# Coupled Modeling of Time-Dependent Full-Chip Heating and Quantum Non-Isothermal Device Operation

Akin Akturk, Neil Goldsman and George Metze†
Department of Electrical and Computer Engineering University of Maryland, College Park, MD 20742, USA
†Laboratory of Physical Sciences, College Park, MD 20742, USA,
akturka@glue.umd.edu, neil@eng.umd.edu

*Abstract*- **A method for predicting full chip temperature heating resulting from device operation is presented. The method couples distributed device simulation with lumped thermal analysis. Predictions show sixty degree Kelvin temperature increases for 0.5cm IC's. A method for reducing chip temperature is also presented.**

## I. Introduction

As devices get smaller on-chip thermal effects become increasingly important. Predictions indicate that chip temperatures will increase exponentially beyond acceptable values[1], prompting researchers to investigate thermal effects[2]. In this work, we developed a mixed-mode simulation technique which predicts temperatures over the entire chip using non-isothermal device modeling. On the device level, we solve the Schrodinger and semiconductor equations coupled with the heat flow equation. From the distributed heat flow equation, a lumped model for heat flow and temperature distribution for the entire chip is derived. A Monte Carlo-type methodology, which connects the device modeling and on-chip temperature calculations, is employed. The model self-consistently calculates the device characteristics along with heating as a function of position on the chip and operation time. Calculations also show how device performance will be affected as the power density increases. We also offer solutions to the over-heating problem by showing the effect of placing thermal contacts on the chip.

## II. Mixed Mode Device Performance and Chip Temperature Model

We combine our quantum device transport model and the lattice heating equation[3], with a complete chip heat transport model. The complete set of equations is shown below.

Our device model includes Schrodinger, Poisson, electron current-continuity, hole current-continuity and lattice heat flow equations shown below, respectively.

$$E\psi(y) = -\frac{\hbar^2}{2m^*}\frac{d^2}{dy^2}\psi(y) - q\phi(x,y)\psi(y) \tag{1}$$

$$\nabla^2\phi = -\frac{q}{\epsilon}(p - n + D) \tag{2}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla.\overrightarrow{J_n} - R_n + G_n \tag{3}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla.\overrightarrow{J_p} - R_p + G_p \tag{4}$$

$$C\frac{\partial T}{\partial t} = \nabla.(\kappa\nabla T) + H \tag{5}$$

Here $\phi$, $\psi$, $n$, $p$, $J_n$, $J_p$, $T$, $H$, $D$, $R$, and $G$ are the electrostatic potential, wave function, mobile electron concentration, mobile hole concentration, electron current density, hole current density, lattice temperature, lattice heating, net dopant concentration, net recombination rate and net generation rate, respectively. In the heat flow equation, $C$ is the heat capacity and $\kappa$ is the thermal conductivity.

To obtain temperature over the entire chip, we transform the differential heat flow equation (5) to a lumped equation. This leads to the following $R^{th}C^{th}$ thermal network of lumped KCL-type equations for heat flow throughout the chip.

$$C_i^{th}\frac{dT_i}{dt} + \frac{T_i - T_j}{R_{i,j}^{th}} = I_i^{th}(T_i) \tag{6}$$

The subscript $i$ represents a specific device and a thermal node. The maximum value of $i$ equals the number of devices on the chip, and a value for the chip temperature $T_i$ is calculated at each node. $C_i^{th}$ is the thermal capacitance which is proportional to the local specific heat, $R_{i,j}^{th}$ is the thermal resistance which is inversely proportional to the local thermal conductivity, and thermal current $I_i^{th}$ is the power generated by Joule heating in the i'th device.

The lumped heat flow equation (6) can be derived by integrating and applying the divergence theorem to the differential heat flow equation (5), over each device on the

 0-7803-7826-1/03/$17.00 © 2003 IEEE

chip[4]. If the heat flow equation is integrated over the volume of each device, the following equation can be derived by using Gauss' theorem and taking heat capacity, $C$, and thermal conductivity, $\kappa$, as constants over the device volume.

$$C\int_V \frac{\partial T}{\partial t}dV = \kappa\int_S \nabla T dS + \int_V H dV \tag{7}$$

Taking $-\kappa\nabla T$ as the heat flux and discretizing $T$ between the devices yield the following equation, which is similar to a full KCL type nodal equation except the capacitive term which is only from the relevant node to ground.

$$\Omega_i C_i \frac{\Delta T_i}{\Delta t} + \sum_j \frac{\kappa_{ij}\Delta x_{ij}\Delta y_{ij}}{\Delta z_{ij}}(T_i - T_j) = \int_{V_i} H dV_i \tag{8}$$

Where subscript $i, j$ represents the value of the corresponding quantity between the nodes $i$ and $j$.

