Efficient Transient Electrothermal Simulation of CMOS VLSI Circuits under Electrical Overstress *

Tong Li
Analysis Product Division
Avant! Corporation

Ching-Han Tsai **Sung-Mo (Steve) Kang**
Dept. of Electrical and Computer Engineering
Univ. of Illinois at Urbana-Champaign

Abstract

Accurate simulation of transient device thermal behavior is essential to predict CMOS VLSI circuit failures under electrical overstress (EOS). In this paper, we present an efficient transient electrothermal simulator that is built upon a SPICE-like engine. The transient device temperature is estimated by the convolution of the device power dissipation and its thermal impulse response which can be derived an analytical solution of the heat diffusion equation. New fast thermal simulation techniques are proposed including a regionwise-exponential (RWE) approximation of thermal impulse response and recursive convolution scheme. The recursive convolution provides a significant performance improvement over the numerical convolution by orders of magnitude, making it computationally feasible to simulate CMOS circuits with many devices.

I. Introduction

Smaller devices, higher packing density and rising power consumption lead to dramatic temperature increases in deep submicron VLSI circuits. Considering that many IC failure mechanisms such as electrical overstress and electrostatic discharge (EOS/ESD), electromigration and hot carrier phenomenon are strongly dependent on operating temperature [1][2], it is essential to perform accurate temperature simulation and study its impact on circuit performance and reliability before a design is committed to silicon fabrication.

Extensive research has been conducted on electrothermal simulation for VLSI circuits. Two distinct areas of interest are full chip-level [3] and transistor-level [4][5] electrothermal analyses. For chip-level applications, the steady-state temperature profile of the chip is needed to assess the impact of circuit temperature on timing, power consumption and reliability. On the other hand, transient thermal response is the focus of transistor-level electrothermal simulation, which is critically important for understanding the behavior of CMOS I/O circuits under electrical overstress (EOS) [2]. The device current during an EOS event can be much higher than several hundreds of mA, causing the device temperature to rise by hundreds of degrees, with steep temperature gradients within the range of micrometers from the heat source [6]. CMOS I/O circuit analysis demands a tightly-coupled electrothermal simulator that can compute the transient electrothermal response of the circuit.

Traditionally, 2D device simulators such as MEDICI [7] are used to study device self-heating effects [6][9]. The computation cost of these simulators makes it infeasible to extend their applications to circuit simulations. Incorporating electrothermal simulation capability into an electrical simulation environment such as SPICE requires both accurate temperature-dependent device models and fast thermal response computation. This work describes such an efficient electrothermal simulator for CMOS I/O circuit simulation. The thermal model is derived from the analytical solution of the 3D heat diffusion equation. Techniques including regionwise-exponential approximation and recursive convolution are proposed to reduce the cost of device temperature evaluation at each time step down to $O(1)$ from $O(n^2)$ for full numerical convolution, where n is the number of time steps.

*This research was supported in part by Semiconductor Research Corp. (SRC97-DP-109), JSEP (N00014-97-J1270) and Rome Laboratory (F30602-97-1-0006).

The remainder of this paper is organized as follows. In section 2, we present the high current electrothermal device models for MOS transistors, semiconductor resistors and diodes. In section 3, the analytical solution for the heat diffusion equation is derived. In section 4, an efficient recursive numerical convolution technique based on the RWE approximation is introduced. Experimental results are given in section 5. Finally, we summarize the paper.

II. Electrothermal Device Modeling

In this section, we briefly introduce the high current electrothermal device models. Please refer to [8] for detailed temperature-dependent device model equations.

A. MOS Transistor

Normally MOS transistors operate in the linear and saturation modes and are governed by standard MOS equations. But these standard equations are no longer applicable for transistors operating under high stress current in the ampere range. Under such stress, the avalanche breakdown and the turn-on of the parasitic lateral BJT transistors come into effect. Fig. 1 shows the equivalent circuit of the temperature-dependent MOS snapback model [10], including the effects of the parasitic BJT. Notice that additional current source I_{th}, modeling the thermal generation current, is added between the drain and internal base node for electrothermal simulation.

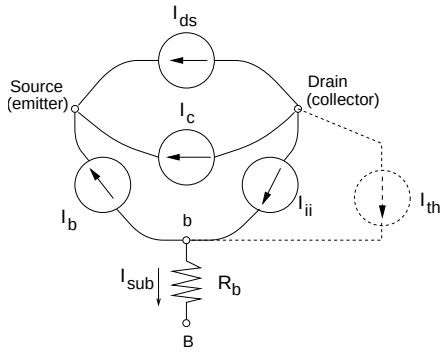


Figure 1: The MOS snapback model with the thermal generation current source I_{th}.

