Transient 3D heat flow analysis for integrated circuit devices

using the transmission line matrix method on a quad tree mesh

T. Smy[*]**, D. Walkey, S.K. Dew**[1]

Department of Electronics
Carleton University

Ottawa, ON, Canada K1S 5B6
[1] Department of Electrical and Computer Engineering
University of Alberta, Edmonton, AB, Canada T6G 2G7

Abstract

This paper presents a 3D transmission line matrix implementation(TLM) for the solution of transient heat flow in integrated semiconductor devices. The implementation uses a rectangular discontinuous mesh to allow for local mesh refinement. This approach is based on a quad tree meshing technique which can have a complex geometry using blocks of varying sizes. Each such block can have a maximum of two adjacent blocks on any vertical side and a maximum of four blocks on the top or bottom.

The TLM implementation is based on a physical extraction of a resistance and capacitance network and then the creation of the appropriate TLM matrix. The formulation allows for temperature dependent material parameters and a non-uniform time stepping.

The simulator is first tested using a 2D example of a heat source in a rectangular region. Using this example the numerical error is determined and found to be less than 0.4%. Next, nonlinearities are included, and a number of non-uniform time stepping algorithms are tested. Then, a 3D problem is also compared to an analytical solution and again the error is very small. Finally, an example of a full solution of heat flow in a realistic Si trench device is presented.

Keywords: Simulation, Device, Thermal, Transient, TLM

[*] Corresponding author

1 Introduction

The solution of the heat diffusion equation using transmission line matrix (TLM) techniques is well established[1, 2]. A specific application of TLM has been the solution of steady state and transient thermal responses of integrated circuit (IC) devices[3, 4]. This problem is of substantial interest due to the need to know maximum operating temperatures for reliability reasons and to develop temperature consistent models of the device.

The basic equation for describing the heat flow in IC devices is[6, 7]:

$$\rho C_h \frac{\partial T}{\partial t} = \nabla \left(\kappa(T) \nabla T \right) + g(x, y, z, t) \tag{1}$$

where $\kappa(T)$ is the temperature-dependent thermal conductivity, C_h the specific heat of the material, ρ the material density, and $g(x, y, z, t)$ describes the local heat generation. A simplistic description of the problem to be solved is that of a small thin heat generation plate at the base/collector junction of the device under simulation which is placed in a large semiconductor substrate characterized by a constant thermal conductivity, material density and heat capacity. The boundary conditions are somewhat problematic, but a fixed temperature on the bottom of the semiconductor and the assumption of either device "tiling" (leading to a reflective boundary condition at the edges) or of a semi-infinite substrate leads to a problem that can be solved analytically using a Green's function approach. However, solving this equation for realistic IC devices is complicated by a number of factors:

1. The problem must be solved in 3D, because of the dominant role of 3D heat spreading from the small source into the substrate.

2. The electrical device equation leads to non-uniform heat generation at the base.

3. The device metalization must be simulated. This is particularly crucial for substrates with poor thermal conductivities such as GaAs.

4. Temperature-dependent material properties must be included. Both GaAs and Si exhibit temperature-dependent thermal conductivities and heat capacities. This makes Eq. 1 a nonlinear equation.

Accounting for the effects described above necessitates the use of a 3D numerical model. Some applications simply require the solution of the steady state problem and then the extraction of an effective thermal

2

resistance. R_{th} which is defined as the steady state temperature above ambient divided by the power dissipated in the device. However, many problems require a full solution of the transient response of the device. TLM has shown itself highly appropriate for solving time dependent transient heat flow problems and is particularly attractive for multimaterial and nonlinear problems as any temperature or material dependence of the thermal parameters can be locally accommodated.

Another constraint on this type of problem relates to the efficiency of the method in the face of a large spatial dynamic range. The heat flow problem for IC devices is characterized by very large thermal gradients localized around the active components which are relatively small regions (~ 1 μm) containing a power source. In these regions, a high density of mesh elements is required to obtain an accurate solution. However, it is also necessary to solve for relatively large volumes due to substrate thicknesses which are typically 50-500 μm. These two requirements make the efficient solution of the heat equation for 3D IC devices only possible with a high degree of mesh refinement around the device junctions and much larger elements towards the edges of the simulation region.

The TLM method is traditionally performed using a continuous rectangular mesh. Clearly, a uniform grid spacing would be inefficient because very many cells would be required if forced to use a fine mesh throughout. A nonuniform mesh can be readily implemented using TLM, but this is usually still continuous so that the fine mesh required near the active devices must also extend through at least part of the substrate, thereby still requiring an unnecessarily large number of cells. Ideally, a multigrid or quad tree approach should be used such that local mesh refinement need not extend beyond where it is absolutely required, thereby allowing the minimum number of cells in the calculation.

In such meshes, two or more small cells may be adjacent along a single side of a larger cell. While this allows the necessary local mesh refinement, it leaves a problem of how to connect all of the transmission lines (TL's) of the small cells to that of the large one. So far, three approaches to this problem have been made, generally treating the 2D case.[8, 9, 10] Of these, Pulko[10] and Ait-Sadi[8] opted to connect all of the small cell TL's together at the cell boundary. These merged TL's are then connected to a single TL from the large cell. In effect, this imposes a constant temperature criterion over most of the large cell boundary which is a qualitative change to the problem being considered. Wong[9] avoids this problem by connecting only the TL of the middle small cell to that of the large one. Dependent voltage sources are then connected to each of the TL's of all of the other small cells to linearly interpolate the temperature along the large cell boundary. This appears to correct some of the inaccuracies introduced by the earlier methods.[9] However,

some potential problems with this approach are that dependent sources can reduce the numerical stability of the method and also that such active sources no longer explicitly conserve energy. The former situation can lead to problems when utilizing dynamic time stepping to improve efficiency, and the latter can introduce systematic errors into the solution. A preferable approach would strictly stay with passive components to avoid these drawbacks.

This paper will describe the development of a non-uniform meshing technique based on a 3D quad tree mesh (QTM) approach and a TLM implementation using this mesh. A distinct feature of this approach is that multiple transmission lines are used if necessary to connect large cells to each of their smaller neighbors. Such an approach avoids imposing constant temperatures along the cell boundary and avoids using unnecessary dependent sources. However, such an approach results in a variable number of TL's for each cell. Further, this TLM implementation will be used for both constant and temperature-dependent material properties, and the time step will ultimately be adjusted dynamically during simulation to reduce execution time. Results confirming the accuracy of this TLM approach will be presented, and finally a full transient simulation of a realistic Si-based trench device will be given.

2   Model Development

2.1   Basic TLM for Heat Flow

TLM modeling represents a physical model of heat flow as a sequence of voltage pulses travelling through a matrix network of transmission lines. By computing the voltage (temperature) at the nodes connecting the transmission lines, an explicit, unconditionally stable solution to the *telegrapher's equation* can be described by a set of equations that are iterated in a straightforward manner.

The telegrapher's equation is given by:

$$\nabla^2 V = R_d C_d \frac{\partial V}{\partial t} + L_d C_d \frac{\partial^2 V}{\partial t^2} \tag{2}$$

where R_d, L_d and C_d are respectively, the resistance, inductance and capacitance per unit length in the transmission line and V is the node voltage. Equations 1 and 2 have a similar structure. Replacing V with T, setting $R_d C_d = \rho C / \kappa$, adding a current source to the left hand side of Eq. 2 and requiring the term involving the second time derivative to be negligible results in a correspondence between the two equations. Once R_d and C_d have been determined by the physical problem, a requirement that the pulses travel from

node to node in a time step Δt results in the complex transmission line impedance $Z = \sqrt{L_d/C_d}$ being equal to $\Delta t/(C_d \Delta x)$ where Δx is the local distance between nodes. The requirement that the second term in Eq. 2 be neglectable then becomes a condition that Δt be sufficiently small.

