Accuracy and Time Resolution in Thermal Transient Finite Element Analysis

Roger Stout, P.E.
David Billings, P.E.
ON Semiconductor
Advanced Packaging Technology

Abstract

For reasonable accuracy in thermal transient FEA, two related, but distinct, goals must be simultaneously attained: (1) the mesh must have sufficient intrinsic temporal and spatial resolution for the problem of interest; (2) the time steps must be sized to yield proper mathematical transient behavior for that mesh. If the distinction between these needs is not fully grasped, or either point simply disregarded, errors may be very large. 1D, 2D-axisymmetric, and 3D models of a simple thermal system are compared with closed-form theory to demonstrate how (and how not) to build a successful model with optimal computational performance.

Introduction

All finite element modeling requires decisions regarding model resolution. In many, if not most, situations, we deal with what is, in the macroscopic real world, a continuous medium. Sometimes there may exist a mathematical closed-form solution for the problem we are trying to solve. More often than not, however, we must resort to a computer model to provide an approximate solution. The use of the computer model generally forces us to discretize the domain of interest. The size of the discretization may significantly affect the accuracy of the results.

In steady state or static analysis, the discretization applies obviously to the spatial dimensions of the problem. It is well understood that in regions where the gradients are steep (whether this be thermal, stress, velocity, electric field strength, or what have you), the model must have a finer grid in order to maintain adequate accuracy. Perhaps this can be ignored if the places of steepest gradients are not the places where a model's predictions are important. Otherwise, some care must be taken in sizing the grid according to the gradients present in the regions of concern. Further, unless computer resources are limitless, the finest grid found necessary at the most crucial location, cannot generally be applied to the entire model. Computer memory or time, eventually, inevitably, constrains the overall model size.

When transient analysis is required, not only is spatial discretization an issue, but also the added time dimension imposes its own discretization constraints. Again, as we make the leap from the macroscopic real world where time is continuous (so we assume), to the discretized finite element model where time comes in measured steps, the accuracy of the results may become very sensitive to the size of the steps. Unfortunately, the effects of spatial and temporal discretization interact with each other, and ultimate model accuracy must take this into account.

Semiconductor Device Thermal Transient as a Context

To put "meat" onto the "bones" of the above discussion, this paper will focus on transient thermal analysis. Even more specifically, in the semiconductor packaging business, a normal requirement is transient thermal characterization. We shall therefore use time scales and dimensions typical of what is current in semiconductor packaging. The specific models constructed for the purposes of this paper might be considered "toy" problems, but it should be readily seen how the issues apply to "real" thermal transient problems, at any scale and with any materials.

A thermal transient analysis can be broken into two fundamentally different steps.
(1) The stimulus driving the model must be analyzed, and some suitable minimum time scale for describing
the stimulus must be identified. For instance, suppose a model will be driven with several cycles of a
square-wave forcing function. A reasonable choice for time discretization might therefore be, say, one
tenth the period. Or if the duty cycle of the square wave is not equal "on" and "off," perhaps one fifth the
smaller of the two portions of the cycle. In semiconductor device packaging, we routinely desire a transient
heating curve. This is accomplished by a simple "step" application of power, taking the model from initial
response to steady state. Here the model's designed temporal resolution should be based on the fastest
response the end user will eventually require. For instance, if the shortest pulses ever applied by the user
will be one millisecond, then the model's target temporal resolution for the heating curve should be, say,
one fifth of that, or 200 microseconds.

(2) A model must now be constructed consistent with the time resolution determined in step (1). This itself
is may be broken into two parts: the spatial grid size, and the time stepping required.

a) The spatial discretization of the model must be established. In semiconductor device modeling, there is
generally a well-defined heat-input location within the model, typically the active surface of the silicon
chip. Sometimes heat input is more properly modeled as volumetric heat generation in some limited
region. In either case, the steepest thermal gradients will occur predictably in immediate proximity to the
surface or region of heat inputs. For thermal transient analysis, if there is a specific time scale at which
accurate response is desired, it is straightforward to calculate the spatial grid necessary to achieve it.

b) Given a specific grid, the transient time stepping may now be determined. This is even more
straightforward than step (2,a), having to do with the intrinsic mathematical/numerical behavior of the
modeling tool being used. ANSYS, in particular, stipulates a certain initial time step based on element size
and material properties.[1] (Be that as it may, one issue to be addressed in this paper is the implication of
this "rule," and what does and does not happen if one violates it.) The emphatic point is that time stepping
properly chosen here is not directly related to the desired time resolution of the model as identified in step
(1), above. It is related only indirectly, and only insofar as the grid is constructed properly for the time
scale of interest. An improperly meshed model will never provide accurate transient response, regardless
of the time stepping!

**Nomenclature and Definitions**

\[ A \] cross section area in direction of heat flow

\[ c_p \] specific heat

\[ k \] thermal conductivity

\[ L \] heat flow path length of interest

\[ q \] surface flux (power per unit area)

\[ Q \] heat (power)

\[ \rho \] density

\[ \alpha = \frac{k}{c_p} \] thermal diffusivity

\[ \tau, \tau_{crit} = \frac{L^2}{\alpha} \] thermal conduction mode characteristic time

\[ R = \frac{L}{k \cdot A} \] 1D thermal path resistance

\[ C = \rho c_p L \cdot A \] thermal capacitance
**Procedure**

First, we present a physically simple model for which a closed-form solution has been determined. The short-time limit of this solution will also be analyzed. Next, a lumped-parameter model will be analyzed and contrasted with the original closed-form solution. We then construct a series of ANSYS models to explore the deviations from closed-form results, as a function of both grid size and time stepping. Time step variations will be controlled entirely through the use of the initial time step size, with ANSYS "auto time stepping" enabled. The first models, though built in 3D, are functionally 1D thermal models. The second set of models cross between 1D and 3D, using the 2D axisymmetric capabilities of ANSYS, mainly to explore whether the use of axisymmetric elements changes the observations gleaned from the first set. The final models are full 3D, which again are compared with the closed-form solution (where it applies).

