iCET : A Complete Chip-Level Thermal Reliability Diagnosis Tool for CMOS VLSI Chips *

Yi-Kan Cheng, Chin-Chi Teng, Abhijit Dharchoudhury†, Elyse Rosenbaum, and Sung-Mo Kang

Coordinated Science Laboratory
1308 West Main Street
Univ. of Illinois at Urbana-Champaign
Urbana, IL 61801

†High Performance Design Technologies
Motorola Inc.,
6501 William Cannon Drive West
Austin, TX 78735-8598

Abstract

In this paper, we present the first chip-level electrothermal simulator, iCET. For a given chip layout, packaging material, user-specified input signal patterns, and thermal boundary conditions, it automatically finds the CMOS on-chip steady-state temperature profile and the resulting circuit performance. iCET has been tested on several circuits and it can efficiently analyze layouts containing tens of thousands of transistors on a desktop workstation.

1. Introduction

Due to the increasing component density in IC devices with higher operation speed and larger scale of integration, the power density and the on-chip temperature increase dramatically. Since the failure rate of microelectronic devices depends heavily on the localized operating temperature, hot spots due to high local power dissipation have become a long-term IC reliability concern. Due to the complexity of active devices in a VLSI chip, verification of electrical functions at various operating temperatures relies on computer simulations. A new fast and accurate chip-level thermal reliability diagnosis tool is therefore required. Once the accurate temperature profile is determined, several important issues can be addressed as shown in Fig. 1.

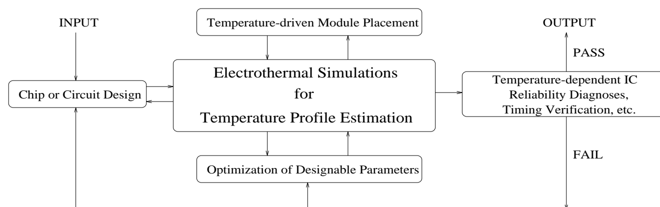

Figure 1: Applications of electrothermal CAD tools.

Although many research efforts have been undertaken to deal with the solution of electrothermal problems in electronic circuits [1][2][3], there have been no attempts at providing electrothermal simulation capability at the chip level. The chip-level tool must avoid the following problems found in existing tools:

- simulation inefficiency resulting from commonly used *coupled* electrical and thermal simulations;
- slow execution speed of SPICE-like simulators for device power calculations;

*This research was supported in part by Intel Corporation, Rome Laboratory (F30602-95-1-0006), JSEP (N00014-96-J-0129), and Semiconductor Research Corp. (SRC95-DP-109).

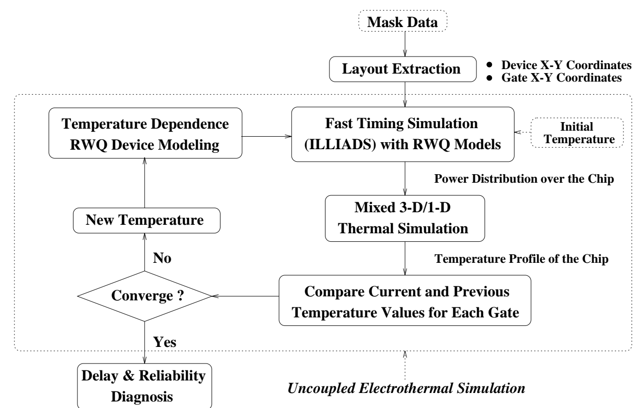

Figure 2: Flowchart of iCET electrothermal simulation.

- improper use of thermal boundary conditions (BCs);
- no automation for top-down procedures.

The work presented in this paper is intended to remove the above bottlenecks by introducing a new methodology for temperature profile estimation and hot spot identification in CMOS VLSI chips.