Using the electrical analogy, temperature and the integrated power density resemble voltage and current, respectively. Thus equivalent thermal resistances and capacitances between devices can be written as follows:

$$R^{th} = \frac{\Delta z}{\kappa \Delta x \Delta y} \tag{9}$$

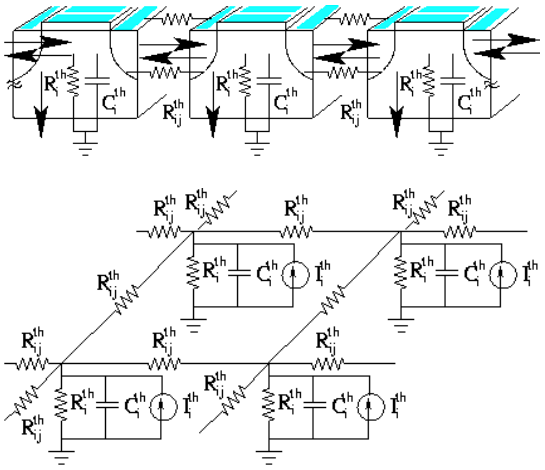
$$C^{th} = \Omega C \tag{10}$$

Note that the thermal resistance is defined between nodes, unlike the thermal capacitance which is only from the given node to ground. Once the appropriate values are used between the devices for the given geometry and materials, a methodology is developed to solve the $R^{th}C^{th}$ thermal network.

The complete solution of the equations (1)-(6) provides the key device quantities (potential, carrier concentrations, wave function, mobility, boundary conditions, etc.) as a function of internal device temperature, as well as the chip temperature profile. Figs. 1a and 1b illustrate the connection between the device and thermal network we solve.

## III. Numerical Approach

We solve the coupled device/chip performance and heating model using a block iteration method as illustrated in Fig. 2. While this may appear straightforward, it is complicated by the necessity to resolve two different fundamental scales of dimension at the same time. The main difficulty is to develop a methodology which allows us to calculate the internal device operation, and the temperature distribution of the entire chip simultaneously. To achieve this, we solve the device transport equations and the device heating equation using the nonlinear distributed device model. To obtain the temperature distribution for the entire chip, we solve the linear lumped model. We combine the distributed and lumped models using a mixed-mode Monte Carlo-type method.




Figs. 1a and 1b illustrate the mixed-mode problem. Fig. 1a shows devices and their thermal connections, Fig. 1b shows KCL-type thermal network

As shown in the algorithm flow-chart in Fig. 2, we first solve the linear $R^{th}C^{th}$ thermal network for the temperature at each node. This requires discretizing the time derivative in equation (8), and then writing it in matrix form. This gives the following equation for the $i, j$ node on the chip at the time step k.

$$C^{th}_{i,j}\frac{(T^k_{i,j} - T^{k-1}_{i,j})}{\Delta t} + \frac{(T^k_{i,j} - T^k_{i,j\pm1})}{R^{th}_{i,j\pm1/2}} + \frac{(T^k_{i,j} - T^k_{i\pm1,j})}{R^{th}_{i\pm1/2,j}} + \frac{T^k_{i,j}}{R^{th}_{i,j}} = I^{th}_{i,j}(T^k_{i,j}) \tag{11}$$

Here $R^{th}_{i,j\pm1/2}$ is the thermal resistance between the nodes $i, j$ and $i, j+1$.

We use a uniform random distribution of power sources, of values between zero and one, as input. We note that each node on the chip corresponds to an individual transistor. Thus, for large IC's the number of nodes is typically greater than one million. To solve a system, with such a large number of nodes, an iterative method (bilateral conjugate gradient method) is adopted. This relieves computational burden of the CPU and facilitates the simulation by eliminating the troublesome matrix inversions.

We next solve the internal nonlinear device equations for a single representative device using a combination of methods involving Scharfetter-Gummel discretizations[5], a QL implicit eigenvalue solver and Newton's method[5]. The results give the quantum corrected device characteristics, including the power generated by the representative device.

We then take advantage of the linearity of the $R^{th}C^{th}$ thermal network, to scale the full-chip temperature distribution using the heating power calculated in the single device simulation.


0-7803-7826-1/03/$17.00 © 2003 IEEE

We then iterate between the lumped chip model and the device model until temperature and current-voltage characteristics converge at the device and chip levels.