To calculate the transistor temperature, the transistor is modeled as heat sources at the junctions, as shown in Fig. 2. The heat source dimensions are approximated by $a = W$, $b = b_o \ln \frac{V_H \mu_{eff}}{b_o v_{sat}}$ and $c = x_j$, where $b_o = \sqrt{\frac{\epsilon_{si} t_{ox} x_j}{\epsilon_{ox}}}$ [11], v_{sat} is the electron saturation velocity, V_H is the holding voltage [10], μ_{eff} is the effective carrier mobility, x_j is the junction depth, t_{ox} is the gate oxide thickness and $\epsilon_{si}, \epsilon_{ox}$ are the dielectric constants of silicon and silicon dioxide.

Both junction temperatures T_{DJ} and T_{SJ} are needed to evaluate the parameters of the parasitic BJT at each step of the transient simulation. The drain junction temperature T_{DJ} controls the thermal generation current I_{th}, while the source junction temperature T_{SJ} affects the saturation currents I_{oe} and I_{oc}.

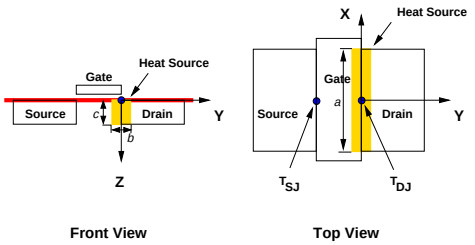


Figure 2: Sectional views of a MOSFET with modeled heat sources.

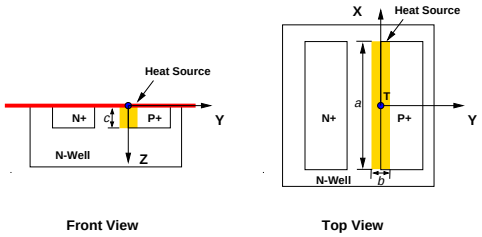


Figure 4: Sectional views of a lateral diode.

B. Semiconductor Resistor

A model for current flow in semiconductor resistors that accounts for both the low and high field effects is given in [13] as:

$$I_n = \frac{qnWX_jV}{L\left(\frac{1}{\mu_n} + \frac{V}{v_{sat}L}\right)} \quad . \tag{1}$$

where I_n is the electron current, V is the voltage across the resistor, L is the effective resistor length, X_j is the effective junction depth, v_{sat} is the saturation velocity of electrons in silicon, and μ_n is the electron mobility which depends on the doping level (N_d).

The temperature dependence in Eq. (1) arises from the mobility, carrier concentration and saturation velocity. To evaluate the semiconductor resistor temperature, the heat source is modeled as shown in Fig. 3. The heat source dimensions

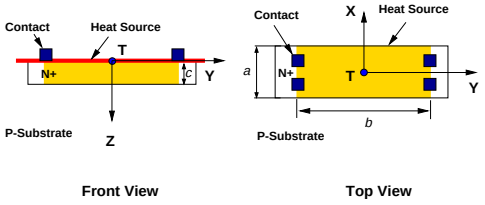


Figure 3: Sectional views of a diffusion resistor.

are approximated by $a = W$, $b = L$ and $c = X_j$, where L is the distance between the centers of the two end contacts. The temperature T is evaluated at the center of the heat source.

C. Diode

The avalanche breakdown characteristic of a reverse biased P-N junction diode is described by the following equation:

$$I_D = MI_s(\exp(V_D/V_T) - 1). \tag{2}$$

where M stands for the empirical *Miller* multiplication factor that models the breakdown phenomenon under fixed temperature and low-level injection.

The dominant temperature effect in Eq. (2) is introduced through the reverse saturation current I_s of the diode, which is a strong function of temperature. The sectional views of a lateral diode are shown in Fig. 4. The heat source dimensions are approximated by $a = W$, $b = x_d$ and $c = x_j$, where W is the width of the diode, x_d is the depletion width of the reverse biased junction and x_j is the effective junction depth. The temperature at the depletion edge is used to control the reverse saturation current.

III. Transient Thermal Modeling

A. Analytical Solution

The heat conduction equation for a homogeneous solid with uniform thermal conductivity is written as [14]:

$$\nabla^2 T(\mathbf{r},t) + \frac{g(\mathbf{r},t)}{k} = \frac{1}{D}\frac{\partial T(\mathbf{r},t)}{\partial t} \tag{3}$$

where $g(\mathbf{r},t)$ is the source power density, k is the bulk thermal conductivity, and D is the thermal diffusivity, defined as $\frac{k}{\rho C_p}$, with ρ being the density of material and C_p the specific heat.

In our analysis we consider a single rectangular heat source with dimensions (a, b, c) as shown in Fig. 5. We assume that the silicon dioxide layer above the bulk silicon is a heat insulator. Furthermore, the chip boundaries on all other sides, including the bottom, are treated as infinite boundaries whose temperatures remain the same as the ambient. The assumptions are valid in our case because the heat source dimensions are small compared with the chip size, and because the heat generated by the source cannot reach the chip boundary within the duration of the EOS simulation which is less than a few milliseconds.

The insulation assumption enables us to apply the method of images by adding an identical heat source symmetrically with respect to the insulation boundary, as illustrated in Fig 5. By doing so the problem is transformed into the equivalent problem of solving the temperature response for a heat resource in a medium extending infinitely in all directions.