With these conditions, the transient heat equation can be solved in a TLM framework by iterating the following steps:

Nodal Temperatures: At each node n, the temperature at the time step k, $_kT(n)$, is calculated from the local sources and the incident temperature pulses, $_kV_b^i(n)$. The subscript b indicates which transmission line the pulse is on.

Scattered Pulses: The scattered temperature pulses $_kV_b^s(n)$, are calculated from the nodal temperatures and incident pulses.

Incident Pulses: The incident pulse for the subsequent iteration, $_{k+1}V_b^i(n)$, are computed from the scattered pulses from the previous time step, accounting for any time step changes and differences in nodal separation.

A substantial amount of work has been done on TLM solutions of heat flow, including the development of 3D and 2D models based on continuous rectangular meshes[3]. One advantage of TLM over conventional finite difference techniques is greater stability. In particular, this allows the time step to be increased as the simulation progresses towards steady state, thereby accelerating the computation substantially. Incorporating this into the TLM equations simply requires some additional computations for scattering effects at cell boundaries due to changes in the time step[10], although care is required to ensure the second time derivative term in Equation 2 remains small. Finally, by updating resistance and capacitance values during the simulation, nonlinear temperature-dependent material properties are easily incorporated into the TLM framework.

2.2 TLM on a Quad Tree Mesh

A full description of the 3D model building code is described elsewhere[11]. This model building code will, given an IC technology description and layout information, automatically generate a 3D model of the device, discretize the model, and solve for the temperature distribution in the device. The basis of this code is a quad tree meshing technique that creates a complex mesh geometry using rectangular blocks of varying

sizes. The model is created such that a block and its neighbors form a relatively simple set of topologies whereby each block can have a maximum of two blocks on any vertical side and a maximum of four blocks on the top or bottom. From these block topologies, a network of thermal resistances and capacitances can be extracted representing the thermal equation to be solved.

The essence of the multigrid quad tree mesh (QTM) is shown in Fig. 1(a). The bottom right corner shows an area of local refinement, and a reduction in mesh density is obtained by increasing the block size as we move away from it. The primary requirement of a 2D QTM is that there can be at most two smaller blocks present on any side of a given block. This allows for a very quick expansion of block size, but still minimizes the number of local block topologies. The extension of this style of meshing to 3D can be done in a number of ways, but perhaps the most obvious is to allow the top and bottom of a block to have at most four smaller blocks adjacent. This results in block topologies such as shown in Fig. 1(b).

The fundamental approach to TLM modeling of the QTM is shown for a 2D case in Fig. 2(a). At the center of each block is a temperature node, and a transmission line from this node to all adjacent nodes is created. Each transmission line consists of two impedances and two resistances, one of each associated with each block. This is shown in Fig. 2(b) for two nodes denoted as n_i and n_j. Although the association of $R_d C_d$ with $\rho C / \kappa$ allows for a number of choices in choosing the values of R and Z of each transmission line, a natural choice is to use the physical values of thermal resistance and capacitance associated with nodal connections.

Figure 3(a) shows the basic geometries used to calculate the two resistor values associated with a link between adjacent blocks in a 2D slab. The resistance, R_1, between node n_1 and the common boundary is simply given by $R = L/(\kappa W D)$ where D is the thickness of the slab, and W and L are the widths and lengths as shown in the figure. For cells which are of different sizes, the same formula holds; however, multiple branches are needed and the W used in the calculation for a given branch in the larger cell is reduced to that of the corresponding smaller cell. This subdivision of the large cells results in several parallel resistors whose equivalent resistance is equal to that which would have resulted had the entire width of the cell been used in a single resistor. Extending this to 3D just requires treating D in the same manner as is done for W.

To calculate the thermal capacitance associated with each node, the total capacitance in a block is determined by $C_{tot} = V_{tot} \rho C_h$ where V_{tot} is the block volume, ρ the material density and C_h the specific heat of

the material. As shown in Fig. 3(b), this capacitance can be apportioned to each transmission line according to the area (or volume) it corresponds to. In 2D, this is given simply as $C_{tot}/4$ divided by the number of adjacent cells along the corresponding edge.

Although the calculation of the line resistances and capacitances was developed for a 2D mesh, very similar expressions are determined for the 3D case. The above approach to the extraction of the transmission line resistances corresponds mathematically to the approach described in [9] with the appropriate reduction for a quad tree mesh and an extension to 3D. Once the capacitances have been determined, the line impedances can be calculated using $Z_n = \Delta t / C_n$ for a given time step.

The use of this formulation in 2D or 3D QTM results in a network of nodes (at the block centers) connected by transmission lines. Unlike previous multigrid[8, 9, 10] approaches, connections only occur at the nodes. We denote each line as a "branch" with the number of branches at each node n being $B(n)$. Each TLM node in the problem can be described as having a node temperature $_kT(n)$ at time step k and associated with each branch b of the node is a resistance $R_b(n)$, an impedance $Z_b(n)$, a scattered temperature pulse $_kV_b^s(n)$ and a incident pulse $_kV_b^i(n)$. Also associated with the node is a current source $G(n)$ if heat is being generated in the block represented by that node. A generalized TLM node is shown in Fig. 4.

Using the standard TLM approach and the above nomenclature, we obtain for the node temperature at the time step k:

$$_kT(n) = \left[\sum_b^{B(n)} \frac{2_kV_b^i(n)}{R_b(n) + Z_b(n)} + G(n) \right] \frac{1}{Y} \tag{3}$$

with,

$$Y = \sum_b^{B(n)} \frac{1}{R_b(n) + Z_b(n)} \tag{4}$$

the summations are over all branches present at that node.

The scattered pulse on each branch is then given by:

$$_kV_b^s(n) = {_kT(n)}\frac{Z_b(n)}{R_b(n) + Z_b(n)} + {_kV_b^i(n)}\frac{R_b(n) - Z_b(n)}{R_b(n) + Z_b(n)} \tag{5}$$

The equation connecting the scattered pulses at the time step k and the incident pulses at $k + 1$ must take into account both the scattering at the discontinuity in Z at the block boundaries and the scattering due to time step changes. Taking these into account we obtain a relation for each branch of the node:

$$_{k+1}V_b^i(n) = {_kV_{b'}^s(n')}\tau_{b'} + {_kV_b^s(n)}\rho_b \tag{6}$$

7

where

$$\tau_{b'} = \frac{(1+\beta)Z_b(n)}{Z_b(n) + Z_{b'}(n')} \tag{7}$$

$$\rho = \frac{Z_{b'}(n') - \beta Z_b(n)}{Z_b(n) + Z_{b'}(n')} \tag{8}$$

and $\beta =_k \Delta t /_{k-1} \Delta t$ is the ratio of the current time step to the previous one. For the above expressions n' and b' denote respectively the node and reverse branch at the end of the transmission line.

The TLM solution of the heat diffusion problem requires only the repeated solution of Eq. 3, 5 and 6 with appropriate initial and boundary conditions. Two basic boundary conditions are needed: 1) a fixed temperature on a block face or 2) a zero heat flow condition across a block face. Implementing these boundary conditions requires replacing Eq. 6 for the branches connected to a boundary. The first condition requires the implementation of the following equation for the branch connected to the face:

$$_{k+1}V_b^i(n) = V_C - {_k}V_b^s(n) \tag{9}$$

where V_C is the fixed voltage on the face. The zero flow condition results in:

$$_{k+1}V_b^i(n) = {_k}V_b^s(n) \tag{10}$$

for any branch with a reflective boundary condition at the end.