**Mathematical Closed-Form Models**

### 1D Surface Flux Heating of Finite Continuous Region

The physical model against which we will benchmark our ANSYS models is depicted in Figure 1. At time zero, a uniform, constant heat flux is suddenly imposed to the free surface of a uniform temperature, uniform material, spanning a width $L$. The "back" surface of the region is thermally grounded. The solution to this problem is presented here without development [2], but it is important to recognize the appearance of the specific quantity $\tau$, the "thermal characteristic time" given in the Nomenclature section as Equation 2. The name is given this term precisely because it appears in this Equation 5; that is, its identity as the "characteristic time" is precisely because it characterizes the solution to the 1D finite continuous region. Plots of this solution, along with plots of the other two theoretical models (next two sections) may be found in Figures 4 and 5.

![Figure 1 - 1D Finite Domain Schematic](image)

**Eq. 5**

\[
T(t) = 2q \frac{L}{k} \left\{ \frac{t}{\pi \tau} + 2 \sum_{n=1}^{\infty} (-1)^n \left[ \frac{t}{\pi \tau} \sqrt{n^2 \pi \tau} - n \operatorname{erfc} \left( n \sqrt{\frac{t}{\pi \tau}} \right) \right] \right\}
\]

Observations (not altogether obvious) to be made about this closed-form solution are that at short times ($t \ll \tau$), temperature rise is proportional to the square root of time. Also, as time gets very large, the final value is

**Eq. 6**

\[
T(\infty) = q \frac{L}{k}
\]

It also happens that at short times, very few terms of the infinite series are required for extreme precision, but as time begins to exceed several characteristic times (say $t > 10\tau$), an increasingly unwieldy number of
terms are required for convergence. Out to four characteristic times, though, ten terms suffices to provide double-precision accuracy.

1D Surface Flux Heating of Semi-Infinite Domain

A second theoretical model of some interest is the semi-infinite domain, shown in Figure 2. This is the equivalent of the finite-domain model (above), analyzed for only its short-time response, that is, before the heat has had time to travel to the "back" (thermally grounded) wall of the domain. The solution [3], also plotted in Figures 4 & 5, is:

Eq. 7  \[ T(t) = \frac{2q}{\sqrt{\pi \kappa \rho c_p}} \sqrt{t} \]

It can be shown rigorously that the short time limit of Equation 5 is exactly Equation 7.

\[ T_{init} = 0 \]

\[ q \]

Figure 2 - 1D Semi-infinite Domain Schematic

Step Heating of Lumped RC System

A third theoretical solution of interest is that of a "lumped parameter" model, utilizing the concepts of thermal resistors and capacitors, and is illustrated in Figure 3. The simplest possible model consists of a thermal resistor and a thermal capacitor in parallel, connecting the point of heat input to thermal ground. The obvious choice for the resistor, if modeling the finite continuous domain of Figure 1, is the steady state thermal resistance of the "slab" of material, given by Equation 3. Less obvious is the best choice of thermal mass (or capacitance), which turns out to be half the thermal mass of the slab (whose full value is given in Equation 4). For now, suffice it to say that it is because we will be comparing this response with the limiting short-time response of our ANSYS models, which we know distribute element masses equally to elements' nodes. In the most trivial possible ANSYS model of this problem, namely, 1 element, half the mass therefore will be lumped at the free surface, and half at the fixed boundary. Consequently, only half of the thermal mass is available to influence the transient response. (Similarly, even as we add nodes to the ANSYS model, the heated free surface will always be associated with just half the thermal mass of the layer of elements adjacent to the surface.) In any event, using half the slab volume for its thermal capacitance, the resulting solution to this single lumped mass problem is:

Eq. 8  \[ T(t) = q \frac{L}{k} \left( 1 - e^{-\frac{2t}{\tau}} \right) \]
Plotted in Figures 4 & 5, this is a pure exponential step response in temperature, to a step input of power. In the limit, as time goes to infinity, the familiar final value of Equation 6 results. At the short time extreme, its limiting behavior is linear in time:

\[ T(t) = \frac{2q}{\rho c_p L} t \]  

Eq. 9

Substituting in for \( \tau \) from Equation 2, and recognizing the capacitance relationship of Equation 4, yields these additional interesting alternative expressions on the short time behavior:

\[ T(t) = \frac{2q}{\rho c_p L} t \]

Eq. 10

\[ T(t) = \frac{2Q}{C} t \]

Eq. 11

Equations 10 and 11 (hence Equation 9, though obscured) are remarkable in their complete absence of dependence on the material thermal conductivity. This tells us that the initial heating of a lumped parameter (or indeed, any finite element) model will be based solely on the local thermal mass; that is, the rate of temperature increase will be in direct (inverse) proportion to the thermal mass present, before the heat has time to "conduct away." This is completely unlike the \( \sqrt{t} \) behavior that occurs in a truly continuous domain, where the heat conducts away in essence even as it arrives, as the previous closed-form solutions have indicated.
Figure 4 - Exact Solutions (linear scale)

Figure 5 - Exact Solutions (log scale)
Comparison of Continuous and Lumped Models

As mentioned already, the solutions to the three preceding models are plotted in Figures 4 and 5, normalized as \( \frac{k}{qL} T(t) \). Clearly, the two "continuous domain" solutions are identical at short times. It is clear that qualitatively, this \( sqrt(t) \) short-time response is "pretty good" through something like 40% of the characteristic time. Quantitatively, its deviation from the exact solution of the finite continuous domain (Equation 5) does not exceed 1% until 0.32 \( \tau \); at 0.5 \( \tau \), the error has grown to 4.25%; and at \( \tau \) to 17%.