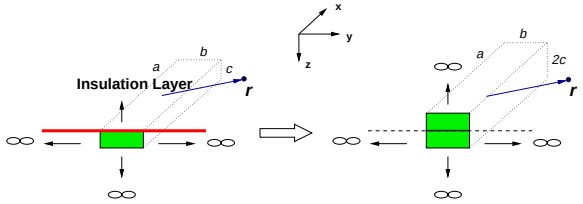


Figure 5: Problem transformation using the method of images. (a, b, c) is the heat source size, and $\mathbf{r}(x, y, z)$ is the temperature monitoring location.

The *Green*'s function $G(\mathbf{r}, t|\mathbf{r}', \tau)$ solution for a point source in an infinite medium for the following heat equation [15][16]

$$\nabla^2 G(\mathbf{r},t) + \delta(\mathbf{r}-r')\delta(t-\tau) = \frac{1}{D}\frac{\partial G(\mathbf{r},t)}{\partial t} \tag{4}$$

is the three dimensional *Gaussian* function

$$G(\mathbf{r},t|\mathbf{r}',\tau) = \frac{1}{[4\pi D(t-\tau)]^{3/2}} \exp\{\frac{-(\mathbf{r}-r')^2}{4D(t-\tau)}\} \tag{5}$$

Thus, the solution to the general time varying problem, Eq. (3), can be written as an integral over both time and the heat volume:

$$T(\mathbf{r},t) = T_a + \int_0^t \frac{P(\tau)d\tau}{\rho C_p \triangle} \int_\triangle \frac{dr'^3}{[4\pi D(t-\tau)]^{3/2}} \exp\{\frac{-(\mathbf{r}-r')^2}{4D(t-\tau)}\} \tag{6}$$

where T_a is the temperature of the surrounding ambient. As a result, the temperature response at an arbitrary point $\mathbf{r} = (x, y, z)$ from the pipelined rectangular source can be found by integrating r' over the ranges $-a/2 \le x' \le a/2$, $-b/2 \le y' \le b/2$, and $-c \le z' \le c$. We obtain

$$\begin{aligned} T(x,y,z,t) \quad = \quad & T_a + \frac{2}{\rho C_p \triangle}\int_0^t P(\tau)G(x,a,t-\tau) \\ & G(y,b,t-\tau)G(z,c,t-\tau)d\tau \end{aligned} \tag{7}$$

where

$$G(x,a,t) = \frac{1}{2}(erf(\frac{\frac{a}{2}+x}{2\sqrt{Dt}}) + erf(\frac{\frac{a}{2}-x}{2\sqrt{Dt}})) \quad (8)$$

$$G(y,b,t) = \frac{1}{2}(erf(\frac{\frac{b}{2}+y}{2\sqrt{Dt}}) + erf(\frac{\frac{b}{2}-y}{2\sqrt{Dt}})) \quad (9)$$

$$G(z,c,t) = \frac{1}{2}(erf(\frac{c+z}{2\sqrt{Dt}}) + erf(\frac{c-z}{2\sqrt{Dt}})) \quad (10)$$

$\triangle = 2abc$, and t is the duration of the transient simulation. Eq. (7) can be rewritten as

$$T(x,y,z,t) = T_a + \int_0^t P(\tau)h(t-\tau)d\tau \quad (11)$$

where P(t) is the power consumed by the device,

$$h(t) = \frac{2}{\zeta\rho C_p \triangle} G(x,a,t)G(y,b,t)G(z,c,t) \quad (12)$$

is the thermal impulse response, and $\zeta \in [1,2]$ is an empirical parameter to account for the non-perfect insulation of the silicon dioxide. $\zeta = 1$ corresponds to the case when the silicon dioxide is a perfect insulator. Eq. (11) is our main equation for calculating the transient temperature response of a device from its characteristics and power dissipation.

The thermal parameters of the substrate material in Eq. (3) are temperature dependent. For silicon, the lattice thermal conductivity and the ρC_p product are given in terms of the local lattice temperature T by [12]

$$k = 1.5486(\frac{T}{300})^{-\frac{4}{3}} \qquad [W/cm\ -^\circ K] \quad (13)$$

$$\rho C_p = 1.574(\frac{T}{300})^{0.1} \qquad [J/cm^2\ -^\circ K] \quad (14)$$

B. Power Monitor and Convolver

Given the thermal impulse response and power dissipation of a device, we calculate its temperature change by implementing Eq. (11) with two special devices, the power monitor and the convolver, as depicted in Fig. 6.

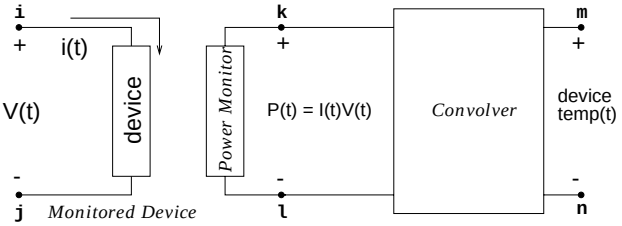


Figure 6: Typical setup of circuit for calculating device temperature from its power dissipation.