The initial condition used was to assume the device was at a uniform initial temperature, V_0. To impose this condition, all initial incident pulses were set to $V_0/2$.

The above formulation assumed constant material parameters. It is, however, straightforward to update temperature-dependent material parameters at each iteration. An additional requirement for an accurate solution is that the second term in Eq. 2 involving a second time derivative is negligible. This is essentially a requirement for a sufficiently small time step. A significant advantage of the TLM method is the ability to increase the time step size as the simulation progresses towards steady state. The implications of increasing the time step on solution accuracy will be discussed below.

3 Results

In order to evaluate the accuracy of the TLM algorithm discussed above, the TLM solution on both a QTM mesh and a uniform mesh were compared to an analytical solution for a number of representative geometries.

3.1 2D slab

The first geometry considered was a 2D slab with a fixed temperature on the left side and a square heat generation region in the center of the slab. The other edges of the slab were assumed to have a zero heat flow condition. The basic geometry is shown for both a QTM and a uniform mesh in Fig. 5(a). The slab dimensions are 64 μm by 64 μm with a 16 μm square uniform heat generation region (100 mW). For these initial simulations, the time step was kept uniform and was at a value of one tenth the smallest RC time constant present in the slab. The slab material was Si with a thermal conductivity of $1.2 \times 10^{-4} W/\mu m\text{-}K$, density of $2.33 \times 10^{-12} g/\mu m^3$ and specific heat of $0.7 J/g\text{-}K$ were used.

A time dependent analytical solution for this problem can be developed[12] using a Green's function approach. This leads to the following expression for the temperature:

$$
\begin{aligned}
T(x,y,t) \;=\; & \frac{g}{\alpha\kappa}\left[\frac{y_2-y_1}{w}+\sum_{n=1,2,\ldots}^{\infty}e^{-\alpha\gamma_n^2 t}\left(\frac{2\cos(\gamma_n y)}{w}\right)\left(\frac{\sin(\gamma_n y_2)-\sin(\gamma_n y_1)}{\gamma_n}\right)\right] \\
& \left[\sum_{m=1,3,5\ldots}^{\infty}e^{-\alpha\beta_m^2 t}\left(\frac{2\sin(\beta_m x)}{l}\right)\left(\frac{\cos(\beta_m x_2)-\cos(\beta_m x_1)}{\beta_m}\right)\right]
\end{aligned}
\tag{11}
$$

where w and l are the slab dimensions, g is the power density over the region defined by the corner points (x_1,y_1) and (x_2,y_2), $\gamma_n = n\pi/w$, $\beta_m = m\pi/2l$ and $\alpha = \kappa/C_h\rho$.

In Fig. 5, the basic slab mesh geometry is shown for both the QTM and a uniform mesh of 32×32. This QTM mesh shows an aggressive attempt at reducing the number of blocks and contains 220 blocks. The characteristic QTM mesh structure can be seen. The uniform mesh has, of course, 1024 blocks. Also shown in this figure are grayscale depictions of the temperature distribution in the slab at times of 0.1 μs and 15 μs. As can be seen, the two meshes produce very similar distributions. In Fig. 6(a), a comparison between the two mesh solutions and the analytical solution is shown for the temperature of a point at the center of the slab. The temperature is shown as a function of time and it can clearly be seen that the solutions for both meshes compare very favorably with the analytical solution. However, the uniform mesh solution took considerably longer to compute.

The residual error (the absolute difference between the analytical solution and the numerical solution normalized to the final temperature) is also shown for a number of cases in Fig. 6(b). Shown in this plot are three TLM solutions and two solutions based on a simple finite difference (FD) approach. (While both FD

and TLM produce very similar results for a given mesh, the TLM method has the potential to do so with considerably less calculation by stretching the time step. This is discussed further below.) The minimal error is found for the highest mesh density and a uniform grid of 1024 elements. In this instance, the FD solution initially shows a larger error for the transient but by 5 μs, the residual errors are very similar. The low density QTM with 220 elements shows a larger steady state error by a factor of about 4, with the FD solution again having a larger error initially. An intermediate QTM mesh of 350 elements using a TLM solution reduces the error to only around 2.5 times that of the high density result. It is important to note that for all cases the residual error is less than 0.4% of the normalized transient and less than 0.5 K for an absolute error.

To enable an easy evaluation of the effect of the quad tree mesh on the spatial temperature distribution, the temperature has been plotted in Fig. 7 in the x and y directions through the center of the slab from one edge to the other. The temperature profile is plotted for three different times during the evolution to steady state. As can be seen, both meshes produce a reasonably accurate solution throughout the region.

The results described above indicate that a 2D QTM mesh essentially produces as accurate a solution for the temperature as a uniform high density mesh, with a reduction in the number of blocks by a factor of four.

3.2 Non-uniform time stepping

One of the main advantages of TLM over FD simulation methods is the ability to reduce execution time by optimizing the time step during the simulation. In this section, we explore some simple methods of altering or controlling the time step during the simulation. To characterize the error present in the simulation during the transient, an error parameter was evaluated[5]. The parameter used was calculated using:

$$e = \left| 2L_d C_d \left(\frac{_k T(n) - 2_{k-1} T(n) +_{k-2} T(n)}{_k \Delta t +_{k-1} \Delta t} \right) \right| \tag{12}$$

which was averaged over the heat generation region to calculate e_{ave}. This parameter is a direct estimation of the size of the second derivative error term in Eq. 2.

A variety of algorithms for step size growth was investigated using the QTM of 220 blocks shown in Fig. 5. The most obvious method of increasing the step size is a uniform growth rate given by $_k \Delta t = (1 + G)_{k-1} \Delta t$ where G is a fixed and typically less than 5%. Fig. 8(a) shows the transient temperature response at the center of the block and the average error, e_{ave}, for three different growth rates. The smallest growth rate of 0.5% follows the analytical curve very well. Increasing the growth rate to 5% causes an error

10

to develop in the steep part of the transient which then disappears as steady state is approached. Further increasing the growth rate by a factor of 2 produces a large overshoot and then oscillatory behavior as the solution moves towards steady state. The plot of average error parameter clearly shows the inaccuracy of the solution at higher growth rates. For $G = 0.5\%$, the error parameter stays essentially below 10^{-4}, and the solution is well behaved. For the large growth rates, e_{ave} quickly becomes rather large introducing errors in the initial part of the transient. For the growth rate of 10%, the error parameter is large for the whole simulation, ranging between 10^{-3} and 10^{-2}.

It was noticed that for longer simulation times, the absolute size of the time step would become too large, producing errors and oscillations in the solution. An approach to minimizing both solution error and execution time is to use a moderate growth rate but limit the time step to a maximum value. In Fig. 8(b), the results of this approach are presented. A growth rate of 1% was used and three simulations are shown corresponding to a maximum time step of 20, 50 or 100 times the smallest RC time constant (τ_{min}) present in the slab. The plot of the temperature transient shows that when limited to 20 times τ_{min} the solution closely follows the analytical result. Using 50 or 100 τ_{min} produces increasing error; however, the solution does stay relatively close to the correct answer and approaches the correct steady state value. The effect of fixing the time step can clearly be seen in the plot of the averaged error parameter with a damped oscillation occurring in e_{ave} from the point at which the time step stops growing. After this point, the error becomes progressively smaller, except in the case of 100 τ_{min}, indicating that this fixed time step is too large.

Finally, a third approach to time step control was tried which involved monitoring the error parameter. In this case, if e_{ave} became larger than 5×10^{-4}, the growth rate was held constant until it again dropped below this threshold, after which it was allowed to increase again. In Fig. 8(c), the results from this approach are shown for four different growth rates of 1, 2.5, 5 and 10%. As can be seen from the plots, this method is quite successful. A high growth rate of even 2.5 or 5% can be used with a good result obtained. A 10% growth rate does produce some error in the initial part of the transient, but even for this extreme case the error is well behaved in the latter part of the transient.