At the other end, as steady state is approached, the two "finite domain" solutions are identical (continuous and lumped). Of at least as most significance, however, is that the continuous response is always "faster" (that is, its curve lies above the other, meaning it reaches any given level of response sooner). And at the short-time end (as Figures 4 and 5 clearly illustrate), the lumped solution simply "disappears" in comparison to the continuous solution. Quantitatively, their ratio (lumped over continuous) is in fact \( \sqrt{\frac{\pi t}{\tau}} \); so, for instance, at 0.1 \( \tau \) the lumped response is but 56% of the continuous response. Since this ratio is itself proportional to \( sqrt(t) \), then at 0.001 \( \tau \) the ratio will be only 5.6%, and so forth.

The importance of these theoretical solutions is actually twofold in our present circumstances. Not only do they provide the benchmarks against which the finite element solutions may be compared, but also they provide the crucial insight we need in order to understand the intrinsic transient limitations of individual finite elements. Viz, whenever a thermal conduction element is subjected to a disturbance at one of its boundaries, the disturbance propagates to the other side of the element on a time scale of order given by Equation 2. Consequently, at time scales shorter than this order, it is unreasonable (or more pointedly, extremely incorrect) to assume any validity to the transient details. Even though the local equilibrium solution for such an element might be perfect, its transient response in getting there may have nothing to do with the real physics. To wit, short time linear response is vastly different than short time \( sqrt(t) \) behavior.

Building the Finite Element Models

Problem Scaling

Before engaging in any model construction, one pitfall to avoid is the numerical lower bound for the tool in use. Regarding transient analysis, in particular, the ANSYS Thermal Analysis Guide cautions the user to avoid time steps smaller than 1E-10 [1]. In the context of discrete semiconductor package modeling, it is not uncommon to be concerned with transient response to microsecond duration pulses. Yet more extreme are state of the art logic/optical devices, wherein speeds in excess of 1GHz are becoming common. Therefore, we can immediately make the decision that "seconds" are in general an unwise choice for the time scale of semiconductor thermal transient analysis. Microseconds make a more likely first choice, which is the scale adopted for the models described in this paper. Likewise, SI units of meters cause instant problems when modeling devices whose surface features are measured in microns (1E-6 meters). Typical silicon "chip" thicknesses range from 200 to 400 microns, and implanted junction depths at the surface are typically a few microns. Convenient spatial units for these analyses are, therefore, microns.

Throughout this paper, the following standard material properties for silicon will be used, and are presented here both in SI units and in the rescaled units necessary to avoid the small-quantity problems inherent in our models:

- \( k \) thermal conductivity, 100 W/m/K = 1E-10 J/microsec/micron/K
- \( \rho \) density, 790kg/m^3 = 7.9E-16 kg/micron^3
- \( c_p \) specific heat, 2400 J/kg/K = 2.4E3 J/kg/K
- \( \alpha \) thermal diffusivity = 52.74 micron^2/microsec
In addition, our models will utilize the following two values typical of discrete semiconductor devices:

\[ L \text{ typical silicon chip thickness, 15 mil} = 381 \text{ micron} \]
\[ q \text{ typical surface heat flux 1E6 W/m}^2 = 1E-12 \text{ J/micron}^2/\text{microsec} \]

With these values taken all together, typical temperature rise for a silicon chip is a few Kelvins. Thus our models' natural temperature units will remain K.

**1D Surface Flux Heating of 3D Finite Element Models**

Figure 6 shows the basic structure of the first series of ANSYS models. Although these are true 3D models insofar as ANSYS is concerned, all of these first models are in essence exact 1D models. They consist of a single row of elements aligned along the Y-axis, with grounded temperature nodes at one end (at \( y=381 \)), and uniform heat flux input at the other end (\( y=0 \)). All other external surfaces are adiabatic. Due to the symmetry of the boundary conditions, there are no gradients in the X or Z directions. In every case, we track the heated surface temperature for comparison with the closed-form solutions. What we vary in these first several models are the element sizes, ranging from the smallest elements of 5.08 micron (and 75 total elements), to a single element 381 microns large. In the various figures, these models will be identified as "e=5 unif", "e=7 unif", etc. Now consider the overall transient properties of this model. According to Equation 2, the characteristic time of a slab of silicon 381 microns thick is about 2800 microseconds \((381^2/52.74)\). We can therefore expect reasonably close \( \sqrt{t} \) behavior to continue through approximately 1000 microseconds (with 1% accuracy or better).

So, why the minimum element size of 5.08 microns? Because following the procedure outlined in the introduction, we have to make a decision on the fastest time response of interest to us in this silicon semiconductor model. Our immediate goal is to generate the traditional "single pulse heating curve," where the abscissa is the pulse length. Supposing that most customers will drive our product slower than 1MHz, we arbitrarily choose a desired fastest time response of 0.5 microseconds (that being the pulse width of a 50% duty-cycle 1MHz wave train). Therefore our ANSYS model must provide good accuracy from times as short as 0.5 microseconds out through steady state. Again, Equation 2 may be used, solving for the element size required for this response time. The result is 5.13 microns. For the convenience of integral multiples of our minimum element size, we recognize that 75 elements of 5.08 microns will fit exactly in the 381 micron solution domain (Figure 6). Varying the uniform element size, we generate these specific models:

### 1D Uniform Mesh Models Analyzed

<table>
<thead>
<tr>
<th>model designation</th>
<th>element size (microns)</th>
<th># of elements</th>
<th>response time (microsec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>e=5 unif (Fig. 6)</td>
<td>5.08</td>
<td>75</td>
<td>0.49</td>
</tr>
<tr>
<td>e=8 unif</td>
<td>7.62</td>
<td>50</td>
<td>1.1</td>
</tr>
<tr>
<td>e=13 unif</td>
<td>12.7</td>
<td>30</td>
<td>3.1</td>
</tr>
<tr>
<td>e=25 unif</td>
<td>25.4</td>
<td>15</td>
<td>12.2</td>
</tr>
<tr>
<td>e=51 unif</td>
<td>50.8</td>
<td>8</td>
<td>49</td>
</tr>
<tr>
<td>e=127 unif</td>
<td>127</td>
<td>3</td>
<td>310</td>
</tr>
<tr>
<td>e=381 unif</td>
<td>381</td>
<td>1</td>
<td>2800</td>
</tr>
</tbody>
</table>
A similar series of models are constructed where the X-Z dimensions of the elements are constant throughout, but the Y-dimension increases at a constant ratio, beginning with a 5.08 micron element at the heat flux end, and proceeding up to the fixed temperature end. It is also interesting to repeat the smallest element multiple times before beginning the fixed ratio expansion. These models are identified in the figures as "e=5 R=r.n", where "r" is the fixed growth ratio (either 2 or 3), and "n" is the number of repeated 5 micron elements at the heat input end. Thus "e=5 R3.2" means there are 2 identical 5 micron elements to receive the heat flux, after which the element sizes increase in the ratio of 3 to 1. Utilizing a constant growth ratio obviously vastly reduces the model size, as Figure 7 (the "R3.1" model) illustrates. The following table shows the number of elements resulting in each of these "constant growth" models. Note they all have the same minimum element size of 5.08 microns, hence the same theoretical response capability of 0.5 microsec.
With these various meshing schemes, we thus have 12 different 1D models to explore. The secondary variable in the exploration is the initial time step. According to the ANSYS Thermal Analysis Guide [1], the initial time step that should be used in a transient analysis is

\[ \tau_0 = \frac{x^2}{4\alpha} \]

where \( x \) is the element size in the direction of maximum gradient of the smallest elements of interest. This recommendation is apparently aimed primarily at nonlinear analyses, where solution oscillations are possible. In any event, this recommended value is precisely one fourth the value we originally used to determine the minimum element size required to meet our time resolution/accuracy objectives, namely, the characteristic time of an element of dimension \( x \) according to Equation 2. The question that naturally arises is, what happens if you use a different factor than four? Indeed, does it do any good to drive an element at four times its theoretical response capability? What, if anything, do you lose if you drive it more slowly? What, if anything, do you gain if you drive it faster? To answer these questions, we define our minimum step size according to

\[ \tau_{0b} = \frac{x^2}{b \cdot \alpha} \]

where \( b \) is an adjustable factor. ANSYS says to use a value of 4. We will try (in most of these models) values of 1, 2, 4, and 10. In the finest mesh model, we will also use values of 100 and 1000, and in the coarsest (1 element) model, we will use values of 4, 10, 30, 100, 300, and 1000. (In this latter case, values of 1 and 2 result in initial time steps larger than 1000 microseconds, and since one of our peripheral goals is to examine the "short time" response where the sqrt(\( t \)) behavior should be approximated, there isn't any point in starting with a larger initial step.) In the various figures, the initial time step factor will be indicated as "\( b=1 \)" or "\( b=4 \)", etc.

### Surface Flux Heating of 2D Axisymmetric Finite Element Models

The next set of ANSYS models to be explored is built using 2D axisymmetric elements. There are two reasons for pursuing these models. First, we should like to confirm that a mere change in element type does not alter any conclusions already drawn from the strictly 1D models. Second, we want an initial assessment of the influence of 3D effects, and how "soon" they disturb a locally 1D problem. (In other words, at short enough times, any surface heating model should "act like" a 1D model for some brief period based on the dimensions of the surface. We should like to verify this in our FEA tool.) We will not bother, however, with the variation in the initial time step factor, presuming that we have learned enough about it from the 1D models.

To answer the question of whether the element type alone significantly modifies any previous conclusions, we define a 2D axisymmetric model, which represents a strict 1D problem (i.e., the boundary conditions will result in purely axial gradients, and no radial gradients). The model is a simple cylinder (Figures 8 through 11), uniformly heated across one end, grounded at the other end, and adiabatic on the exterior cylindrical wall. External domain size remains 381 microns in the direction of heat flow, and an arbitrary
radius of 381 microns is chosen for the radial boundary. (In the semiconductor device context, this model would approximate in circular geometry a medium-small, square piece of silicon, about 675 microns, or 27 mils, on a side.) Two different uniform mesh densities are examined (the e=5 and e=8 element sizes used in the previous 1D models).

To address the second issue, we "spoil" the purely axial problem by applying the uniform heat flux across only the inner half-radius of one end (i.e., 25% of the axial area). We then observe the transient response of the center surface node. In this second case, the meshing will be somewhat different in the radial and axial directions. Given that "edge effects" of the heated surface may be important, we will mesh these models in two ways. First, entirely uniformly in both axial and radial directions. Second, graded in the axial direction exactly as before (constant element ratio from one row to the next, again using R=2 and R=3 ratios), and split between uniform and graded in the radial direction. A uniform mesh will be maintained throughout the heated radius (out to 191 microns, the inner half of the domain), at which point grading will begin and continue at constant ratio until the external radius of 381 microns is reached.