Referring to the node number assignments in Fig. 6, we know the power dissipated by a device is given by

$$P_{device}(t) \equiv v_{kl}(t) = v_{ij}(t)i_m(t) \quad (15)$$

which can be linearized for the time step t_n to give the characteristic equation of the power monitor as:

$$\mathbf{v}_{kl}(t_n) = v_{ij}(t_{n-1})\mathbf{i}_m(t_n) + i_m(t_{n-1})\mathbf{v}_{ij}(t_n) - v_{ij}(t_{n-1})i_m(t_{n-1})$$

The other special device convolver, which implements the convolution in Eq. (11), is described next.

IV. Fast Thermal Simulation

In this section, we first introduce a simple numerical method to perform the convolution in Eq. (11). Then we give a brief review of *Prony*'s method [17], which is used in our RWE technique to approximate the thermal impulse response $h(t)$ in Eq. (12) regionwisely with sums of exponentials. RWE approximation enables us to perform the convolution recursively, vastly improving the computation efficiency over the simple numerical scheme.

A. Numerical Convolution

Assuming that $P(t)$ is piecewise-linear, we can simply perform the convolution in Eq. (11) numerically using the trapezoid method:

$$\begin{aligned}\mathbf{T}(t_n) &\approx T_a + \mathbf{P}(t_n)\tfrac{F(t_n-t_{n-1})}{t_n-t_{n-1}} + \textstyle\sum_{i=1}^{n-1} P(t_i)\cdot \\ &\quad \left[\tfrac{F(t_n-t_{i-1})-F(t_n-t_i)}{t_i-t_{i-1}} - \tfrac{F(t_n-t_i)-F(t_n-t_{i+1})}{t_{i+1}-t_i}\right] \\ &= a(t_n)\cdot\mathbf{P}(t_n) + b(t_n) \end{aligned} \quad (16)$$

where $F(\cdot)$ is defined as

$$F(t) = \int_0^t \int_0^\tau h(\tau')d\tau'\, d\tau$$

Eq. (16) is implemented as a time-varying voltage-controlled voltage source whose coefficients $a(t_n)$ and $b(t_n)$ are updated at each time step. Clearly the evaluation of $a(t_n)$ and $b(t_n)$ becomes increasingly expensive as the simulation time increases.

B. Prony's Method

In linear system theory, the impulse response of a system can be described by its poles and the corresponding residues. The impulse response can be described by the summation of all the residues multiplied by exponentially damped sinusoids:

$$f(t) = \sum_{i=1}^{N} C_i e^{a_i t} \quad (17)$$

where a_i's are the poles, C_i's are their corresponding residues and N is the number of poles or residues.

Suppose that values of $f(t)$ are specified at a set of 2N *equally spaced* points, Eq. (17) can be rewritten as

$$\begin{aligned} f(t_n) &= \sum_{i=1}^{N} C_i e^{a_i n\Delta t}, n = 0,1,\cdots,2N-1 \\ &= \sum_{i=1}^{N} C_i \mu_i^n, n = 0,1,\cdots,2N-1 \end{aligned} \quad (18)$$

where Δt is the size of the time stepping interval, and $\mu_i = e^{a_i\Delta t}$. Eq. (18) becomes a set of $2N$ nonlinear equations in $2N$ unknowns.

The problem of interpolating a function using the sum of exponentials with unknown exponents was solved by *Prony* for the case of equally spaced data samples [17]. The method is based on the fact that the f_n in Eq. (18) must satisfy the following difference equation of order N:

$$\begin{aligned} f_N + f_{N-1}\alpha_1 + f_{N-2}\alpha_2 + \cdots + f_0\alpha_N &= 0 \\ f_{N+1} + f_N\alpha_1 + f_{N-1}\alpha_2 + \cdots + f_1\alpha_N &= 0 \\ &\cdots \\ f_{2N-1} + f_{2N-2}\alpha_1 + f_{2N-3}\alpha_2 + \cdots + f_{N-1}\alpha_N &= 0 \end{aligned} \quad (19)$$

where the roots of the algebraic equations

$$\mu^N + \alpha_1\mu^{N-1} + \alpha_2\mu^{N-2} + \cdots + \alpha_{N-1}\mu + \alpha_N = 0 \quad (20)$$

are $\mu_i (i = 1,\cdots,N)$.

Eq. (20) can be solved exactly for the α_i's $(i = 1,\cdots,N)$. Then, the roots of Eq. (20) can simply be found and the poles are obtained by

$$a_i = \frac{ln\mu_i}{\Delta t} \quad (21)$$

To obtain the residues C_i , the matrix contained in Eq. (18) is in the form of a transposed *Vandermonde* matrix whose inverse can be computed in a closed form. Thus, *Prony's* algorithm involves the solution of two matrix equations and a solution of the zeros of an Nth degree polynomial, N being the number of desired poles.