2. Simulation Method in iCET

The flowchart of iCET electrothermal simulation procedures is shown in Fig. 2. The main features of iCET are listed as follows: First, to achieve the computational time efficiency required by large circuits, iCET uses a fast timing simulator, ILLIADS, to calculate the power dissipated by each *logic gate*. The gates are then viewed as heat sources of the chip. ILLIADS provides the following advantages: (i) The speedup of ILLIADS over SPICE-like programs increases linearly with the circuit size as measured in terms of the transistor count; (ii) The speedup is further enhanced by introducing the *incremental* simulation technique; (iii) An accurate *temperature-dependent* modeling method for MOS device is developed by using the regionwise quadratic (RWQ) modeling technique [4]. The accuracy of power dissipation estimated by ILLIADS is comparable to SPICE for a wide range of temperatures.

Secondly, existing *coupled* electrothermal simulators are inefficient in nature [1][2]. The total simulation period is first divided into many small time intervals, and the transient power and temperature values are then updated and coupled for *every* time interval. Since iCET is designed to find the chip-level steady-state temperature distribution, a much more efficient simulation methodology is used: ILLIADS first finds the *average* power for each gate. Next, power values are fed to our thermal simulator to estimate the *pseudo*-steady-state temperature distribution. The distribution is then used to update the device model parameters for the second-round power calculation. This process continues until convergence is obtained and the *real* steady-state temperature is found. Note that our approach decouples power and temperature calculations; this is valid because the thermal time constant (in the range of

33rd Design Automation Conference ®
Permission to make digital/hard copy of all or part of this work for personal or class-room use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copying is by permission of ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permssion and/or a fee.
DAC 96 - 06/96 Las Vegas, NV, USA ©1996 ACM, Inc. 0-89791-833-9/96/0006..$3.50

milliseconds) is much larger than the typical signal switching period.

Thirdly, a novel thermal simulator, iTEMP, is developed to find the on-chip temperature distribution. It solves *3-D* heat equations for the chip substrate but treats the packages and heat sinks as *1-D* thermal resistances. A hierarchical approach is also developed to efficiently identify the hot spots. iTEMP accepts a variety of thermal BCs at any side of the chip.

Fourthly, by using the RWQ modeling technique, power estimation can be done by iCET even when only measured data is available and the MOS models have not been fully developed or characterized. This makes iCET adaptive to advanced device technologies.

Referring back to Fig. 2, the primary input to iCET is the layout description file of the target VLSI chip. A layout extractor has been developed to obtain the electrical circuit which the layout represents, as well as to identify the location of each device. Next, iCET calculates the geometrical bounds and the average power dissipation (at room temperature) for each logic gate. iTEMP then takes as input the power values and coordinates of the heat sources to calculate the on-chip temperature distribution, and the average temperature of each gate is found. Now, each gate has updated local temperature so that ILLIADS must be re-executed to find the new average power values under the new temperature distribution. From our experiments, the number of iterations required for convergence is usually less than 4, since the temperature rise only affects the short-circuit power which is only about 10% of the total power consumption.

3. RWQ Modeling Procedure

The RWQ modeling procedure takes a set of data points (V_{ds}, V_{gse}, I_{ds}) as input, where $V_{gse} = V_{gs} - V_{t0}$ and V_{t0} is zero-bias threshold voltage of a MOS device. Next the (V_{ds}, V_{gse}) plane is optimally partitioned into a number of regions and a quadratic model of I_{ds} in terms of V_{ds} and V_{gse} is numerically fitted in each region. For a given regionwise partition, the following quadratic model of I_{ds} is fitted to the data in the kth region where $k = 1, \ldots, n_r$

$$\frac{I_{ds}}{\beta} = (\alpha_0^{(k)} + \alpha_1^{(k)} V_{gse} + \alpha_2^{(k)} V_{ds} + \alpha_3^{(k)} V_{gse}^2 + \alpha_4^{(k)} V_{gse} V_{ds} + \alpha_5^{(k)} V_{ds}^2) \quad (1)$$