Figures 8 through 11 depict four variations of this model (as indicated in the table following). An identical "model designation" scheme as in the 1D models is used. The number of radial elements is listed in the table as "x+y", where "x" is the number of uniform elements in the inner half (always heated), and "y" is the number of uniform or graded elements in the outer half (heated or unheated). Total element count is, of course, the product of the axial and radial counts.
Figure 10 - 2D-Axisymmetric Graded Model (e=5 R2.1 half heat)

Figure 11 - 2D-Axisymmetric Graded Model (e=5 R3.1 half heat)

2D-Axisymmetric Models Analyzed

<table>
<thead>
<tr>
<th>model designation</th>
<th># of radial elements</th>
<th># of axial elements</th>
<th>total # of elements</th>
<th>response time (microsec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>e=5 unif full heat (Fig. 8)</td>
<td>38+37</td>
<td>75</td>
<td>5625</td>
<td>0.49</td>
</tr>
<tr>
<td>e=8 unif full heat</td>
<td>25+25</td>
<td>50</td>
<td>2500</td>
<td>1.1</td>
</tr>
<tr>
<td>e=5 unif half heat (Fig. 9)</td>
<td>38+37</td>
<td>75</td>
<td>5625</td>
<td>0.49</td>
</tr>
<tr>
<td>e=8 unif half heat</td>
<td>25+25</td>
<td>50</td>
<td>2500</td>
<td>1.1</td>
</tr>
<tr>
<td>e=5 R2.1 half heat (Fig. 10)</td>
<td>38+6</td>
<td>7</td>
<td>308</td>
<td>0.49</td>
</tr>
<tr>
<td>e=5 R3.1 half heat (Fig. 11)</td>
<td>38+4</td>
<td>5</td>
<td>210</td>
<td>0.49</td>
</tr>
</tbody>
</table>
Surface Flux Heating of 3D Finite Element Model

The final ANSYS models are full 3D models with 3D gradients. They relate closely to the axisymmetric models of the preceding set, in that one quarter of the "free" end of the model receives the heat input. Again, there are two purposes in this pursuit. The first is to replicate in rectangular geometry the same effects of short time and "center" position within the model, as were explored already in the 2D axisymmetric cases with the central 25% of the surface heated. Second, now that a full 3D model is being analyzed, we feel the attraction of reducing the total element count from what would result were a uniform mesh to be utilized.

As in the 2D axisymmetric model, we will use a 381 micron dimension along the Y-axis, which is the primary heat flow direction (surface heat on the inner 25% of the area at one end of the model, grounded at the other end of the model). In the transverse directions (X and Z axes), the domain width is again 381 microns. Both XY and YZ surfaces are adiabatic. Since the heat flux is not symmetric about the geometric center of the rectangular volume modeled, the model is thus "quarter symmetry." In the real world, it would represent a silicon chip 30 mils square by 15 mils thick, fairly typical for a small, discrete device.

Driven by a desire to reduce total element count without unwittingly sacrificing overall accuracy, we also modify slightly the meshing scheme yet again, as compared to the 2D axisymmetric models. Consider that a 381 micron cube, meshed uniformly with 5.08 micron elements (75 on a side), would contain 421,875 elements! Increasing element size slightly to just 7.62 microns, for instance, reduces the element count down to 125,000, but it sacrifices transient response resolution by a factor of 2. Clearly some scheme of grading the mesh is worth considering: even a modest growth ratio of 1.2 requires a mere 16 elements, beginning with 5.08 micron, to fill a 381 micron space. If done in all three spatial directions, this yields just 4,096 elements, a full two orders of magnitude smaller than the uniform mesh based on the starting element dimension!

We would like to demonstrate, therefore, that a graded mesh can still provide acceptable accuracy. Having learned something by now of the limitations of the constant ratio models, however, here in the 3D models we will explore additional (smaller) growth ratios of 1.2 and 1.5, along with the ratios of 2 and 3 analyzed in the previous 1D and 2D-axisymmetric models. Also, the meshing scheme will not be quite so simplistic as perhaps hinted at in the preceding paragraph. Reasoning that the response at the center of the heated region behaves as if it is essentially 1D, but assuming that "edge effect" inaccuracies similarly affect only the "edge" elements of the heated region, we will use the smallest element dimensions both at the center (where we care the most to observe the response), and also at the edge of the heated region (half the distance from center to exterior wall). The grid will then grow at constant ratio outward from the center, and inward from the edge of the heated region, until the two meshes meet at the midpoint of the heated area. Likewise, the mesh will grow at constant ratio outward from the heated edge through the entire unheated area until the outer dimension is reached. Careful study of Figure 12, built with the 1.2 element ratio, should clarify the mesh grading scheme. We also need to tweak our model notation due to the use of non-integral ratios, as follows: "r=x:1", where "x" is the constant growth ratio (and may be expressed as a decimal number). Thus, Figure 12 depicts the "e=5 r=1.2:1" model. Because we continue to be interested in 0.5 microsecond response capability, all meshes will begin with the same 5.08 micron smallest elements.