C. Regionwise Exponential (RWE) Approximation

Although *Prony*'s method provides an elegant solution to the problem of approximating a system using exponential terms, it is not directly applicable to approximating the thermal impulse response. Fig. 7 shows an impulse response example in log-log scale. Due to the requirement of equally-spaced sampling points, direct application of *Prony*'s method across the entire

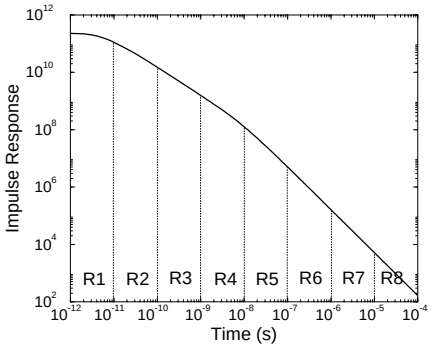
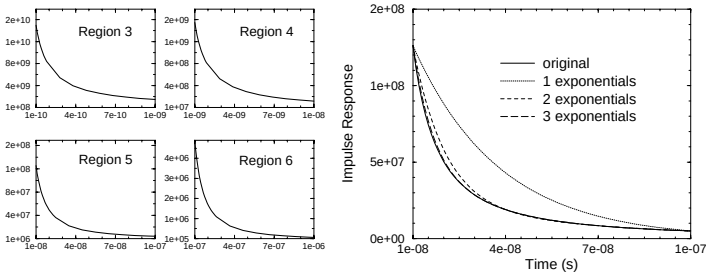


Figure 7: The thermal impulse response from Eq. (12). The heat source dimension (a, b, c) is $(20, 0.5, 0.5)$ and the temperature monitoring point (x, y, z) is $(0, 0, 0)$.

impulse response domain will result in under-sampling in the short time range and over-sampling in the long time range. Also, because the thermal response differs dramatically across the time domain, directly applying *Prony*'s method will result in non-negligible round-off errors. Furthermore, *Prony*'s method tends to generate positive poles when high orders of exponential terms are required.

The proposed regionwise exponential (RWE) approximation technique is as follows. First the entire thermal response is partitioned into a number of fixed regions, as shown in Fig. 7. Then *Prony*'s method is applied to approximate each region with a number of exponential terms. The more partitions used, the less the exponential terms required for each region. We find that partitioning into the following regions $[10^{-12}, 10^{-11}]$, $[10^{-11}, 10^{-10}]$, $\cdots$, $[10^{-6}, 10^{-5}]$ and $[10^{-5}, 10^{-4}]$ is both adequate and easy to implement. In Fig. 8(a), four regions of the impulse response are drawn in linear scale. Each region can

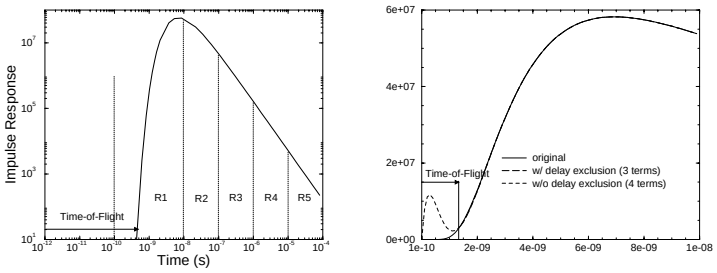


(a) (b)

Figure 8: (a) Several regions of the impulse response in Fig. 7 is shown in linear scale. (b) Several exponential approximations for region 5 using *Prony*'s method are shown. It is clear that fitting with three exponential terms provides the best approximation with the poles $(-0.21, -0.95, -2.95)$ and the corresponding residues $1e7 * (1.32, 4.45, 6.85)$.

be well approximated by three exponential terms, as shown for region 5 in Fig. 8(b). Note that the modeling duration for the impulse response must be long enough to cover the entire simulation time, which is normally no longer than several milliseconds.

Often we want to monitor the temperature at a location away from the heat source (see Fig. 5). For instance, we would be interested in the NMOS source temperature change due to the heat dissipated at the reverse-biased drain junction. In this case the temperature at the monitoring location does not respond instantaneously to the power dissipation. As a result, the thermal impulse response will exhibit a delay t_d, called time-of-flight, and our scheme of partitioning the impulse response will have to be adjusted accordingly to accommodate this delay effect. An example impulse response with a time-of-flight t_d is shown in Fig. 9(a), with the first partitioned region now beginning at t_d to avoid fitting the delay region.



(a) (b)

Figure 9: (a) The impulse response. The heat source dimension (a, b, c) in micron is $(20, 0.5, 0.5)$ and the temperature monitoring point (x, y, z) is $(0, 10, 0)$. (b) The exponential approximation for region 1. Better approximation can be achieved with the exclusion of the time-of-flight $(1.3ns)$.

D. Recursive Convolution

As mentioned earlier, the transient thermal simulation involves performing a convolution at each time step, which can become prohibitively expensive as the simulation duration increases. Here we employ the well-known recursive convolution method [18] to improve the computation efficiency significantly. Our approach differs from the previous work in that the we apply *Prony*'s method with our RWE technique to approximate the thermal impulse response with great accuracy, even with the presence of time-of-flight in the impulse response.

