimensional factored equations. Hence, in order to specify ΔT^* on a surface where ΔT^n is known, the equations are solved in reverse order. Namely, the equation relating ΔT^n and ΔT^{**} (i.e., Eq. (6c)) is used to calculate ΔT^{**} on the fixed temperature surfaces, and then the equation relating ΔT^{**} and ΔT^* (i.e., Eq. (6b)) is used to calculate ΔT^*. After calculating ΔT^{**} and ΔT^* on the fixed temperature surfaces, a recursion scheme to solve for the remaining unknown nodes in the domain becomes straightforward.

8. Results for validation

A sample problem, with a known analytical solution, was constructed to study the adaptive nesting concept and provide a framework to assist in drawing conclusions related to defining nest boundaries, solution improvements, and interpolating values for the new nodes. The sample problem consists of a two-dimensional solid made of a single material and possessing one heat source. The problem in this section was chosen to be two-dimensional solely for the ease of displaying the results in tabular and graphical forms. The self-adaptive solution of a fully three-dimensional transient problem will be demonstrated at the end of the article. The overall problem dimensions are designated by L_x and L_y and the heat source dimensions by Q_x and Q_y. The geometry for the sample problem is illustrated in Fig. 4(a).

The analytical solution is determined by numerically integrating over time where the integrand is the volume-integrated Green's function in infinite space used to represent the response of the heat equation at a point. The physical boundaries and the fixed temperature surface are represented through the use of images. The result of the numerical integration is the time history of the value at the evaluated location as well as the steady-state solution [19]. Steady-state results with temperature-dependent thermal conductivity may be obtained by the use of Kirchhoff's transform [20], but an analytical solution to the transient portion of the problem with temperature-dependent properties is not available. Using superposition, the transient response under linear (constant thermal conductivity) conditions is used to predict the response to a pulsed condition. The use of an analytical problem with a known solution makes it possible to assess the error estimation technique as well as demonstrate convergence.

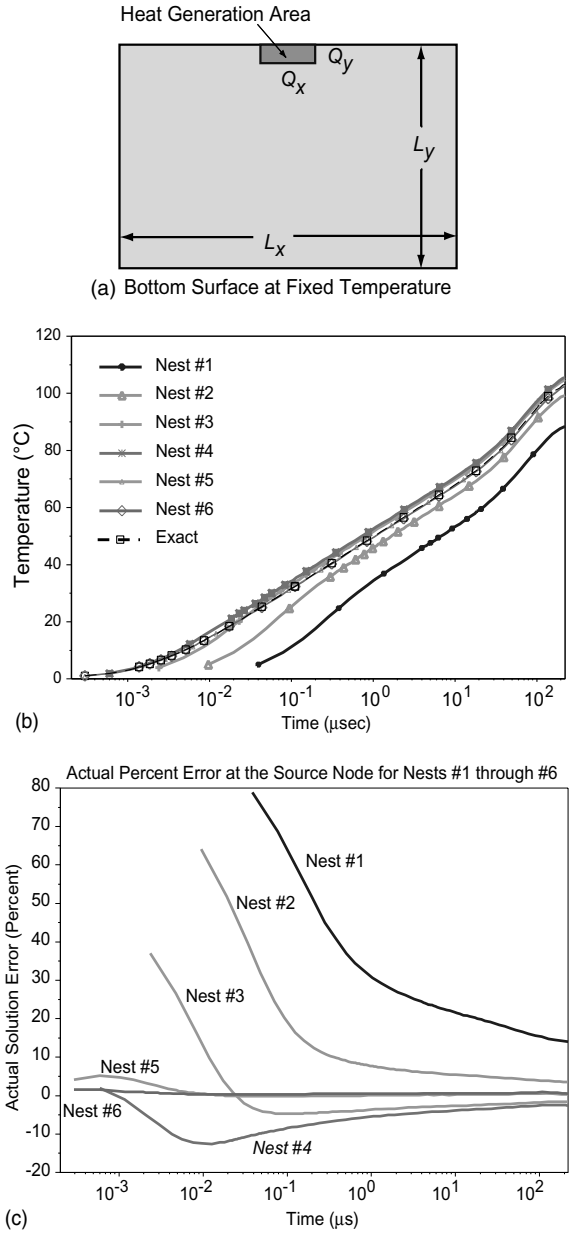


Fig. 4. Analytically solvable problem: (a) geometry, (b) peak temperature, (c) actual error.