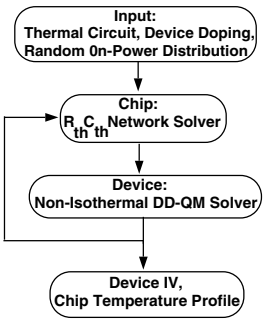



Figure 2. Coupled Flowchart

## IV. Simulation Methodology

We investigate non-isothermal device operation by incorporating temperature dependencies of some of the parameters used in the device equations (1)-(5). The following parameters are explicitly allowed to vary with temperature; intrinsic carrier concentration, thermal voltage, saturation velocity, carrier mobilities, boundary voltages, recombination lifetimes and the bandgap. The analytical expressions we use for the temperature variation of carrier mobilities, saturation velocity and boundary voltages are shown below.

$$\mu(T) = \mu(300K)\left(\frac{T}{300K}\right)^{-2.5} \quad (12)$$

$$\nu_{sat}(T) = \nu_{sat}(300K)\left(\frac{1+e^{\frac{T}{600K}}}{1+e^{0.5}}\right) \quad (13)$$

$$\phi(T) = V_{applied} + \frac{kT}{q} ln\left(\frac{n}{n_o(T)}\right) \quad (14)$$

After solving the device equations, including lattice temperature equation (5), we find that the temperature variation within a bulk MOSFET is usually less than one percent. This result is shown in Fig. 3. We take advantage of this small variation to facilitate obtaining the temperature of the entire chip, and approximate the temperature within a single device as uniform. The key result here is that we can treat the entire device as a single thermal node for the purposes of complete chip thermal analysis.

To solve for the heat flow and temperature throughout the chip, we need the power produced by each device. However, there may be well over one million devices. We circumnavigate this difficulty by choosing a representative device and activity profile. For the current work, we assume a random activity profile, based on a uniform activity distribution throughout the chip. If a snapshot of our simulated chip is taken at a random time instant, we would observe such a uniform random distribution in terms of the power consumed by each device. We utilize this activity profile as a zero order approximation to real world situations. Normally the activity profile for each chip would differ depending on its architecture and the application running on it. For example, a random access memory (RAM) would consume less power than a central processing unit (CPU), because considerably less power is used for storage purposes than computational purposes. Due to the nature of our algorithm, we can easily integrate different activity profiles and chip configurations to our model. Also if a transient solution is required, the chip temperature profile as well as the activity profile can be stepped in time.

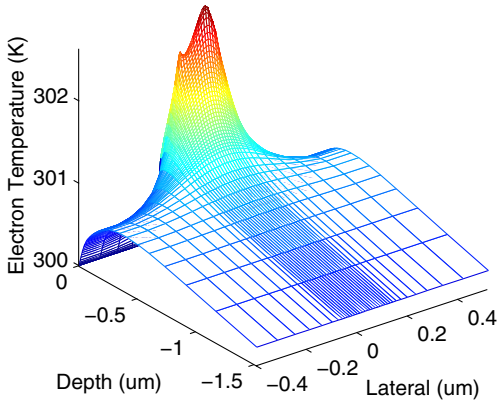



Figure 3. Temperature Profile in a 0.1$\mu$m MOSFET $V_{GS}$=1.0V, $V_{DS}$=1.0V

In this work, we take the hottest device as the representative one. It can also be chosen randomly from the chip. However our investigations show that most of the temperature variation takes place near the chip edges and on the chip package. Thus for the case we simulate, the resulting temperature profile would be almost the same no matter which device we choose to represent the ensemble of transistors.

Additionally, we take the average power consumed by each device as one tenth of the power consumed for the given bias conditions, which were $V_{GS}$=$V_{DS}$=0.7V. It is reasonable to assume that the typical device is in the switching state for one tenth of the clock cycle for logic applications.

## V. Simulation Results

Fig. 3 shows the electron temperature profile within an NMOSFET. The temperature reaches its peak value near the drain side, which is the far corner on the figure. The key observation is that the device is so small, and the thermal resistance is sufficiently low, that its internal temperature variation is negligible for full chip heating calculations.

In Fig. 4 we show the simulation results of a 0.1$\mu m$ MOSFET at local chip temperatures of 300K and 400K


0-7803-7826-1/03/$17.00 © 2003 IEEE

for three gate bias values. As expected, for the linear and saturation region current decreases with higher $T$. This is mainly due to decreased mobility.