Of the above techniques monitoring the error parameter was the most successful. Of course, the point of growing the time step is to reduce the number of iterations used to simulate the transient response to a particular value of t. In Fig. 9, the total number of iterations as a function of t is presented for four cases which produce an accurate solution. These cases are: a uniform time step of $0.1\tau_{min}$, a uniform growth rate of 0.5%, a growth rate of 1% and a maximum time step of $20\tau_{min}$, and a growth rate of 2.5% and

a constraint of 5×10^{-4} on the error parameter. As can be seen, the last case produced the minimum number of iterations and therefore minimal execution time; however, all the non-uniform time stepping algorithms produce a very dramatic decrease in the simulation time. The above algorithms are by no means either optimal or comprehensive, but do indicate the large advantage of using TLM with non-uniform time stepping.

3.3 Nonlinear materials

To investigate the accuracy of the simulation with respect to nonlinear material properties, the 2D slab problem described above was simulated with a temperature-dependent thermal conductivity given by:

$$\kappa = \kappa_0 (1 + k_{c1}(T - T_0) + k_{c2}(T - T_0)^2) \tag{13}$$

where $\kappa_0 = 1.45 \times 10^{-4} W/\mu m\text{-}K$, $k_{c1} = -4.5 \times 10^{-3} K^{-1}$, $k_{c2} = 10^{-6} K^{-2}$ and $T_0 = 300K$ for Si. In Fig. 10(a), the transient response in the slab center for constant and temperature-dependent thermal conductivity is shown. The importance of including the nonlinearity can clearly be seen. Unfortunately, there is no analytic solution to the nonlinear transient problem, so that numerical techniques are required for even the simplest geometries. The second plot in Fig. 10 shows the time evolution of the temperature profile in the x-direction through the slab center. Although it is not possible to get an analytical solution of the transient response of the non-linear temperature-dependent problem, it is possible to get a steady state solution using a Kirchoff transformation[13]. This analytical result is also shown in the second plot, and it can be seen that the simulations do smoothly approach the steady state analytical solution, providing confidence in the non-linear TLM solution.

4 3D block

The advantage of the QTM is most dramatically illustrated for a 3 dimensional problem. In Fig. 11, a QTM for a 3D block of height of 50 μm and a width and length of 32 μm is shown. The bottom face of the block was fixed at 300 K, and a thin heat source region was situated near the top of the block. The heat source of 25 mW is centered at the block, is 8$\times$8 μm in its lateral dimensions, has a thickness of 0.2 μm and is 0.2 μm from the top face. The block material is Si.

As with the 2D slab, an analytical transient solution of this problem is available (assuming temperature independent material parameters), resulting in a solution similar to Eq. 11, but with a triple sum. Figure

11 presents the temperature distribution of the block using a QTM model with only 1974 blocks. However, if a uniform x-y mesh were to be used, the number of nodes would be 122,880, assuming the same non-uniform meshing in the z direction. A comparison of the time response of the model and the analytical solution is shown in Fig. 12. As can be seen from both the plot of the temperature transient response and the residual error, the QTM TLM solution is very accurate over the entire response time. A spatial temperature distribution is shown in the second figure. Both the analytical and QTM TLM solution for the temperature at the center of the block down through the depth of the block are presented. The agreement is excellent, lending confidence to the 3D extensions of the QTM TLM approach. The situation analyzed here is very similar to the final device simulations to be investigated, consisting of a block of Si with a backside fixed temperature and a small thin heating plate at the top of the block representing the heating at the base/collector junction of a microelectronic device.

5 Full Si trench simulation

Finally, a model of a realistic Si trench device is shown in Fig. 13(a). The figure shows 1/4 of the device which has a central 1 μm x 20 μm emitter with a collector contact on each side. The technology has a PtSi/W plug contact structure and two aluminum metal layers were included in the model with a W plug via. The simulation region is 64 μm square and the wafer thickness was 100 μm. The backside temperature was 300 K and the power generated in the device was 1.75 mW, uniformly generated in a region under the emitter. The ends of the emitter and collector metalization were also fixed at 300 K. The total model consists of 15,000 blocks.

A simulation of the transient response of this Si device is shown in Fig. 13(b). The temperature of point at the center of the base/collector junction is presented as a function of time. The Fig. 14 shows the change in the temperature profile at four different times.

6 Conclusions

This paper presents 2D and 3D TLM implementations of the solution for heat flow, ultimately in realistic integrated semiconductor devices. The implementation is based on a quad tree mesh which is very efficient at localized mesh refinement. Model building code will, given an IC technology description and layout information, automatically generate a 3D model of the device and discretize the model. The basis of the

quad tree meshing technique is to create a model consisting of differing size blocks, such that a block and its neighbors form a relatively simple set of topologies, whereby each block can have a maximum of two blocks on any vertical side and a maximum of four blocks on the top or bottom of a block.

From this physical model, a generalized TLM implementation is formulated by extracting a physical resistance and capacitance network. The formulation allows for non-uniform time stepping and temperature-dependent material parameters.

The TLM implementation was first tested for a simple 2D problem for which an analytical solution is obtainable. A direct comparison between the QTM TLM solution, uniform mesh solutions and finite difference solutions showed the QTM mesh TLM solution works very well. A number of algorithms for optimizing the time stepping were investigated with the best found to be a straightforward monitoring of an error parameter. It was found that increasing the time step during simulation could reduce execution time by up to 100 times. This is on top of the reduction factor resulting from using the QTM over a continuous mesh.

A simple 3D situation was also analyzed and it was found that a QTM of 1974 blocks (versus 122,880 blocks for a uniform x-y mesh) produced a very accurate solution with less than 0.5% error. Finally, an example of a full simulation of a Si trench device using 15,000 blocks was presented, displaying the full capabilities of the simulator.

Acknowledgement

References

[1] Peter B. Johns, A simple explicit and unconditionally stable numerical routine for the solution of the diffusion equation, Int. J. Numer. Eng. **11**, 1407-1328, (1977).

[2] M.Y. Al-Zeben, A.H. Saleh and M. A. Al-Omar, TLM modelling of diffusion, drift and recombination of charge carriers in semiconductors, Int. J. Numer. Model.: Electron. Netw. Devices Fields, **5**, 219-225, (1992).

[3] Xiang Gui, Paul W. Webb, and Guang-bo Gao, Use of three-dimensional TLM method in the thermal simulation and design of semiconductor devices, IEEE Trans. Electr. Dev., **39**(6) 1295–1302 (1992).

[4] Xiang Gui, Guang-bo Gao, Hadis Morkoc, The effects of surface Metallization the Thermal Behavior of GaAs Microwave Power Devices, IEEE Trans. Microwave Th. Techniques, bf 42(2) 342–344, 1994.

[5] Paul W. Webb and Ian A. D. Russell, Application of the TLM method to transient thermal simulation of microwave power transistors, IEEE Trans. Electr. Dev., **42**(4) 624–631 (1995).

[6] J. Higgens, Thermal Properties of Power HBT's, IEEE Trans. Electr. Dev., **40**(12) 2171-2177 (1993).

[7] H. Carslaw, *Conduction of heat in solids*, Ch.1, Oxford, 1959.

[8] R. Ait-Sadi, A.J. Lowery and B. Tuck, Combined fine-coarse mesh transmission line modelling for diffusion problems, Int. J. Numer. Model.: Electron. Netw. Devices Fields, **3**, 111-126, (1990).

[9] C.C. Wong and H. Xiao, A combined fine-coarse mesh method for the transmission line modelling of diffusion, Int. J. Numer. Model.: Electron. Netw. Devices Fields, **9** 405-415 (1996).