The recursive convolution formulation is described as follows. By using RWE approximation, the impulse response in Eq. (12) is expressed as the sum of N exponential terms:

$$h(t) \quad = \quad \sum_{j=1}^{N} h_j(t) = \sum_{j=1}^{N} C_j e^{a_j(t-t_j)} u(t-t_j) \qquad (22)$$

where t_j is the delay for the j-th exponential term. As a result, Eq. (11) can be rewritten as

$$T(t) = T_a + \sum_{j=1}^{N} T_j = T_a + \sum_{j=1}^{N} \int_0^t P(\tau) h_j(t-\tau) d\tau \qquad (23)$$

In other words, $T(t)$ can be obtained by summing up the contributions of individual exponential terms in the above convolution. Consider $T_j(t_n)$, the temperature change at time step t_n owing to the jth exponential term:

$$\begin{aligned} T_j(t_n) &= \int_0^{t_n} P(\tau) h_j(t_n - \tau) d\tau \\ &= \int_0^{t_n} P(\tau) C_j e^{a_j(t_n - t_j - \tau)} u(t_n - t_j - \tau) d\tau \\ &= \int_0^{t_n - t_j} P(\tau) C_j e^{a_j(t_n - t_j - \tau)} d\tau \\ &= e^{a_j(t_n - t_{n-1})} \int_0^{t_{n-1} - t_j} P(\tau) C_j e^{a_j(t_{n-1} - t_j - \tau)} d\tau \\ &+ \int_{t_{n-1} - t_j}^{t_n - t_j} P(\tau) C_j e^{a_j(t_n - t_j - \tau)} d\tau \\ &= e^{a_j(t_n - t_{n-1})} T_j(t_{n-1}) \\ &+ \int_{t_{n-1} - t_j}^{t_n - t_j} P(\tau) C_j e^{a_j(t_n - t_j - \tau)} d\tau \end{aligned} \qquad (24)$$

The first term of Eq. (24) can be calculated in constant time. If we approximate any unknown value of $P(t)$ in Eq. (24) with the linear interpolation of $P(t_{n-1})$ and $P(t_n)$, we can integrate the second term of Eq. (24) into a first-order function of $P(t_n)$. Then Eq. (24) can be rewritten as:

$$\mathbf{T}_j(t_n) = \left[e^{a_j(t_n - t_{n-1})} T_j(t_{n-1}) + A_j(t_n)\right] + B_j(t_n)\mathbf{P}(t_n) \quad (25)$$

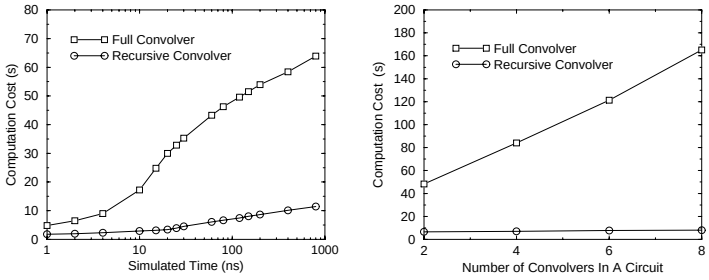
which is the characteristic equation of the recursive convolver. Similar to the case of numerical convolver, Eq. (25) is also implemented as a time-variant voltage-controlled voltage source with coefficients $A_j(t_n)$ and $B_j(t_n)$ updated at each time step. The difference is that the coefficient evaluation time is now reduced to only $O(1)$.

V. Experimental Results

In the simulator, two circuit components are implemented for thermal simulation, namely a full convolver which uses numerical convolution and a recursive convolver which uses the recursive convolution method. In this section, we first compare the performance of two thermal convolvers. Then, we present the electrothermal simulation examples of an individual device. Last, we present circuit models for simulating the thermal coupling effects among devices.

A. Convolution vs. Recursive Convolution

Since the RWE approximation accurately models the thermal impulse response, simulation using the recursive convolver produces almost identical results to those using the full convolver. However, recursive convolution has significantly reduced the computation time over full convolution as shown in Fig. 10. The computation cost is the run time in seconds on a SUN



(a) (b)

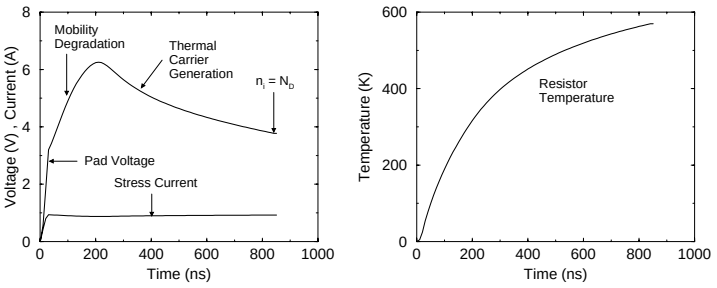
Figure 10: (a) Computation costs as a function of simulation time. (b) Computation costs as a function of the number of thermal convolvers.