Only four distinct models (with the aforementioned growth ratios) are analyzed for this 3D exploration, as listed in the following table. Again we use only the ANSYS recommended initial time step factor of 4. Transverse element counts (X and Z) are again expressed as "x+y" where "x" is the element count in the inner, heated model region, and "y" is the element count in the outer, unheated region. The total element count is the product of the Y-axis count and the square of the transverse count. If even only the 1.2 ratio model produces acceptable accuracy, still we've reduced the model size by a factor of nearly 600!
Full 3D Models Analyzed

<table>
<thead>
<tr>
<th>model designation</th>
<th># of transverse (X-Z) elements</th>
<th># of axial (Y) elements</th>
<th>total # of elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>e=5 r=1.2:1 half heat (Fig. 12)</td>
<td>18+12</td>
<td>16</td>
<td>14400</td>
</tr>
<tr>
<td>e=5 r=1.5:1 half heat</td>
<td>12+8</td>
<td>9</td>
<td>3600</td>
</tr>
<tr>
<td>e=5 r=2:1 half heat</td>
<td>10+6</td>
<td>7</td>
<td>1792</td>
</tr>
<tr>
<td>e=5 r=3:1 half heat</td>
<td>8+4</td>
<td>5</td>
<td>720</td>
</tr>
</tbody>
</table>

Analysis Results and Discussion

1D Surface Flux Heating of 3D Finite Element Models

Influence of the Initial Time step

The results of the 1D models can be presented in a couple of ways. For a given element size (that is, meshing scheme), the family of response curves with varying initial step size may be plotted together. Figures 13-15 present three of these families, for the e=8, e=13, and e=25 models, for a broad assessment of the similarities and trends. Now consider Figure 16, which zooms in on the 0.1 to 10 microsecond portion of Figure 13. Undeniably interesting things are happening right around the critical element resolution time, $\tau_{crit}$. For $L=7.62$ microns, $\tau_{crit}$ is 1.1 microsec, and the b=4 ANSYS recommended initial time step is 0.27 microseconds. Clearly the solution at the first time step, even using the recommended ANSYS value, falls far short of the exact solution (by a factor of about 3!). Indeed, the model solution does not come within 10% of the exact solution until $2\tau_{crit}$. 
Figure 13 - 1D $e=8$ Uniform Mesh Heating Curves

Figure 14 - 1D $e=13$ Uniform Mesh Heating Curves
Figure 15 - 1D $e=25$ Uniform Mesh Heating Curves

Figure 16 - 1D $e=8$ Uniform Mesh Zoomed
Another, perhaps more useful way, to see these results is shown in Figure 17, an error plot (with respect to the exact solution) of the same four "b value" trials of the e=8 model. It is evident that changing the initial time step makes a difference, though regardless of the initial step, the solution eventually "catches up with" (and in fact, overshoots) the exact solution -- after about 50 $\tau_{\text{crit}}$. On the one hand, $b=1$ delays the 10% error point until about 4 $\tau_{\text{crit}}$, or a factor of 2 later than the recommended $b=4$. On the other hand, $b=10$ shifts the 10% error point earlier by a factor of 2. Will that improvement continue if $b$ is increased further?

![Figure 17 - 1D e=8 Uniform Mesh Errors](image)

To answer this, consider Figures 18 and 19, the heating curve and error plots of the e=5 model, which was solved for additional $b$ values of 100 and 1000. Particular attention should now be paid to the "linear" response curve in Figure 18 (which also was present in earlier figures, but not discussed). Each "linear" response is that expected (according to Equation 9) of a lumped RC model having the dimensions of the individual elements of the corresponding uniform mesh (where $C$ is based on half the element volume). It begins to appear that it is this limit being approached as $b$ increases, not the exact (continuous domain) solution proportional to $\sqrt{t}$. Consequently one concludes that if the mesh is fixed, continuing to reduce the initial time step is never going to get us closer to the exact solution.
Figure 18 - 1D e=5 Uniform Mesh Heating Curves

Figure 19 - 1D e=5 Uniform Mesh Errors
Short Time Behavior of FEM is Linear, NOT sqrt(t)

Lest there be any lingering doubt, Figures 20 and 21 present corresponding families of "b" curves for the extreme, 381 micron, single element, finite-element model (which was analyzed at b values of 4, 10, 30, 100, 300, and 1000). Note that the intrinsic response time of a 381 micron element is precisely that of the 381 micron domain itself, which we recall is about 2800 microseconds. The main difference in Figure 20 (as compared to Figure 18) is that rather than plotting just the linear short-time limit of the RC model, we have plotted the actual entire lumped RC solution (Equation 8, identified as "exp(t)" on the chart). Figure 21, the error with respect to the exact solution, also makes it clear that further decreases in initial time step are pointless, at least if the objective is to accurately simulate a surface-heated, continuous domain. With the recommended b=4 initial step, the error falls below 10% at about $\tau_{\text{crit}}$ (slightly earlier than for the e=5 model), but reducing the initial step beyond that provides rapidly diminishing returns. On the other hand, Figure 22, the ANSYS error with respect to the lumped RC solution, demonstrates unequivocally that this is the limiting case for the finite element ANSYS model. The shorter the initial time step, the earlier the ANSYS model hits the lumped-RC solution perfectly. Indeed, with b=4, we fall below 10% error a little earlier than $\tau_{\text{crit}}$, and we can move the 10% error point to as early a time as we desire simply by decreasing the initial time step (without modifying the mesh at all). This also affirms the earlier presumption that only half the thermal mass of the element is effectively responding to the heat input.

Figure 20 - 1D Single Element Heating Curves
Figure 21 - 1D Single Element Errors as Compared to Continuous Exact Solution

Figure 22 - 1D Single Element Errors as Compared to Lumped Exact Solution
Influence of the Element Size

The 1D analyses, looked at somewhat differently, also permit us to assess the deep significance of the
element size. Rather than plot various b-value experiments on a single chart for a given mesh, we can plot
all the different mesh results for a fixed b-value (which is not to say a fixed initial time step, since the
initial time step continues to be a function of the element size). Figure 23 is such a plot, taken from the
recommended b=4 results of each uniform grid analyzed. We have chosen a linear vertical scale for once,
to emphasize that the overall solution behavior suffers noticeably in detail as the element size approaches a
significant fraction of the overall domain. The short time behavior, on the other hand, is more easily
perceived on log-log plots (as before); Figure 24 is the same data as Figure 23, so plotted. What stands out
in Figure 24 is that as the element size decreases, the FEM solution actually tracks very nicely the $\sqrt{t}$
behavior of the exact solution. It is now clear that so long as the initial time step is adjusted to go along
with the element size, we can get arbitrarily good agreement with the exact solution to any time scale we
desire -- all we have to do is make the mesh suitably fine. It is not even so critical what specific b-value is
chosen, though the ANSYS recommendation of 4 works well. Using that value, Figure 25 (solution error)
shows that across all element sizes, the error is somewhere between 10 and 15% at $\tau_{\text{crit}}$; so with an initial
time step one fourth of that value (i.e. b=4), we are controlling the error in a reasonably consistent manner.