[10] S. H. Pulko, I.A. Halleron and C.P. Phizacklea, Substructuring of space and time in TLM diffusion applications, Int. J. Numer. Model.: Electron. Netw. Devices Fields, **3** 207-214 (1990).

[11] T. Smy, D. Walkey and S. Dew, A 3D thermal simulation tool for integrated devices – *Atar*, Accepted by IEEE Trans on Computer Aided Design (Sept. 2000).

[12] M. Necati Ozisik, *Heat Conduction*, Ch. 2-4, John Wiley and Sons, 1980.

[13] W.B.Joyce, Thermal resistance of heat sinks with temperature-dependent conductivity, Solid-State Electronics, **18** 321-322 (1975).
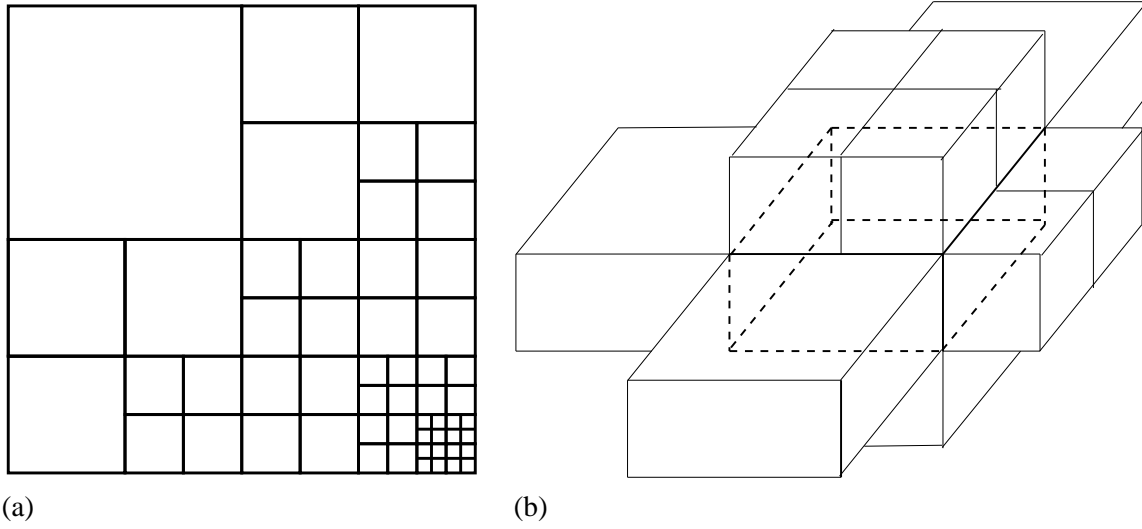
List of Figures

Figure 1: (a) A 2D quad tree mesh showing refinement and (b) a single 3D element with 4 blocks above, 2 on the right and 1 on all other sides.
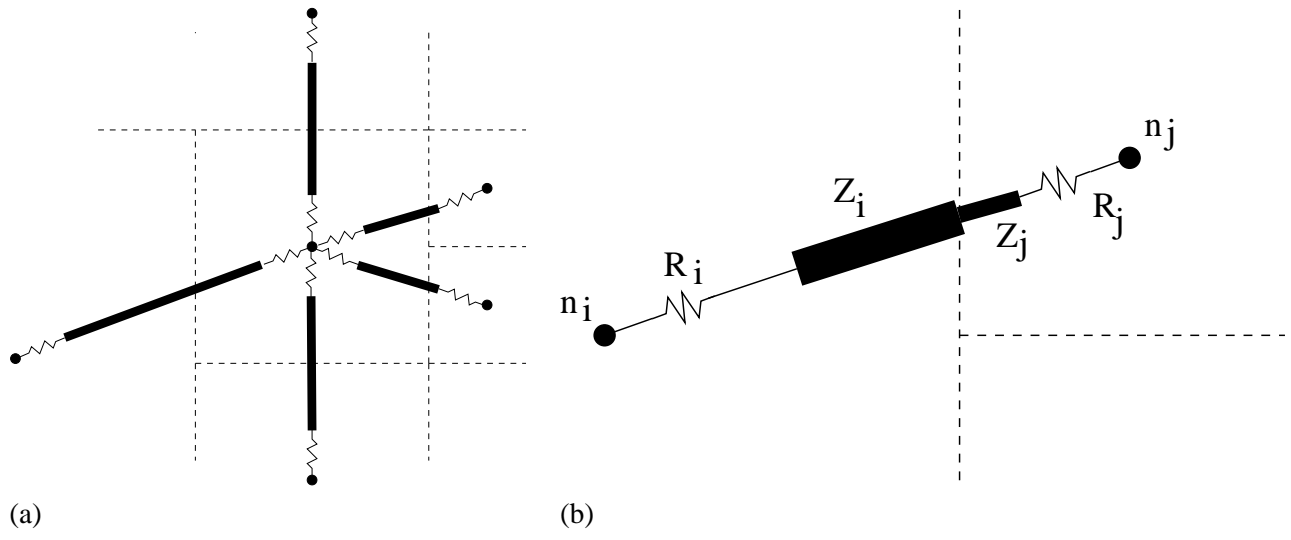

Figure 2: (a) TLM transmission line links in a quad tree and (b) details of a single link in the quad tree mesh.
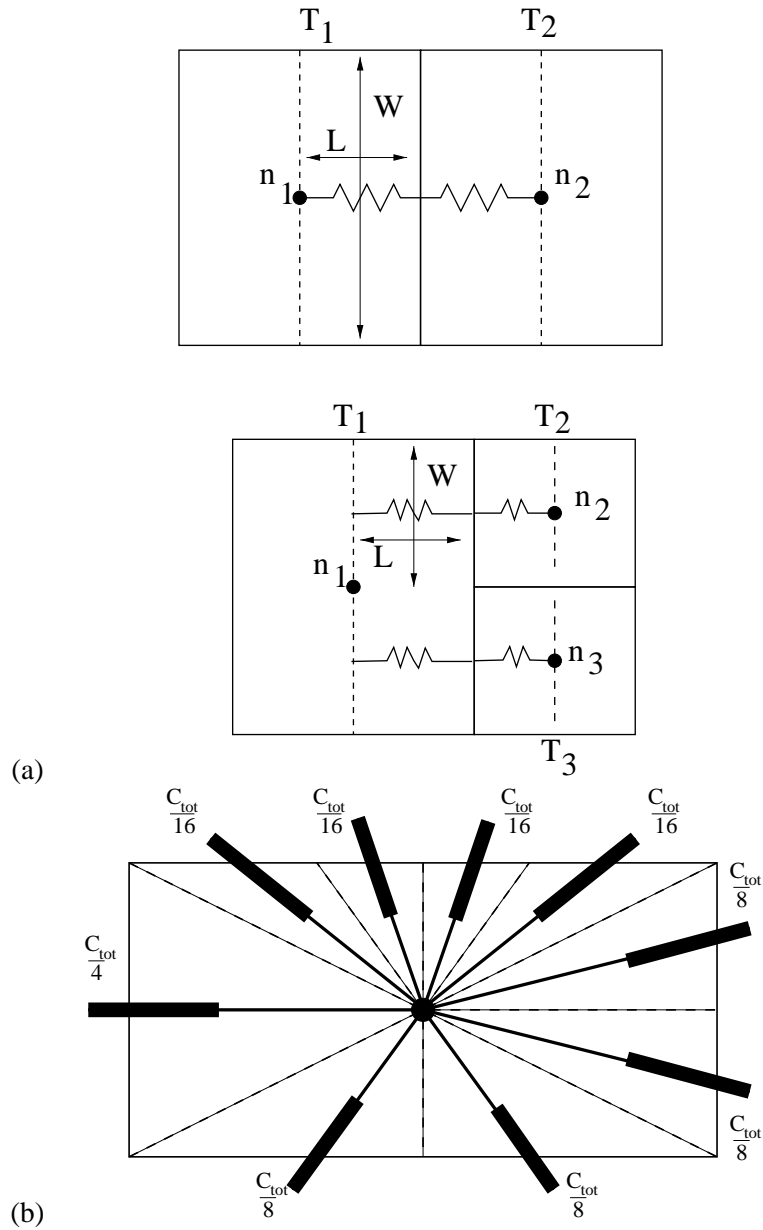
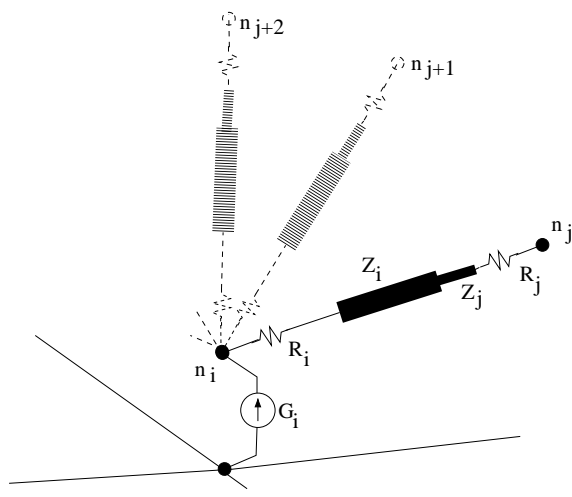Figure 3: Geometry for a 2D mesh for calculating (a) resistances and (b) capacitances.