where $\beta = \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}$, n_r is the number of regions chosen for best fitting, and α's are fitting parameters in the kth region. In (1), μ_0 and V_{t0} are two strongly temperature-dependent physical parameters and $\alpha_0 \sim \alpha_5$ are weakly temperature-dependent. Here we employ the same temperature dependence equations of μ_0 and V_{t0} as those used in SPICE, and denote μ_0 and V_{t0} calculated at temperature T as $\mu_0(T)$ and $V_{t0}(T)$ respectively. Provided that we RWQ-fit a set of experimental data at T and obtain the corresponding $\alpha_0 \sim \alpha_5$, we denote this set of α's as RWQ(T). To accurately model I_{ds} at T, the following approach is adopted: Users first need to fit the experimental I_d-V_{ds} data at three (or any user-specified number) different temperatures (T_1, T_2, T_3). Thus RWQ(T_1), RWQ(T_2) and RWQ(T_3) are obtained. If the local temperature T of a device satisfies $T_1 < T < T_2$ during the electrothermal simulation, $\mu_0(T)$ and $V_{t0}(T)$ as well as RWQ(T_1) are used in (1). Similarly, $\mu_0(T)$, $V_{t0}(T)$ and RWQ(T_2) are used if $T_2 < T < T_3$; $\mu_0(T)$, $V_{t0}(T)$ and RWQ(T_3) are used if $T_3 < T$.

To demonstrate the accuracy of the above approach, we RWQ-fit the I_d-V_{ds} data generated by PSPICE at $T_1 = 27\ °C$, $T_2 = 77\ °C$, and $T_3 = 127\ °C$. RWQ(27), RWQ(77) and RWQ(127) are thus obtained. The resulting I_{ds}-V_{ds} characteristics at $T = 100\ °C$ is shown in Fig 3, which shows good agreement with the SPICE data. Note that in our approach, we implicitly take into account V_{t0}'s degree of freedom which is originally suppressed in the RWQ fitting procedure. If instead, we use only *one* set of RWQ parameters fitted at room temperature (*i.e.,* RWQ(27)) for all higher temperatures, it will give less accurate results as shown in Fig 4. When simulating VLSI chips, we measure and fit the device I-V curves under a few operating temperatures and generate the corresponding

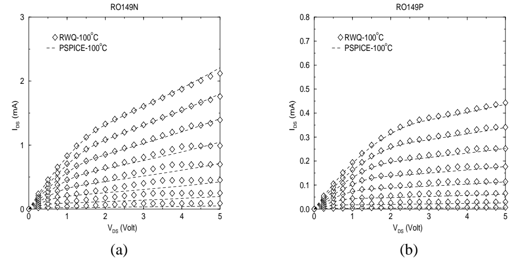

Figure 3: I-V characteristics at 100 $°C$ using RWQ(77).

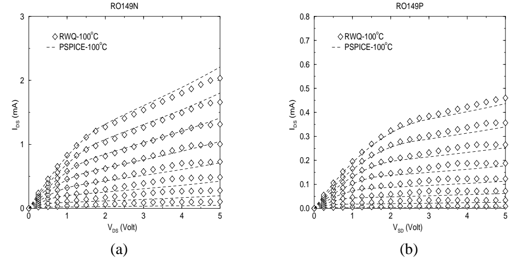

Figure 4: I-V characteristics at 100 $°C$ using RWQ(27).

RWQ(T)'s. Appropriate RWQ(T) will be called during the iCET simulation and highly accurate temperature-dependent power values can be estimated.

4. Numerical/Analytical Thermal Simulator - iTEMP

4.1 Modeling

The heat diffusion equation is the governing equation for heat conduction and temperature calculation. The general equation is written as [5]

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} + \frac{g}{k} = \frac{1}{\alpha}\frac{\partial T}{\partial t} \quad (2)$$

subject to the general thermal BC:

$$k\frac{\partial T}{\partial n_i} + hT = f_i \quad (3)$$

where T is temperature, g is the power density of the heat source(s), k is thermal conductivity, α is thermal diffusivity, h is the heat transfer coefficient, f_i is an arbitrary function of position, and n_i is the outward direction normal to the surface i.