Building on the previous discussion, if we want to get better accuracy than 10% at some specified time, the
mesh must be refined to a smaller $\tau_{\text{crit}}$, not the initial time step reduced to a smaller fraction of $\tau_{\text{crit}}$
(though for sure the initial time step must be reduced in an absolute sense, to go along with the refined
mesh). For instance, to achieve 1% model accuracy at any selected target time, Figure 25 suggests that
$\tau_{\text{crit}}$ of the mesh must be something like one-tenth of that time value.

![Figure 23 - 1D b=4 Uniform Mesh Heating Curves, Linear Scale](image-url)
Figure 24 - 1D b=4 Uniform Mesh Heating Curves, Log Scale

Figure 25 - 1D b=4 Uniform Mesh Errors as Compared to Exact Continuous, Zoomed
Influence of Mesh Grading

Heretofore we have been concerned with uniform meshes. Clearly the short time response of a model is intimately tied to the size of the elements, but do all the elements in model have to be as small as the ones we choose in specific regions where we care the most? Can we achieve similar accuracy in a transient thermal analysis with fewer elements, by creating a graded mesh? In Figure 26, we have collected together the $b=4$ results of all the $e=5$ "graded mesh" models. The common feature, obviously, is that the element receiving the surface heat input is identical in all these models, having a size of 5.08 microns (and a $\tau_{\text{crit}}$ of 0.49 microseconds). What differs is the ratio at which the element size increases from this layer as we move away from the heated surface (1-to-1, 2-to-1, and 3-to-1), and in the 3-to-1 ratio, we have also made a slight variation where the 5.08 micron layer is repeated a few times before the mesh begins to grow.

Figure 26 shows immediately that all these models behave remarkably similarly to each other. That is, they all begin with exactly the same amount of deviation below the exact solution, and after about $2 \tau_{\text{crit}}$, they all follow the $\sqrt{t}$ behavior of the exact solution fairly closely. However, if we plot the actual error (as a percent of the exact solution, as in Figure 27) we see some subtle details of interest. Again, the initial error is identical in all six models, but the 10% error limit is reached over a somewhat large range of times. The R3.1 model, in particular, doesn't clear 10% error until after several $\tau_{\text{crit}}$. Worse, it hovers between 5% and 10% error for a very long time; if it weren't for the grounded boundary condition and finally reaching the (apparently correct) steady-state value, it might never have achieved 1% accuracy at all! The R2.1 mesh does considerably better, though 3-5% accuracy might be considered its realistic limit. On the other hand, it is somewhat incredible that 5 (R3.1) or 7 (R2.1) total elements can get within 10% or even 3% of the true answer, over four orders of magnitude of time.

Figure 26 - 1D $b=4$ Graded Mesh Heating Curves
Finally, Figure 28 (same data as Figure 27, but with an expanded vertical scale) shows a little more clearly the effect of the repeated layers of the $\tau_{\text{crit}}$ elements. Even in the relatively poor 3-to-1 ratio model, the repetition of the smallest elements has a significantly beneficial effect -- at least for a while. Indeed, the error plots for these several variations follow the uniform-mesh error; the more thin elements, the longer the error tracks the uniform mesh error (no big surprise). Once past the thin uniform layer, however (think of $\tau$ for the uniformly meshed layer as increasing with the square of the number of repeated elements), the error actually grows again for a while, until the grounded boundary finally reins it back in. The R2.1 model, by contrast, is at least monotonically decreasing in error with time. This raises a question beyond the scope of this paper: can this error be controlled by manually controlling the time steps (as opposed to relying on the ANSYS auto-time stepping algorithm)? There is independent evidence to support the use of 3-to-1 element growth ratios in transient thermal modeling (with far better accuracy than observed here), but it is beyond the scope of this paper to try to reconcile the difficulty [4]. In any event, we will analyze two additional growth rates smaller than 2-to-1, in the subsequent 3D models.
Figures 29-31 present the results of the several 2D-axisymmetric models solved. What is clear from Figure 29 is that the short-time behavior of all the models is similar, and all the models with identical smallest elements have identical starting errors. In fact, careful comparison with the uniform mesh 1D curves in Figure 24, shows that the 2D-axisymmetric short-time solution (and error) is identical to that of the 1D models. Figure 29 also shows us that the "full heat" and "half heat" models diverge from each other at somewhere around 100 microseconds. Unfortunately, since we have no exact solution to the "half heat" problem, the error plots of the "half heat" models in Figure 30 don't mean much after that point. (Looking at Figure 30, we see more precisely that all the half-heat models actually begin to increase in error at 60-70 microseconds.) This is considerably before the end of the $\sqrt{t}$ behavior of the purely axial solution. Since we're looking at the temperature at the center of the heated disk, this tells us that the edge of the disk (at a radius of 191 microns) influences the center temperature much sooner than we would have guessed based on pure 1D conduction. (Equation 2 gives a 700 microsec characteristic time for a 191 micron conduction length.) Presumably this is explained by the radically 3-dimensional heat spreading occurring in this model. The other unsurprising observation to be made from Figure 29, is that the final temperature reached at the center of the disk, when only 25% of the end of the cylinder is heated, is considerably lower than when the entire disk is heated. Figure 31 zooms in on the final values of the half-heat models. Although we do not have a closed-form reference for that final value, the models do all agree with each other within 1.7%.