8.1. 2D comparisons

In this section, the numerical approach is validated by comparing the numerical results with the analytical solutions for the cases of steady-state, transient, and pulsed operation.

8.1.1. Space

A heat input of 1 unit (results in this section are in dimensionless form) was placed into the heat source with dimensions $Q_x = 0.004L_x$ and $Q_y = 0.003L_y$. The analytical solution for the temperature at the top of the heat source is 121.9. A coarse grid solution (5×5 nodes) predicted a heat source value of 68.9 while a once finer grid solution (9×9 nodes) predicted a heat source temperature of 79.7. The *estimated* error at the source node (13.55% by a comparison between fine and coarse grid solutions) is below the *actual* error in the fine grid solution (34.7%) but the source node result is correctly identified as unsatisfactory.

Table 1 lists the actual error in the fine grid solution, indicating that the fine grid solution is within 2% of the exact answer for all of the nodes except the source node. The reason for the larger discrepancy at the source node is that the fine grid spacing is $0.125L_y$ while the smallest heat source spacing is $0.003L_y$. Hence, the fine grid mesh is not small enough to resolve the heat source. The error estimates in Table 1 correctly identify the insufficiently resolved region as being centered on the source node. The fact the errors decay away from the source node makes it possible to use a search routine that begins at source node(s) and progresses outwardly until the local error has decreased below the threshold established by the user. Hence, it is not necessary to waste time searching the entire computational domain.

Given a field of predicted error, a region (*child nest*) may be created that contains the nodes that have predicted error values exceeding a user-specified criterion. The location of the child boundary must, at a minimum, be such that the predicted errors on the child boundary are less than the specified criterion. However, it is likely that the location of the child boundary should be extended farther away from the region of predicted error. This extension is designated as *node overlap*. Increasing

Table 1
Actual percentage error in the fine grid solution

Numbers in ( ) refer to fine grid node coordinates									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	−0.59	−0.72	−1.15	−1.49	34.67	−1.49	−1.15	−0.72	−0.59
(2)	−0.32	−0.26	0.12	1.08	−1.48	1.08	0.12	−0.26	−0.32
(3)	0.03	0.11	0.30	0.20	−1.14	0.20	0.30	0.11	0.03
(4)	0.20	0.23	0.19	−0.15	−0.67	−0.15	0.19	0.23	0.20
(5)	0.25	0.20	0.12	−0.16	−0.43	−0.16	0.12	0.20	0.25
(6)	0.22	0.16	0.05	−0.16	−0.31	−0.16	0.05	0.16	0.22
(7)	0.16	0.16	0.03	−0.13	−0.24	−0.13	0.03	0.16	0.16
(8)	0.16	0.09	0.02	−0.01	−0.16	−0.01	0.02	0.09	0.16
(9)	0	0	0	0	0	0	0	0	0

the node overlap will increase the computational cost because all the child meshes will be larger.

Six model evaluations for the sample problem were made to demonstrate the benefits of nesting. The results presented in Table 2 indicate that the same answer may be obtained (to within a specified tolerance) by starting with a coarse initial grid and nesting as compared to a solution that uses a single and much finer grid resolution. In the cases presented, the requested accuracy was fixed at 1%, with each case doubling the resolution of the previous one. The first column in Table 2 lists the grid spacing, and the numbers in brackets refer to the nest levels. The finest grid spacing of 0.000977 was sufficiently small for all six cases, and at this resolution, all six cases met the accuracy criteria. The last row in the table shows the total number of nodes for all the nests in each case. Case 1 used about 100 times fewer nodes than case 6 and still met the accuracy goal. Since the computational cost is a strong function of the number of nodes used, case 1 is clearly preferable. The main observations from these simulations are that this problem may be effectively solved by a nesting approach and that the total number of nodes required to achieve an accurate solution is significantly smaller with the nesting approach.