SPARCstation 5. Fig. 10(a) compares the run time of the two methods for the electrothermal simulation of an NMOS transistor. The run time increases as the simulated time increases. The full convolver takes longer time because the cost for evaluating the full convolution increases significantly as the number of time steps increases. Fig. 10(b) shows the computation cost for 100ns simulation time. Each circuit consists of a certain number of NMOS transistors; each transistor needs two convolvers. In the case of the full convolver, the computation time increases drastically as the number of convolvers increases. However when the recursive convolver is used, the run time increases at a lower rate.

B. Device Electrothermal Simulation

N-Well Resistor

The results of the electrothermal simulation for an n-well resistor (W = 50 μm and L = 0.5 μm) are shown in Fig. 11 [8]. Electrothermal simulation has confirmed that initially the resistance increases as a result of mobility degradation due to increased lattice scattering. However, at higher temperatures, the



(a) (b)

Figure 11: (a) Transient simulations of an n-well resistor subjected to a 1 A EOS stress. The regions where mobility degradation and thermal carrier generation dominate are clearly indicated. The resistor becomes intrinsic at around 800 ns. (b) Temperature at the resistor center as a function of stress time.

increase in thermally generated carriers causes the resistance to decrease. Temperature-induced negative differential resistance can cause current localization leading to resistive thermal runaway and eventual failure under EOS conditions.

NMOS Transistor

With the thermal convolver, electrothermal simulation can be performed for devices under arbitrary stress waveforms such as the inputs generated by Human Body Model (HBM) events [2]. Fig. 12 shows the transient responses of an NMOS device to an HBM input. We observe that the drain junction temperature follows the drain stress current instantaneously. The source temperature follows the input current with a certain delay because the source junction is outside the heat source. As

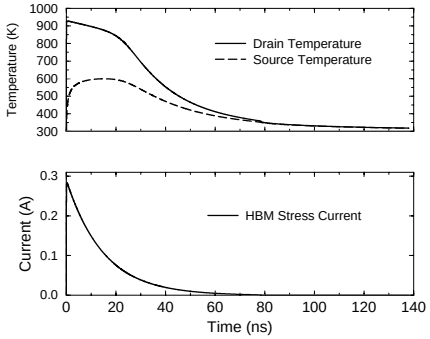


Figure 12: The transient responses (junction temperatures and drain current) of an NMOS device (W=20μm and L=0.6μm) under 400 V HBM-ESD input.

the device failure under HBM is induced by heating, we can determine the HBM-ESD level by performing the electrothermal simulation [8].

C. Thermally-Coupled Simulation

With the computational efficiency of the recursive convolver, it becomes feasible to simulate multiple devices, and their transient thermal coupling effects can be accurately modeled.

Here, the multi-finger NMOS device is chosen as a typical example to illustrate the thermally-coupled simulation method. In CMOS I/O circuits, the NMOS device is commonly used to withstand EOS/ESD stress. The channel width of the NMOS transistor in the output driver is hundreds of micron meters in

deep submicron technologies. The device is often laid out in a multifinger fashion. To accurately study the effect of layout on the NMOS device's EOS reliability, the cross coupling of the heat sources at each drain finger needs to be accurately modeled. As an example, a two-finger device and the temperature circuit are shown in Fig. 13. The power dissipated in each finger is monitored by the elements $Pm1$ and $Pm2$. To model the complete self-heating and coupling effects, it requires eight convolvers in total. Then the temperature summation property is used to obtain temperatures $T_{DJ1}, T_{DJ2}, T_{SJ1}$ and T_{SJ2}, which are fed back to the electrical models of individual fingers.

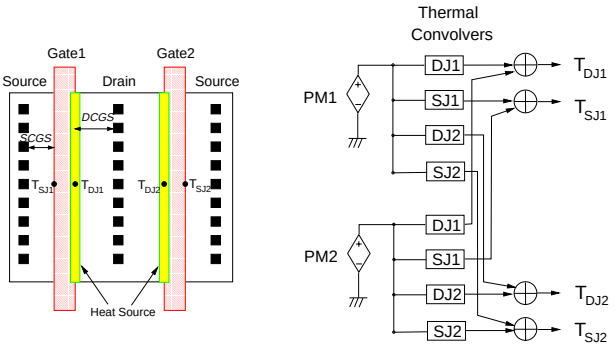


Figure 13: A two-finger NMOS device layout and the temperature monitoring circuit for the fully coupled electrothermal simulation. DCGS stands for drain contact-to-gate spacing and SCGS stands for source contact-to-gate spacing.