iTEMP simulates a chip by the following mixed 3-D/1-D strategy: (i) 3-D simulation is performed for the chip substrate; (ii) To enhance the simulation efficiency, packages and heat sinks are treated as *1-D thermal resistances.* We call strategy (ii) the 1-D *effective* heat transfer method. The idea is to serially combine the thermal resistance of the *bulk of packages or heat sinks* (R_k) with the one *from packages to ambience* (R_h), and to find the effective heat transfer coefficient h^e as given by

$$h^e = \frac{1}{A_c}\frac{1}{(R_h + R_k)} \quad (4)$$

where $R_k = \frac{L}{k_p A_c}$, $R_h = \frac{1}{h_p A_c}$, L is the thickness and k_p is the thermal conductivity of packages or heat sinks, h_p is the heat transfer coefficient from packages or heat sinks to ambience, and A_c is the chip area normal to the direction of heat flow. In other words, we merge the package and heat sink effects into the h term in (3) and form an effective h^e.

4.2 Formulation

Equation (2) and (3) can be solved by numerical or analytical methods. By combining these two methods, iTEMP can not only estimate the on-chip temperature profile, but can also hierarchically identify the hot spots and the corresponding

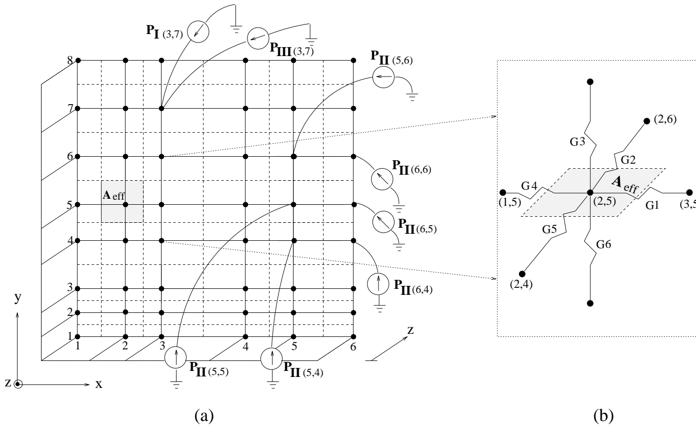

Figure 5: Analogous thermal circuit.

temperatures efficiently. We first present the numerical finite-difference (FD) method used in iTEMP. Since heat conduction in a thermal circuit is analogous to current conduction in an electrical circuit, we can transform a FD heat conduction problem into an electrical RC network problem. Because the gate count in a VLSI chip is large, it is impractical to allocate one or more grids to each gate or device [1][2] in the FD method. Instead, the grid number and spacings in iTEMP are determined by taking into account the chip size and the gate density, as well as the temperature field density. We employ an *adaptive* meshing technique to determine the grid spacing in iTEMP. First, iTEMP coarsely deploys the on-chip grids according the gate density and user-specified initial grid number. After obtaining the initial estimation of the temperature distribution, iTEMP further refines or redistributes the grids by sensing the temperature gradient and adds extra grids for the places with larger gradient, based on the following *weight function* and *equidistribution* [6]:

$$w(r) = \sqrt{1 + \alpha^2 (\frac{\partial T}{\partial r})^2} \quad (5)$$

$$\int_{r_i}^{r_{i+1}} w(r) dr = Constant \quad (6)$$

where α is a user-specified parameter and r denotes x or y. The stopping criterion of grid refinement is by comparing the temperature solutions between current and previous grid systems. If the solution difference is less than the prescribed threshold, then the error of the current grid system is within a certain bound and the refining process is terminated. Note that in order to avoid the speed penalty of full-substrate simulation, only a few grids are introduced in the z-direction (thickness) with most of the grids near the heat sources. The above heuristic makes use of the concept of *heat diffusion length*, beyond which the temperature gradient becomes small. Adding only a few grids near the heat source in the z-direction makes iTEMP even more efficient.