8.1.2. Validation of error estimation

Solving the same sample problem geometry with a heat source that is about two orders of magnitude smaller further illustrates the benefits of self-adaptive nesting. The heat source dimensions were decreased to $Q_x = 0.00004L_x$ and $Q_y = 0.00002L_y$ and the requested spatial error criterion was set at 1%. Thus, in this case, two criteria must be met for completion. The first is that the final grid size must be less than at least one-half the source size and the second is that the requested error criterion must be satisfied. Table 3 lists the grid spacing compared to the heat source size in the second column and the error estimate in the fourth column. Both criteria are met in Nest #13 with the grid size to heat source ratio of 0.2545 and the predicted error of 0.133%. The actual error in the solution is listed in the fifth column and a comparison to the predicted error in the fourth column shows that the initial nests significantly underestimate the error, but the last nest (Nest #13) actually meets the 1% requirement. The last column lists

Table 2
Actual percentage error versus grid size and nest level

Numbers in [ ] are nest levels for the particular case						
Grid spacing ratio ($\Delta X/L_x$)	Case 1 (9×9)	Case 2 (17×17)	Case 3 (33×33)	Case 4 (65×65)	Case 5 (129×129)	Case 6 (257×257)
0.125	[1] 34.7					
0.0625	[2] 25.7	[1] 25.7				
0.03125	[3] 16.7	[2] 16.7	[1] 16.7			
0.015625	[4] 7.74	[3] 7.71	[2] 7.65	[1] 7.58		
0.007813	[5] 1.24	[4] 1.27	[3] 1.33	[2] 1.40	[1] 1.47	
0.003906	[6] 2.75	[5] 2.78	[4] 2.84	[3] 2.91	[2] 2.98	[1] 3.0
0.001953	[7] 0.62	[6] 0.59	[5] 0.51	[4] 0.44	[3] 0.36	[2] 0.31
0.000977	[8] 0.46	[7] 0.43	[6] 0.35	[5] 0.28	[4] 0.20	[3] 0.07
Total no. of nodes	674	695	1432	4505	16,858	66,215

Table 3
Nesting approach with a small source volume

Nest level	(Grid spacing)/(source size) ($\Delta x/Q_x$)	Fraction of original area	Error estimate at source (%)	Actual error at source (%)	Actual error on nest boundary (%)
#1	1040	1.0	11.3	51.2	0.0591
#2	520	0.0139	10.1	45.7	0.724
#3	260.5	0.00347	9.2	40.2	0.507
#4	130	0.868×10^{-3}	8.42	34.7	0.549
#5	65	0.217×10^{-3}	7.77	29.2	0.585
#6	32.55	0.543×10^{-4}	7.21	23.7	0.614
#7	16.3	0.136×10^{-4}	6.72	18.2	0.637
#8	8.15	0.339×10^{-5}	6.3	12.7	0.655
#9	4.07	0.848×10^{-6}	5.93	7.16	0.671
#10	2.035	0.212×10^{-6}	5.59	1.66	0.684
#11	1.015	0.530×10^{-7}	5.3	−3.85	0.696
#12	0.51	0.132×10^{-7}	4.68	0.798	0.706
#13	0.2545	0.331×10^{-8}	0.133	0.666	0.758

the actual peak error on the child boundary created from the parent grid (Nest #1 creates a boundary for Nest #2 that has a peak error of 0.0591%). The results in this last column indicate that the child boundaries are actually created in regions with acceptable solution error. The second column in Table 3 shows the decreasing ratio of nest area to total problem area. The significance of this ratio is a demonstration that the very fine grid resolution is confined to a small portion of the problem geometry. By the time the solution has progressed to Nest #13, the area of the problem involved in the calculations has decreased by seven orders of magnitude. In addition, note that the actual maximum error on the boundary for nest #13 (0.758) is less than the specified 1%.