Simulation results for a ten finger device under a 0.5 A stress current are plotted in Fig. 14. Initially all the fingers conduct the stress current equally. However, as the temperature rises,

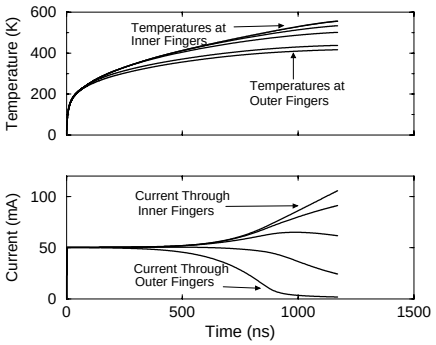


Figure 14: Transient temperature and current responses under a 0.5 A EOS current input.

the inner fingers become hotter than the outer ones. The non-uniform temperature distribution across the fingers causes a redistribution of the current. The thermal generation current in the innermost fingers increases due to the higher temperature; as a result, they conduct more of the stress current. This causes a further increase in the temperature of the innermost fingers and eventually leads to thermal failure of them at around 1200 ns. As seen from the simulation results, the effective total device width will be reduced due to the non-uniform current distribution. With accurate electrothermal simulation, the NMOS device layout in terms of finger width, DCGS and SCGS can be optimally designed to meet the EOS performance requirement.

VI. Summary

In this paper, we have proposed a new thermal simulation algorithm for efficient electrothermal simulation. The algorithm is based on a regionwise exponential (RWE) approximation technique and a recursive convolution scheme. With efficient closely-coupled electrothermal simulation, transient thermal effects on the CMOS circuits, especially the circuit failure due to electrical overstress (EOS) can be studied. With the high simulation accuracy and short turn-around time, the simulator can be effectively used for circuit and layout optimization for real circuits before the design is committed to silicon.

References

[1] P. Yang and J. Chern, " Design for Reliability : the Major Challenge for VLSI," *Proc. of the IEEE,* vol.81, no.5, pp. 730–743, May 1993.

[2] C. Duvvury and A. Amerasekera, "ESD: A Pervasive Reliability Concern for IC Technologies," *Proc. of the IEEE,* vol. 81, no. 5, pp. 690–702, May 1993.

[3] Y. K. Cheng, C. C. Teng, A. Dharchoudhury, E. Rosenbaum and S. M. Kang, "iCET : A Complete Chip-Level Thermal Reliability Diagnosis Tool for CMOS VLSI Chips," *Proc. ACM/IEEE Design Automation Conf.*, pp. 548–551, 1996.

[4] C. Diaz, S. M. Kang and C. Duvvury, "Circuit-level Electrothermal Simulation of Electrical Overstress Failures in Advanced MOS I/O Protection Devices," *IEEE Trans. on CAD*, vol. 13, no. 4, pp. 482–493, 1994.

[5] J. Bielefeld, G. Pelz, H. B. Abel and G. Zimmer, "Dynamic SPICE-Simulation of the Electrothermal Behavior of SOI MOSFET's," *IEEE Trans. on Electron Devices,* vol. 42, no. 11, pp. 1968–1974, November, 1995.

[6] C. Diaz, C. Duvvury and S. M. Kang, "Studies of EOS Susceptibility in 0.6 μm nMOS ESD I/O Protection Structures," *EOS/ESD Symp.*, pp. 205-211, 1993.

[7] Technology Modeling Associates, Inc., Palt Alto, California, *MEDICI, Two Dimensional Device Simulation Program*, 1992.

[8] S. Ramaswamy, "*Modeling, Simulation and Design Guidelines For EOS/ESD Protection Circuits in CMOS Technologies*," Ph.D. Thesis, University of Illinois at Urbana-Champaign, 1996.

[9] A. Amerasekera, M. Chang, J. A. Seitchik, A. Chatterjee, K. Mayaram and J. Chern, " Self-Heating Effects in Basic Semiconductor Structure," *IEEE Trans. on Electron Devices*, Vol. 40, pp. 1836–1844, Oct., 1993.

[10] A. Amerasekera, S. Ramaswamy, M. Chang and C. Duvvury, "Modeling MOS Snapback and Parasitic Bipolar Action for Circuit-Level ESD and High Current Simulations," *International Reliability Physics Symposium*, pp. 318–326, 1996.

[11] P.-K. Ko, *Advanced MOS Device Physics*, ch. 1. Approaches to Scaling. San Francisco: Academic Press, 1989.

[12] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Velag, New York, 1984.

[13] G. Krieger and P. Niles, "Diffused Resistors Characteristics at High Current Density Levels – Analysis and Applications," *IEEE Trans. on Electron Devices*, Vol. 36, pp. 416–423, Feb., 1989.

[14] M. N. Ozisik, *Boundary Value Problems of Heat Conduction*, Dover Publications, New York, 1968.

[15] G. F. Roach, *Green's Function*, Cambridge, 1982.

[16] V. Dwyer, A. Franklin and D. Campbell, "Thermal Failure in Semiconductor Devices," *Solid State Electronics*, vol. 33, no. 5, pp. 553–560, 1990.

[17] F. B. Hilderbrand, *Introduction to Numerical Analysis*, McGraw-Hill, 1973.

[18] S. Lin and E. S. Kuh, "*Transietnt Simulation of Lossy Interconnect*," *Proc. ACM/IEEE Design Automation Conf.*, pp. 81–86, 1992.