Figure 30 shows us detail in the error. We see the degree of improvement all the $e=5$ models hold over all the $e=8$ models, as before, at the intrinsic element response time scale. Also again, obviously, the error in the "half heated" models grows rapidly once the edge effects of the heated area come into play. And yet again, in a fashion quite similar to that of the 1D graded meshes (Figures 27 and 28), the error in the R3.1 model struggles to get below 10%; the error in the R2.1 model dips briefly as low as 3%, but never better.
Figure 29 - 2D-Axisymmetric Heating Curves

Figure 31 - 2D-Axisymmetric Final Values
Surface Flux Heating of 3D Finite Element Model

Figures 32-24 present the results of the four 3D models solved. Again, unfortunately, we have no exact solution good at steady state, so the error plots of Figure 33 don't mean anything after about 100 microseconds. (We can say that between the four models, the final values agree within 1.2%. Figure 34 zooms in on those values.) However, as with the 2D-axisymmetric model, we continue to observe that the short-time solution for the 1D problem continues to apply perfectly well, at least until 3D edge effects come into play. Even the difference between rectangular and circular domain shapes makes but a small change in the steady state result for the "half heat" models, as comparison of Figures 31 and 34 demonstrates. (We no doubt could have made them even more similar by following the usual routine of making the axial areas equal, that is, by shrinking the X and Z dimensions to 338 microns, and heating only the inner 169 microns.)

Finally, the error plots of Figure 33 simply reaffirm what we've seen before in the 2:1 and 3:1 growth models. If 10% accuracy is adequate, 3:1 gets you there with very few elements. If 3-5% accuracy is your goal, 2:1 is necessary. Moving beyond that, if 1% accuracy is the goal, a ratio of 1.5:1 comes close (it actually bottoms out at 1.1% in this model). Amazingly, the "finest" mesh considered in this 3D model, 1.2:1 (which, recall, had only 3% of the total elements that would have been present in a uniform mesh with the same base element size), bottoms out at 0.1%, at about 35 microseconds. Compare this with the 1D b=4 e=5 uniform solution (refer to Figure 27), which although it actually crosses through zero error at about 10 microseconds, also dips negative (to as much as -0.7%) as late as 200 microseconds. In other words, it is not clear that the 1.2:1 mesh actually would be any less accurate than a uniform mesh at all, though its short-time response might be somewhat more sluggish.
Figure 33 - 3D "Half Heat" Errors as Compared to Exact 1D Solution

Figure 34 - 3D "Half Heat" Final Values
Conclusion

Short time response of surface heat flux driven, continuous media thermal problems, is mathematically characterized by temperature changes on the order of $\sqrt{t}$. On the other hand, finite element models, akin to lumped parameter RC models, are characterized by linear behavior in the short time limit.

Therefore, it is impossible to force a finite mesh to respond correctly to surface disturbances, at time scales below its intrinsic characteristic time, which is dictated by its finest elements.

Following the ANSYS guidelines for initial time stepping, ensures only that the model will be responding reasonably well to its own intrinsic capability and mathematical response (which is linear), not necessarily to the quantitative accuracy requirements of the analyst. The only way the analyst can guarantee correct and accurate short-time behavior of a model, then, is to be sure that the spatial grid is sized consistent with his needs, and to use initial time stepping in turn consistent with that properly designed mesh. If the initial time step is vastly larger than the ANSYS recommended value, the resulting errors may be large; however, if the initial time step is vastly smaller than the recommend value, there is no corresponding improvement in the model's ability to correctly respond to correspondingly briefer disturbances.

The starting point for a correctly constructed grid, then, is the smallest element size needed to produce the desired speed of response, in the area of interest. Moving away from that point in the model, the choices are broad. If overall model size is an issue, a graded mesh may be used to great advantage. However, the growth rate of the grid has a direct impact on the resulting accuracy of the transient solution from initial excitation through steady state, at least presuming the use of the ANSYS AUTOTS (auto time stepping) algorithm. As it is a great convenience to thermal transient analysis, we have presumed its use throughout this paper, and in the normal course of modeling we pursue in semiconductor device characterization. We find that even a modest growth ratio of 1.2:1 (which over large domains results in vast reductions in model size) is essentially as good as a uniform mesh. A growth ratio of 2:1 may do no better than 5%, and a growth ratio of 3:1 may yield only 10% accuracy. However, the extreme reductions in model size resulting from these larger ratios, should not be overlooked. Many situations may not require better accuracy (and bear in mind that this accuracy is over the entire transient time domain of the problem, which may cover several orders of magnitude).

One issue touched upon, but not really dealt with in this paper, is the intrinsic mathematical behavior of volumetric heat driven thermal transient problems. These turn out to be linear at the short time limit. There is therefore the intriguing possibility that for volumetrically heated problems, a single, properly designed grid, may actually be suitable for arbitrarily short thermal transient response. In the real world, surface heating may itself be simply a convenient approximation to an even smaller scale, truly volumetric, heating situation. In other words, it is conceivable that a grid, shrunk to a sufficiently fine level of detail to reproduce theoretical "surface" behavior, may be shrunk too far -- for at some point, the real physics says the even shorter-time behavior returns to a linear model! So in semiconductor thermal transient modeling, in particular, the analyst must also keep in mind the lower bound of what is actually known about his device construction, and the nature of the structures responsible for heat generation within the device.

References

[1] ANSYS 5.7.1 Help, Thermal Analysis Guide, Section 3.5.3.2, SASIP 2001