8.1.3. Time

The sample problem geometry may also be used to demonstrate that the self-adaptive nesting approach affords efficient simulation of the transient equation. Prior to discussing time step selection and transient accuracy, it is desirable to demonstrate that the transient solution for each nest is actually converging to the true solution. After gaining confidence that the technique works, issues related to accuracy and speed will be discussed.

The sample problem was solved for the transient response to a step change in dissipated power, from an initially uniform temperature condition. The time interval was set at 200 μs. The initial time step for each nest level was set equal to five times the time step limit determined from a von Neumann stability analysis for an Euler-explicit method. Subsequent time steps were controlled by comparing the results at each time step with those obtained by taking two equivalent half steps, and insisting that the source node temperatures from the two solutions agree to within 1%. The temperature at the peak node in each nest is shown in Fig. 4(b). The curves show that Nests #1 and #2 under-predict the source temperature while the predictions from Nests #3 through #6 are closer to the exact answer. Fig. 4(c) shows the actual error present as a function of time for each nest solution and also illustrates that by Nest #6, the agreement between the predicted and analytical answers has reached the order of 1%. In fact, the maximum error at the source node in Nest #6 is approximately 1.6% for the first time step and decreases to 0.6% as the solution nears a steady-state answer at 200 μs. The difference between the predicted and exact answers is 0.02 °C at the first time step and 0.3 °C at 200 μs.

8.1.4. Validation of time step selection

Initial simulations during the development of the technique used a two-time-step method to control the local solution error. The two-time-step method makes use of a comparison between a predicted solution (T_{2S}) based on two steps of $\Delta t/2$ and a predicted solution (T_{1S}) based on Δt. The local accuracy with this approach was defined by $(T_{2S} - T_{1S})/T_{2S}$ at the particular time level. While the two-step method of controlling the temporal errors works well from an accuracy standpoint, an examination of a one-step method was made to see if the computational time could be reduced. Shampine and Witt [21] reported that limiting the change in the function value is both a reasonable and efficient method of selecting the time steps. The advantage of this approach is that a confirmation run to compare solutions is not needed. Limiting the temperature change was found to be more computationally efficient than performing a comparison evaluation, but the ability to compute a temporal error estimate was lost.

8.2. Issues related to computational speed

During the development of the adaptive transient solution technique, several discoveries were made to support the goal of minimizing solution time

requirements. Choices were required in interpolation schemes, grid refinement and overlap, and iteration with variable thermal properties. As a general rule, the simplest and most straightforward method proved to be the best choice. The nesting scheme allows consistency between grid resolution and heat generation by smearing the heat dissipation into the nearest nodes. As a result, the initial (coarse) grids approximate the source distribution over a larger volume than in the later refined grids. Since the predicted values at the nodes representing the heat source will be found to be in error in the initial nests, the smeared treatment of the heat source is both acceptable and appropriate.

The first observation was that linear interpolation for the values of nodes on the boundaries was sufficient. This statement applies to both the steady-state and linear cases. While higher-order interpolation is possible, no additional accuracy is realized since the boundaries of the child nest have already been specified in a region with acceptable accuracy.

Since the problems of interest are generally three-dimensional, significant reductions in the computational cost may be achieved by selectively refining the grid in only one or two of the spatial dimensions. A reasonable indicator of the dimension most needing refinement is the dimension with the largest gradient of the predicted solution error. As a result, computational costs may be reduced by refining the grid spacing in directions consistent with the smaller dimensions while maintaining a constant grid spacing in the long dimension.

In the case of temperature-dependent thermal conductivity, iteration within a time step is required and one iteration within the time step was found to be sufficient. Additional iterations do not increase the accuracy enough to justify the additional computational cost.