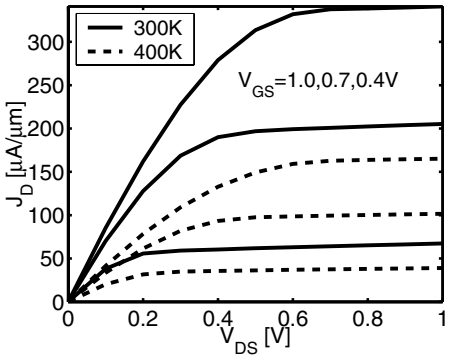



Figure 4. MOSFET IV Curves for T=300K and 400K; $V_{GS}$=0.4, 0.7, 1.0V

In Fig. 5 we show the calculated temperature profile for an entire chip. The chip is 0.5cm by 0.5cm, and has 1,000,000 transistors. The contour graph implies that the chip reaches its maximum temperature of 360K near the center. Note that the temperature distribution will depend highly on the boundaries. We take the external surfaces of the chip to be at room temperature and account for thermal conductivity of package.

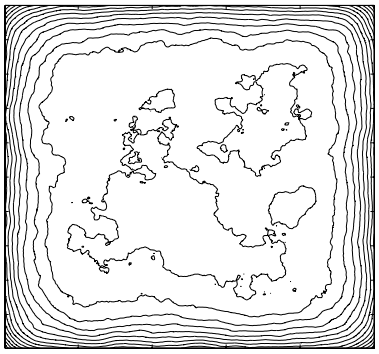


Figure 5. Temperature Profile: Temperature Isotherms range from 300K at chip edges to 360K inside chip

Fig. 6 shows that the effects of chip heating are significantly reduced if thermal contacts are fabricated in the chip. The thermal contacts are modeled as ideal heat sinks which set the chip temperature at the contact locations to room temperature.

Fig. 7 shows the increase in chip temperature as the number of devices and chip size increase.

The thermal capacitances of the lumped model are fundamental in determining how quickly a chip heats up to its steady state temperature value. In Fig. 8, we show the transient heating characteristics of a chip. The figure indicates that the time constant for chip heating is on the order of microseconds.

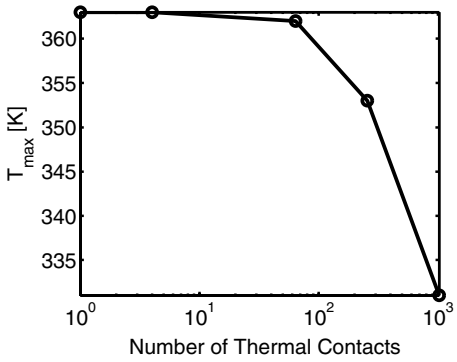



Figure 6. Maximum chip temperature as a function of uniformly distributed thermal contacts

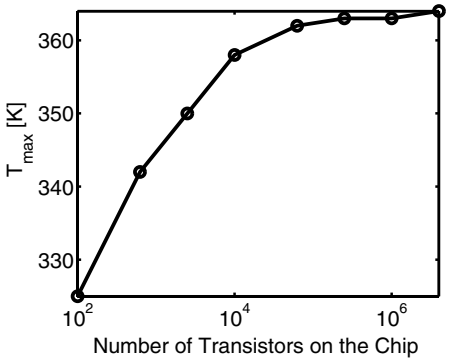



Figure 7. Maximum chip temperature for different device counts

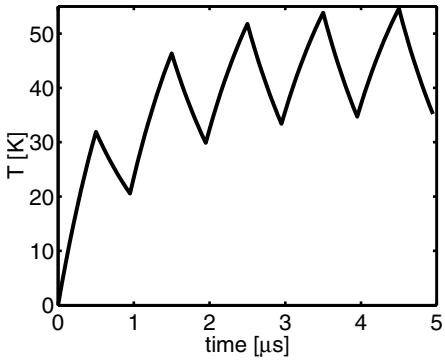



Figure 8. Thermal transients showing time to heat up according to average power

REFERENCES

[1] P. Gelsinger, Microprocessors for the New Millenium, *ISSCC 2001.*
[2] A. Stach, R. Sabelka and S. Selberherr, *Proc. 16th IASTED Int. Conf. on Mod. Ident, and Cont.* pp. 16-19, 1997.
[3] P. Wolbert, G. Wachutka, B. Krabbenborg, T. Mouthaan, *IEEE CAD of Integrated Circuits and Devices*, vol. 13, pp. 293-301, 1994.
[4] S. Lee, D. Allstot, *IEEE Solid-State Circuits*, vol. 28, pp. 1283-93, 1993.
[5] C.H. Chang, *Current and Next Generation TCAD: Advance in 2-D Semiconductor Device Modeling by the Hydrodynamic and Spherical Harmonic Boltzmann Methods*, Ph.D Thesis, University of Maryland, 1999.


0-7803-7826-1/03/$17.00 © 2003 IEEE