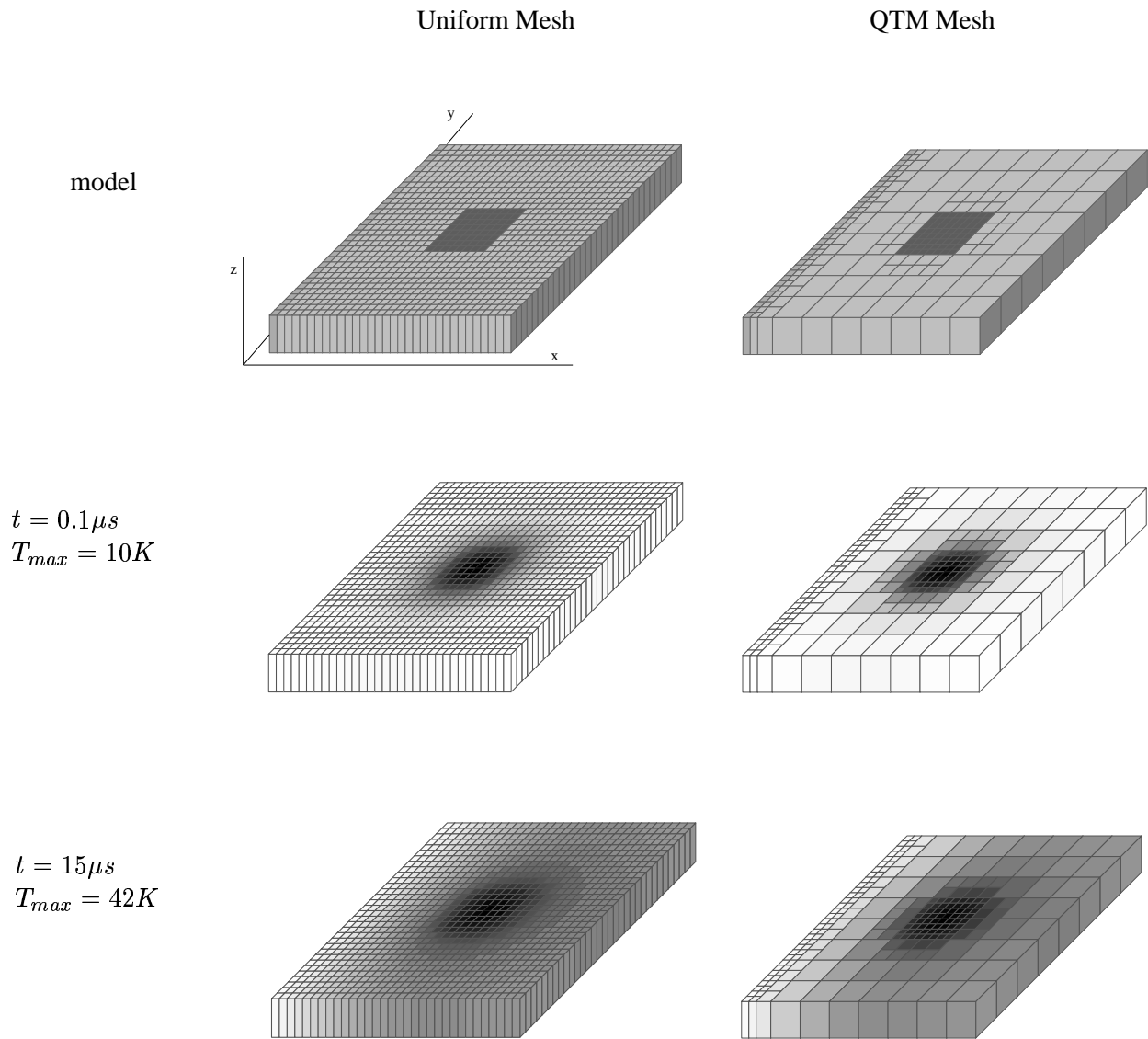Figure 4: A general TLM node for a QTM mesh.
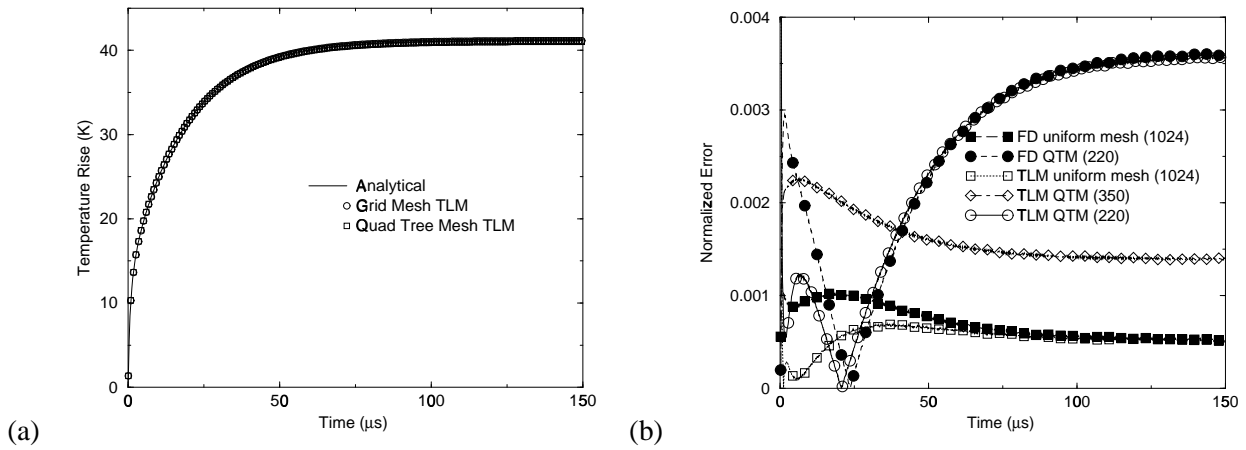
Figure 5: Mesh geometries and temperature distributions for a 2D rectangular slab.