9. Motivating example—revisited

The motivating three-dimensional example problem in Section 3 is now solved with the self-adaptive technique to illustrate the significant gains in computational efficiency. The comparison values are given in Table 4, which lists both the raw data and the speed increase from the conventional method to the adaptive method. As noted in Section 3, the steady-state solution required 32 min of CPU and the transient solution for a single pulse required over 8400 min of CPU time. The self-adaptive method required 4 min for a steady-state simulation and 41 min to solve one pulse cycle. The adaptive method reduced the simulation time by a factor of over 200. In addition, a sequence of four pulses was solved with the adaptive technique with a total simulation time of 127 min. The raw CPU solution times were recorded on a 275 MHz UltraSparc Sun workstation with full compiler optimization. The agreement between the steady-state results from the two simulations was within 3%.

Table 4
Computational Time Comparison

Simulated condition	Conventional method (min)	Nesting method (min)	Speed increase
Steady-state; $k(T)$	34	2.8	12
Pulsed; 1 cycle, $k(T)$	8400	41	205
Pulsed; 4 cycles, $k(T)$	–	127	–

An isometric view of a portion of the model from the new technique is illustrated in Fig. 5(a). This model used a steady-state accuracy criterion of 3% and contained 160,000 nodes in nine nests. The cross-sectional views illustrate the strong thermal gradients around the multiple gate fingers. The power density for this simulation was 0.046 W per finger, which corresponds to a FET dissipated power density of 460 mW/mm of FET periphery. An additional evaluation with a steady-state accuracy criterion of 1% resulted in less than a 1% change in the result (54.6 °C temperature rise compared to 54.4 °C) but required 380,000 nodes. The transient simulations used the nest templates defined by the 3% steady-state criterion. A time history plot for the four pulse cycles is shown in Fig. 5(b). As discussed, the agreement between the conventional and self-adaptive techniques is evident for the first pulse cycle but the self-adaptive approach allows the simulation of multiple pulse cycles in a reasonable time.

It should be noted that while the specific comparison just described relates the self-adaptive technique to a conventional finite difference (or finite volume) solution scheme, a comparison to other conventional modeling schemes (boundary element, finite element, etc.) would be expected to have similar results. The self-adaptive technique is essentially a solution methodology in which a discrete mathematical representation of a physical problem can be solved over successively smaller computational domains, whose initial and boundary conditions are passed from the larger subdomains (with coarser grids) to the smaller subdomains (with finer grids). An error in the computed variable (e.g., temperature) is calculated at common grid points between the parent and child subdomains, making it possible to identify any computational region within the larger subdomain where the error is higher than specified. The method then sheds those regions where the solution has been determined within an acceptable error and focuses the entire available computational power on the smaller subdomain(s) where the error is still unacceptable. The reader is reminded that significant portions of the physical problem are described adequately within the first or