A schematic representation of the grid points into which the heat flows and the analogous thermal circuit are shown in Fig. 5, where P_i denotes the heat flow coming from source i. The solid lines in Fig. 5(a) represent the chosen grid lines and dashed ones are in the middle of two adjacent grid lines. A_{eff} is the effective area of a grid point. Every heat source that overlaps the effective area of some grid point serves as a power source feeding into that grid, and the corresponding power value is calculated based on the ratio of the source area within A_{eff} to the total area of the source. $G_1, \ldots, G_6$ in Fig. 5(b) are the thermal conductances. The resulting thermal circuit is solved by the sparse matrix routine and the on-chip temperatures are obtained. We next compute the average temperature of each gate by averaging the temperature values of grids that a gate covers, and then use the updated values as the input to the fast timing simulator for the next simulation run. If a gate does not cover any grid points, a *2-D polynomial interpolation* scheme based on the temperatures at nearby grid points is adopted to find the average temperature.

Equation (2) and (3) can also be solved by the application of analytical Fourier-integral and Fourier-inversion formulae in the finite range of $0 \leq x \leq a$, $0 \leq y \leq b$ and $0 \leq z \leq c$, where a, b, c are the chip dimensions. Due to the expression length, we only present the solution of the case where all four sides and top surface of the chip are insulated, while the bottom surface is convective. In this case, the corresponding *eigenfunctions* $(K(\beta_m, x),\ K(\nu_n, y),\ K(\eta_p, z))$ and *eigenvalues* $(\beta_m,\ \nu_n,\ \eta_p)$ can be found. By taking the Fourier integral of (2) we have

$$\overline{T}(\beta_m, \nu_n, \eta_p) = \frac{A(\beta_m, \nu_n, \eta_p)}{(\beta_m^2 + \nu_n^2 + \eta_p^2)} \quad (7)$$

where

$$A(\beta_m, \nu_n, \eta_p) = \frac{1}{k}[\overline{g}(\beta_m, \nu_n, \eta_p) + K(\eta_p, z)|_{z=0} \cdot \int_{x'=0}^{a} \int_{y'=0}^{b} (h_z^e \cdot T_a) K(\beta_m, x') K(\nu_n, y') dx' dy'] \quad (8)$$

and $\overline{g}$ is the integral transform of the heat sources, which can be expressed as

$$\overline{g}(\beta_m, \nu_n, \eta_p) = \frac{16 Q_p}{\sqrt{ab}} \frac{1}{\beta_m \nu_n \eta_p} \sum_{i=1}^{n_r} g_i \cos(\beta_m x_{ic}) \sin(\frac{\beta_m}{2} x_{id}) \cdot \cos(\nu_n y_{ic}) \sin(\frac{\nu_n}{2} y_{id}) \cos(\eta_p (c - z_{ic})) \sin(\frac{\eta_p}{2} z_{id}) \quad (9)$$

where $(x_{ic},\ y_{ic},\ z_{ic})$, $(x_{id},\ y_{id},\ z_{id})$ and g_i are the center coordinates, dimensions, and power density of heat source i, respectively. n_r is the number of heat sources. $Q_p = \sqrt{2}[\frac{\eta_p^2 + H^2}{c(\eta_p^2 + H^2) + H}]^{1/2}$ and $H = h_z^e / k$ where h_z^e is the bottom effective heat transfer coefficient. Once $\overline{T}$ has been determined, on-chip temperature at any position can be found subsequently by applying the Fourier-inversion formula.

4.3 Hierarchical hot-spot identification

Although the analytical approach for solving the heat equation provides an analytic expression for every point (x,y,z), it is computationally more efficient than the numerical approach only when the number of points whose temperatures need to be calculated is small. Our experiments show that the numerical approach is generally more efficient than the analytical one in iCET simulations. By using the analytical method, however, we can hierarchically find the hot spots and the corresponding temperatures in an extremely efficient way.