Figure 6: Transient response of the 2D rectangular slab showing (a) temperature and (b) residual error.

Figure 7: Spatial temperature distributions through the center of the rectangular slab (a) in the x-direction and (b) in the y-direction.
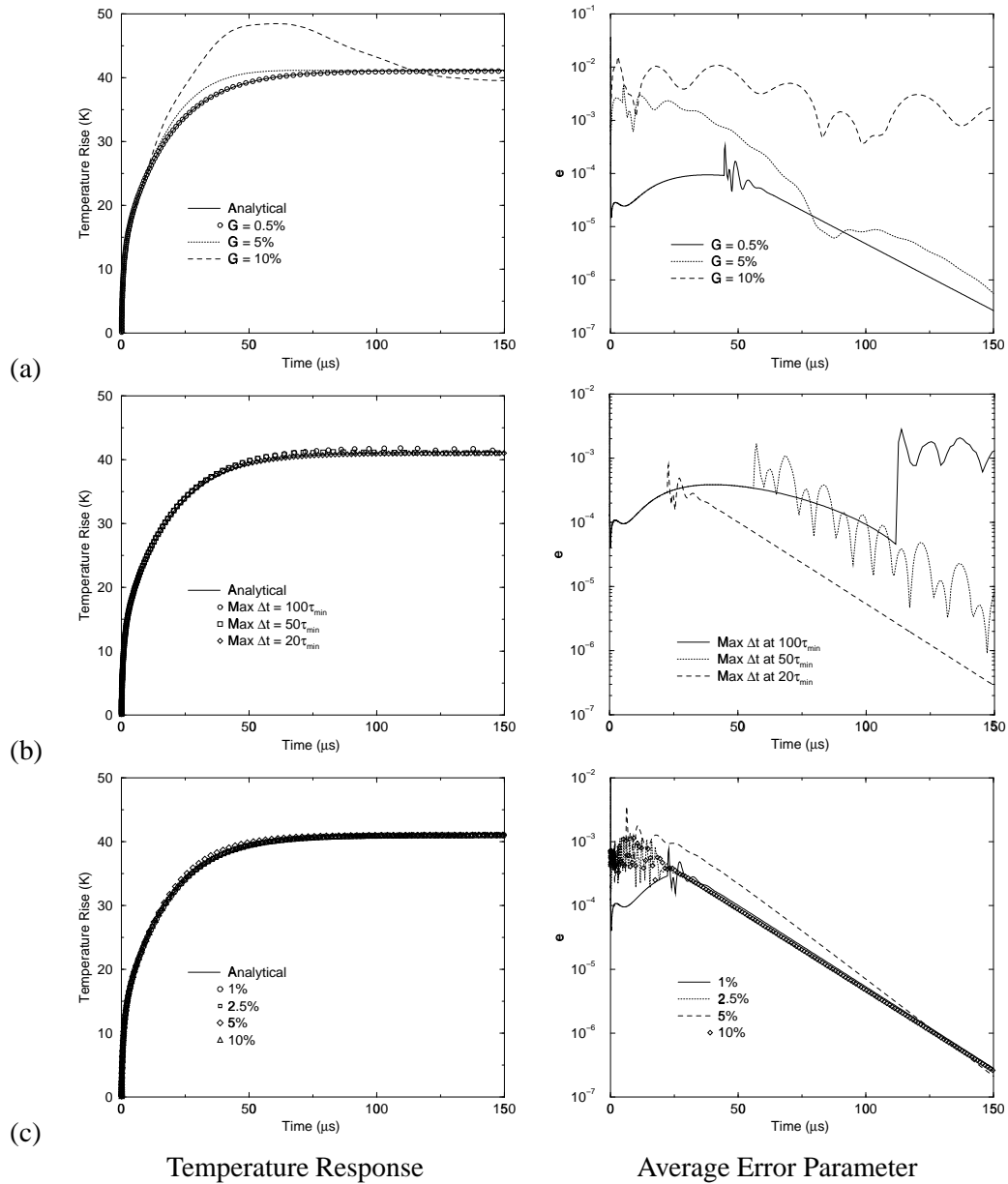
21

Temperature Response   Average Error Parameter

Figure 8: Transient response of the slab using a QTM with different strategies for non-uniform time stepping such as (a) unconstrained growth, (b) 1% growth constrained to a maximum time step and (c) growth constrained by limiting e_{ave} to 5×10^{-4}.

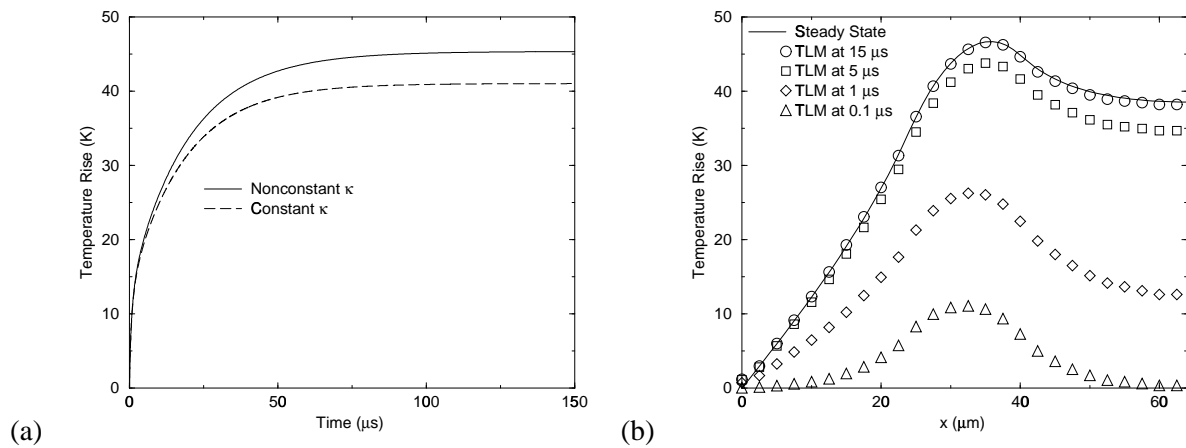Figure 9: Total number of iterations as a function of simulation time.

Figure 10: Effect of a temperature-dependent heat conductivity on (a) transient response in the slab center and (b) temperature profile in the x-direction.
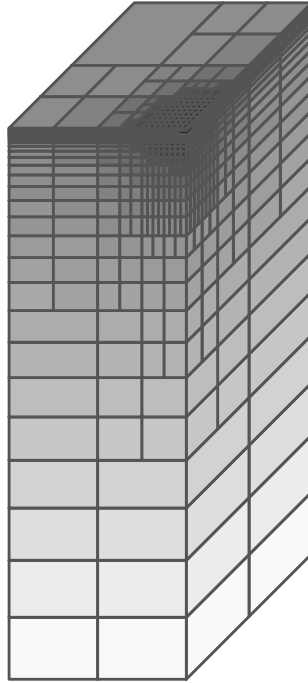
Figure 11: Grayscale representation of the temperature distribution at 100 μs in a rectangular 3D solid using a QTM mesh. The darkest regions correspond to a temperature rise of 65 K. Due to the symmetry, only one quadrant of the block is shown.
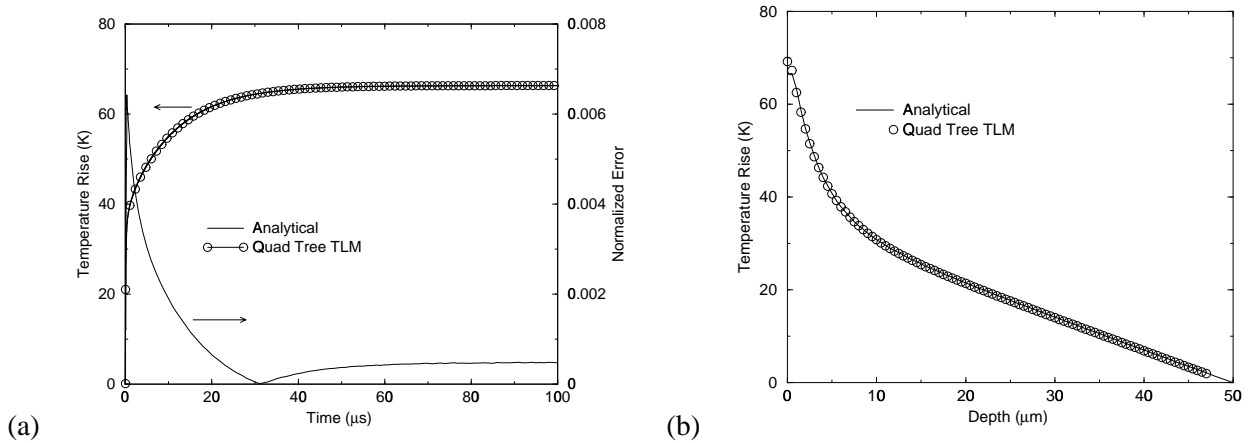