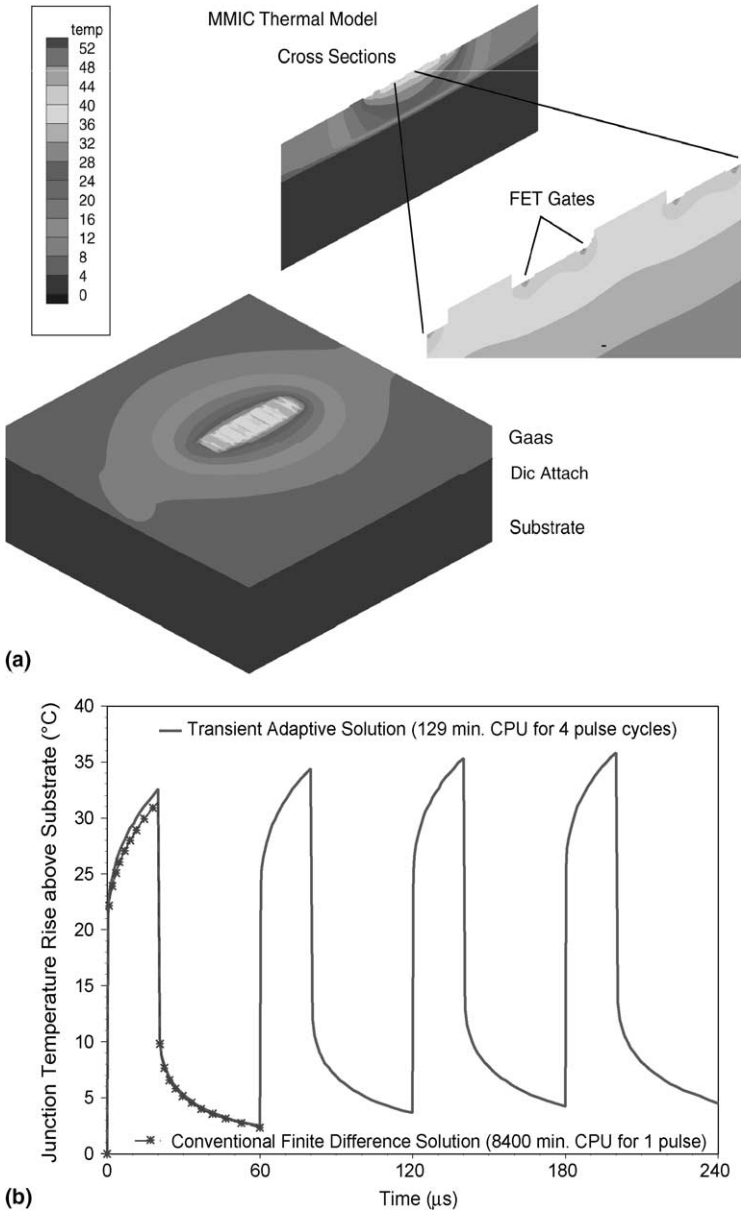


Fig. 5. Full three-dimensional thermal model: (a) steady-state contours (see color version at http://engr.smu.edu/setsl/ijhmt), (b) comparison of transient conventional and adaptive solutions.

second mesh and thus do not need to be recalculated in the framework of this self-adaptive approach. The two key and differentiating aspects of this approach then are that (i) the physics, and not the geometry, drive the grid refinement, and (ii) there is no need to solve again over regions where the solution has already reached the desired accuracy level. A tremendous computational advantage ensues by thus forcing the mesh refinement to be driven by the physics of the problem and to be confined to the region of interest, especially in the temporal domain. This process is entirely independent of which discretization scheme is used to approximate the nonlinear, partial differential equations being solved. Thus, the efficiency and effectiveness of the self-adaptive methodology should be the same for any applicable method. Indeed, the authors' unreported experience with a boundary element method early on in the investigation confirmed these advantages.

10. Concluding remarks

A novel self-adaptive nested grid method has been developed and presented. The method is capable of accurately and efficiently solving thermal transient problems that are characterized by a large range of spatial and temporal scales. The method has also been demonstrated to be significantly faster than conventional thermal modeling techniques. The developed approach is adaptive, requiring that the user specify only the physical geometry, boundary conditions, and an accuracy criterion. Integration of thermal design into the electrical design process is now for the first time possible due to the dramatic reduction in computational time requirements. The results of an associated experimental validation effort are contained in [19,22,23].

Acknowledgements

The authors acknowledge the financial support of Texas Instruments and Raytheon, in the form of a Ph.D. Fellowship for JSW, and research funding for PER. The authors also acknowledge the support of SMU and the National Science Foundation under grants DMII-9632798 and ECS-9601570. We also gratefully acknowledge the support and contributions of our colleague and friend, Dr. Donald C. Price, in initiating and securing the necessary funding for this effort and the associated experimental investigations.