We first use the numerical approach to obtain the on-chip temperature profile and to further identify the areas containing hot spots. Because no rigorous temperature calculations are necessary at this stage, a looser error bound for adaptive grid generation is specified and the wider grid spacings are generated. After locating these areas, we next use the analytical approach to *pinpoint* the hot spots and to determine their temperatures. The advantages of the hierarchical approach for hot spot identification are: (i) We use numerical approach to scale down the problem size for analytical thermal simulation. (ii) Since numerical simulation is used only to screen out the areas not containing the hot spots, the grid number needed will be small. By this hierarchical approach, hot-spot temperatures are accurately and efficiently determined by taking advantage of both numerical and analytical methods.

5. Incremental Electrothermal Simulations

As we mentioned previously, the uncoupled electrothermal simulation method usually takes less than 4 iterations to find the steady-state temperature; the iteration process stops when the current temperature converges to the previous value. This is illustrated in Fig. 6, where a 9-stage ring oscillator is simulated by iCET. The power and temperature values were recorded during the iteration process. In Fig. 6, the temperature difference between two successive simulation runs becomes smaller as the number of iterations increases. This information indicates the feasibility of *incremental simulation*, in which the circuit parameters vary by small amounts about their previous nominal values.

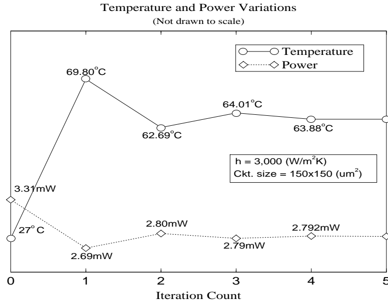

Figure 6: Convergence plot for power and temperature.

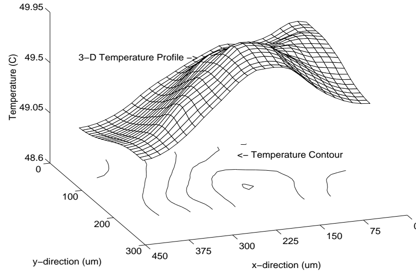

Figure 7: Temperature profile of the 10-bit adder.

Incremental Simulation Procedure

Consider a circuit containing blocks which are ordered for simulation based on the fanin-fanout relationship. In iCET, if the temperature difference in a block between current (perturbed) and previous (nominal) runs exceeds a prescribed threshold (T_THRLD), then we say that the block has local temperature variation and it will be marked with T_VAR. For a block marked with T_VAR, it is not considered to be *latent* and it needs to be re-simulated. The re-simulated waveforms then serve as the nominal waveforms for the next simulation run. However, if a block is not marked with T_VAR, then the nominal and perturbed waveforms for all of the inputs to the block are compared. If the difference between them is less than a user-specified threshold, the block is marked as being incrementally latent and is not simulated (*i.e.*, its perturbed solution is the same as its nominal solution). On the other hand, if the difference in any of the inputs is larger than the threshold, the block is not considered to be latent and is simulated. For increasing number of iterations in the iCET simulation, the advantage of the incremental approach will be even more since it is expected that a large number of latent cases will be detected. Moreover, for larger circuits, we expect that the computational savings of latency will be more pronounced because a larger number of blocks are going to be latent. Note that iCET identifies the blocks having local temperature variations *dynamically* for each new simulation run. In other words, a block can be marked with T_VAR in one run but is not marked in another. Moreover, the T_THRLD value which is initially specified by users for determining if a block has the local temperature variation, is also dynamically updated.

5. Simulation Results

In this section, we demonstrate iCET simulation results for a number of circuits. Here we assume that the top and four sides of the circuits are insulated, while the h value of the bottom surface is assumed to be 3,000 for all test circuits. We first consider a 10-bit negative adder. Simulation results are listed in the first row of Table 1 and Table 2. The resulting 3-D and contour temperature profile is shown in Fig. 7. In Table 1, n_{tran} and n_{hsrc} are the numbers of transistors and heat sources (gates) in circuits; P_{avg} is the total average power dissipated in circuits. In Table 2, T_{avg} is the average temperature of circuits; n_{run} is the number of repeated simulation runs; n_{tot_lat} is the total number of blocks being latent during the *whole* electrothermal simulation; S_{incr} is the speedup factor of the timing simulation, which is computed as the ratio between the total transient analysis time with and without incremental simulation. Output waveforms of bit 0 and bit 10 of the adder are shown in Fig. 8, where more output delay due to thermal effect is in bit 10 than in bit 0 after iCET electrothermal simulation. Table 1 and Table 2 also present simulation results for several other circuits: HIGHWAY [7], ALU and control, and a 16-bit multiplier.