(a)

(b)

Figure 12: Temperature response within a 3D rectangular solid showing (a) transient response of maximum temperature and (b) temperature profile as a function of depth near the steady state.
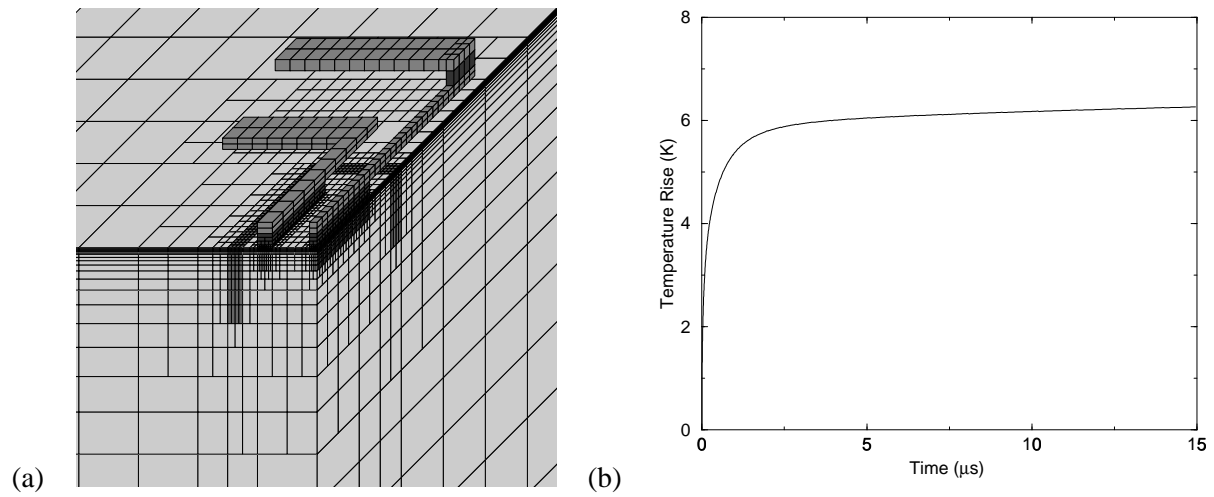
(a)   (b)

Figure 13: (a) Details of a Si trench device showing structure and metalization and (b) transient thermal response of the device.

t = 0.05 μs, $\Delta T_{max} = 3.0$ K   t = 0.25 μs, $\Delta T_{max} = 3.8$ K

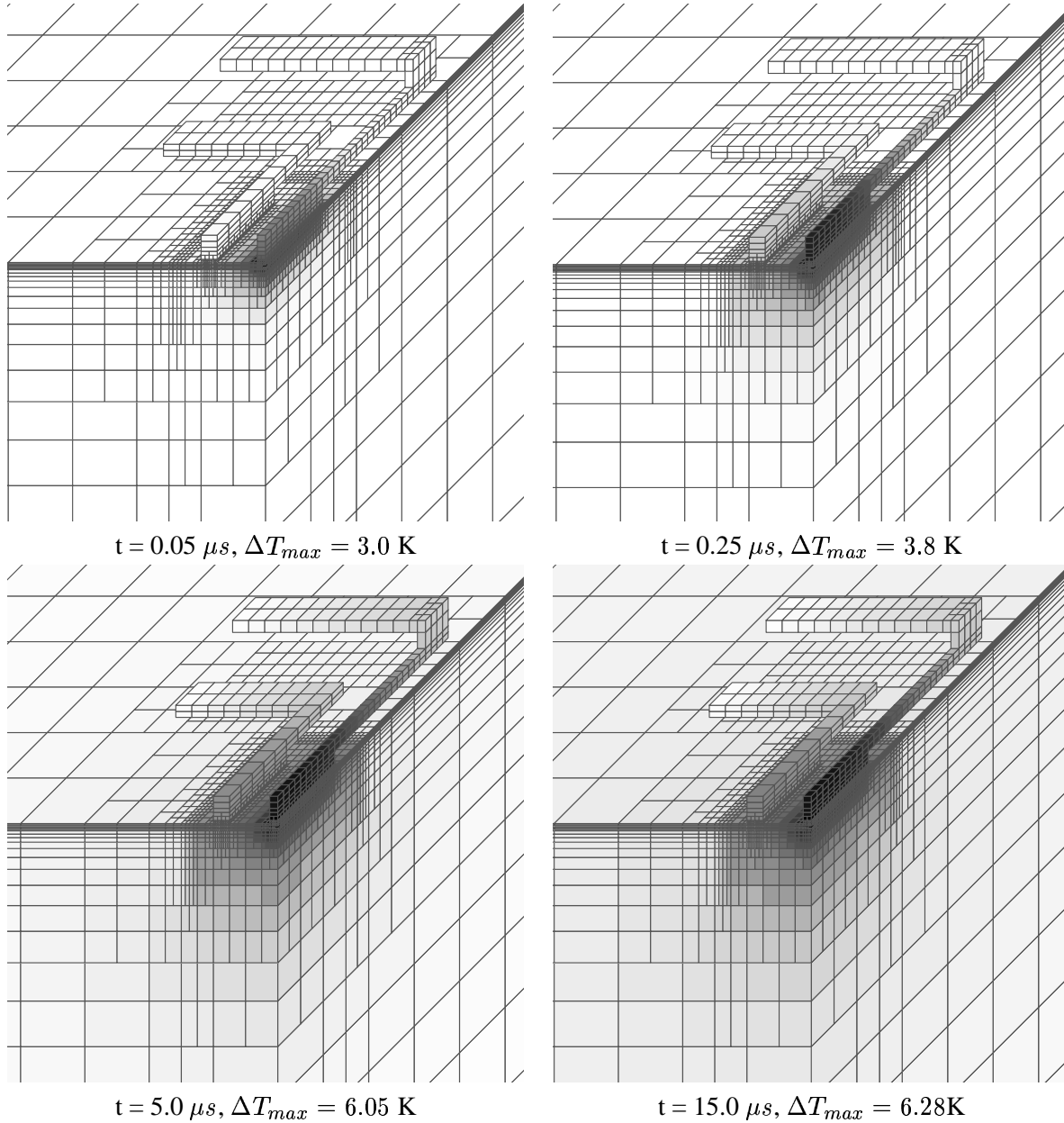t = 5.0 μs, $\Delta T_{max} = 6.05$ K   t = 15.0 μs, $\Delta T_{max} = 6.28$K

Figure 14: Four temperature distributions in a Si trench device at increasing times.