References

[1] S.M. Sze, Physics of Semiconductor Devices, second ed., Wiley, New York, 1981.
[2] H.F. Cooke, Precise technique finds FET thermal resistance, Microwaves & RF 25 (1986) 85.
[3] H. Kabir, A. Ortega, A new model for substrate heat spreading to two convective heat sinks: application to the BGA package, in: Proceedings of SEMI-THERM 14, San Diego, CA, 1998.
[4] V. Kadambi, N. Abuaf, An analysis of the thermal response of power chip packages, IEEE Trans. Electron Devices ED-32 (1985) 1024.
[5] ABAQUS Standard User's Manual, Version 5.6, Hibbitt, Karlsson & Sorensen, Inc., Pawtucket, RI, 1995. Available from <http://www.abaqus.com>.
[6] TAS User's Manual, Thermal Analysis System, Version 4.0, Harvard Thermal Inc., Harvard, MA, 1999. Available from <http://www.harvardthermal.com>.
[7] D. Dawson, Thermal modeling, measurements and design considerations of GaAs microwave devices, in: Proceedings of IEEE GaAs IC Symposium, Cleveland, OH, 1994.
[8] M.J. Berger, Adaptive mesh refinement for hyperbolic partial differential equations, Ph.D. thesis, Department of Computer Science, Stanford University, Stanford, CA, 1982.
[9] M.J. Berger, J. Oliger, Adaptive mesh refinement for hyperbolic partial differential equations, J. Comput. Phys. 53 (1984) 484.
[10] R.F. van der Wijngaart, Composite-grid techniques and adaptive mesh refinement in computational fluid dynamics, Ph.D. thesis, Department of Mechanical Engineering, Stanford University, Stanford, CA, 1989.
[11] M.J. Berger, P. Collela, Local adaptive mesh refinement for shock hydrodynamics, J. Comput. Phys. 82 (1989) 64.
[12] J.B. Bell, M.J. Berger, J.S. Saltzman, M. Welcome, Three-dimensional adaptive mesh refinement for hyperbolic conservation laws, SIAM J. Sci. Comput. 15 (1994) 127.
[13] A.W. Cook, A consistent approach to large eddy simulation using adaptive mesh refinement, J. Comput. Phys. 154 (1999) 117.
[14] A.M. Roma, C.S. Peskin, M.J. Berger, An adaptive version of the immersed boundary method, J. Comput. Phys. 153 (1999) 509.
[15] K.G. Powell, P.L. Roe, T.J. Linde, T.I. Gombosi, D.L. De Zeeuw, A solution-adaptive upwind scheme for ideal magnetohydrodynamics, J. Comput. Phys. 154 (1999) 284.
[16] TEMP (Thermal Electronics Modeling program) User's Guide, 1997, Document Number UG810097, Raytheon Electronic Systems (proprietary general purpose heat transfer code).
[17] P.E. Raad, J.S. Wilson, D.C. Price, 1996, System and Method for Predicting the Behavior of a Component, US Patent No. 6,064,810, 2000.
[18] R.M. Beam, R.F. Warming, An implicit factored scheme for the compressible Navier–Stokes equations, AIAA J. 16 (1978) 393.
[19] J.S. Wilson, Self-adaptive dynamic thermal simulation of microwave integrated circuits, Ph.D. thesis, Department of Mechanical Engineering, Southern Methodist University, Dallas, TX, 1997.
[20] M.N. Özisik, Heat Conduction, John Wiley, New York, 1980, p. 440.
[21] L.F. Shampine, A. Witt, A simple step size selection algorithm for ODE codes, J. Comput. Appl. Math. 58 (1995) 345.
[22] J.S. Wilson, P.E. Raad, D.C. Price, Transient adaptive thermal simulation of microwave integrated circuits, in: Proceedings of SEMI-THERM 14, San Diego, CA, 1998.
[23] P.E. Raad, J.S. Wilson, D.C. Price, Adaptive modeling of the transients of submicron integrated circuits, IEEE Trans. Components, Packaging, Manufact. Technol. 21 (1998) 412.