Circuit	n_{tran}	n_{hsrc}	Freq.[MHz]	$P_{avg}[mW]^{\dagger}$
10-bit Neg. Adder	868	216	100	12.07
HIGHWAY	248	17	25	1.48
ALU and Control	5842	1656	200	67.67
16-bit Multiplier	11016	3001	100	289.32

†Under steady-state temperature distribution.

Table 1: iCET simulation results.

Circuit	$T_{avg}[\ ^\circ C]$	n_{run} [1]	n_{tot_lat}	S_{incr}	CPU[sec] [2]
10-bit Neg. Adder	49.40	3	136	1.35	26.22
HIGHWAY	41.76	3	13	1.34	12.82
ALU and Control	35.03	2	591	1.31	248.54
16-bit Multiplier	45.75	3	2093	1.37	914.08

[1] Convergence criterion: $(\Delta T/T) < 1\%$. [2] On SUN SPARC 10.

Table 2: iCET simulation results.

6. Conclusions and Future Work

In this paper, a new chip-level thermal reliability diagnosis tool is presented. The objective of iCET is to find the on-chip temperature profile, hot spots, as well as the resulting circuit performance for VLSI chips. In the future, we will extend iCET as a general purpose circuit reliability diagnosis program to provide more accurate information for many temperature-dependent reliability problems such as electromigration (EM).

REFERENCES

[1] S. S. Lee and D. J. Allstot, “Electrothermal Simulation of Integrated Circuits,” *IEEE J. Solid-State Circuits*, pp. 1283-1293, Dec. 1993.

[2] C. H. Diaz, S. M. Kang, and C. Duvvury, “Circuit-Level Electrothermal Simulation of Electrical Overstress Failures in Advanced MOS I/O Protection Devices,” *IEEE Trans. Computer-Aided Design*, pp. 482-493, Apr. 1994.

[3] V. Koval, I. W. Farmaga, A. J. Strojwas, and S. W. Director, “MONSTR: A Complete Thermal Simulator of Electronic Systems,” *Proc. ACM/IEEE Design Automation Conf.*, pp. 570-575, 1994.

[4] A. Dharchoudhury, S. M. Kang, K. H. Kim, and S. H. Lee, “Fast and Accurate Timing Simulation with Regionwise Quadratic Models of MOS I-V Characteristics,” *Proc. ACM/IEEE Int. Conf. Computer-Aided Design*, pp. 208-211, Nov. 1994.

[5] M.N. Ozisik, *Boundary Value Problems of Heat Conduction,* London, U.K. Oxford, 1968.

[6] J. F. Thompson, Z. Warsi, and C. W. Mastin, *Numerical Grid Generation,* North-Holland, 1985.

[7] Microelectronics Center for North Carolina, *VPNR Users Guide.*

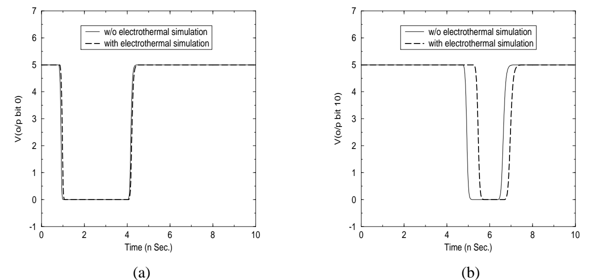

Figure 8: Output waveforms of: (a) bit 0 (b) bit 10