Mansoura Engineering Journal, (MEJ), Vol. 25, No. 2, June 2000. M. 46

A LUMPED TRANSIENT THERMAL MODEL FOR SELF-HEATING IN MOSFETS

نموذج حرارى غير مستقر و مدمج للتسخين الذاتى فى الترانزستور من نوع معدن-عازل-شبه موصل

MN Sabry*, W. Fikry**, Kh. Abdel Salam***,
M.M. Awad*, A.I. Nasser***

* Mech. Eng. Dept., Mansoura University, Egypt
, ** Eng. Physics Math Dept, Ain Shams Univ., (on leave to Mentor Graphics Egypt),
***. 10th of Ramadan Technological Institute, Egypt

ملخص البحث:

يدرس هذا البحث التسخين الذاتى فى الترانزستور من نوع معدن- عازل-شبه موصل، بغية إنشاء نموذج حرارى مدمج مناسب. أخذت الظواهر المستقرة و غير المستقرة فى الاعتبار. نتيجة لتحليل أهمية العوامل المختلفة و باستخدام أدوات تحليل مبسطة، أمكن الخروج بكل من طريقة لحساب آثار التسخين الذاتى و أيضا هيكل عام للنموذج الحرارى المبسط الذى يمكن له أن يأخذ فى الاعتبار مختلف الظواهر الأساسية المتعلقة بالتسخين الذاتى. تمت مقارنة النموذج الحرارى المدمج بنتائج تجارب معملية أجريت فى بحث منشور، و تبين من المقارنة جودة النتائج مع بساطة النموذج.

Abstract

Both static and dynamic effects related to self-heating in MOSFETs (Metal-Oxyde-Silicon Field Effect Transistor) are studied in order to construct an adequate compact thermal model. An available 2D electrical device simulator in addition to a simple 2D finite difference code for the heat equation are used as analysis tools. These tools are used both to justify the proposed model topology as well as to extract model parameters. Both static and dynamic effects predicted by the model are compared with existing experimental results

Accepted June 30, 2000

1- Introduction

Self-heating effects (SHE) in MOSFETs were first reported by Takacs & Trager, 1987. It is quite well established that the increase in lattice temperature due to SHE degrades electron mobility, may result in a negative differential resistance, alters threshold voltage, increases source and drain series resistance, and greatly modifies leakage current at zero bias. The effects are even more pronounced in the case of Silicon-On-Insulator (SOI) CMOS devices (Workman et al. 1998) due to the thermal insulation of the buried oxide layer. The increased awareness of electro-thermal interactions on different levels (device, circuit, PCB) did not yet lead to a systematic use of electro-thermal simulation. This is due to some common misunderstandings, as well as a lack of adequate tools and models.

Although most designers would recognize the importance of thermal effects, their interest is usually directed towards the "average" chip temperature, both in space and time. Temperature gradients over the chip surface are in fact usually moderate, which results in a common belief that it would be sufficient to know the space average temperature. However, a temperature difference as low as 0.5K between 2 transistors in a current mirror for example would result in an error of about 5%, which is enough to degrade performance. Hence it is important to allocate an accurate temperature per transistor that would of course be layout dependant.

Another, common wrong belief is that thermal transient phenomena occur at a big time constant (~10μs to 1ms) which would make them uncoupled with electric transient phenomena (time constant ~1ns to 1μs). This thermal time constant estimate is usually based on the whole chip size, or at best on a whole transistor. Hence it is common to model SHE by an equation of the form (Su et al. 1994, Chen et al. 1995):

$$T_j = T_{amb} + R_{th}\, I_d V_d \qquad (1.1)$$

where T_j and T_{amb} are respectively junction and ambient temperatures, I_d and V_d drain current and potential and R_{th} the so-called thermal resistance. Heat generation, however, is concentrated in a very narrow zone close to the chip surface. The channel temperature thus quickly rises, which directly impacts transistor performance, before heat dissipates through the whole transistor and hence the chip. The time constant, which is related to the early stage of channel temperature rise, is proportional to the channel size. Hence it is in fact much smaller than that of the whole transistor, and may be comparable to electrical time constants.

As for the tools and models availability, many contributions appeared concerning simulation tools and compact thermal models at different levels. For the package level, key contributions are those of Bar-Cohen et al. (1989) and Rosten & Lasance (1994). At the smaller scale level, which is the chip level, there have been a lot of recent contributions among them Szekely et al. (1995), Szekely (1996), Digele et al. (1996), Sabry et al. (1996) and Sabry (1999). At the finest detail level, which is the transistor level, compact models are, to the author's knowledge, still lacking. The main reason is the distributed character of thermal effects, as opposed to the lumped approximation commonly accepted for electrical phenomena. This requires complicated 3D transient electro-thermal simulations to obtain the temperature distribution as well as its time evolution, which is quite heavy.

The objective of the present work is to propose a simplified compact (lumped) transient thermal model for MOS transistor self-heating, which would be able to predict major first order effects, enabling thus electro-thermal simulation at the circuit level and design at a reasonable cost. Design tools usually offer the possibility to incorporate behavioral models. Through this feature, transistor thermal models could be easily incorporated on top of well-known electric models. The simple thermal model proposed here could thus be the first building brick in a "thermal library" that would greatly facilitate electro-thermal analysis.

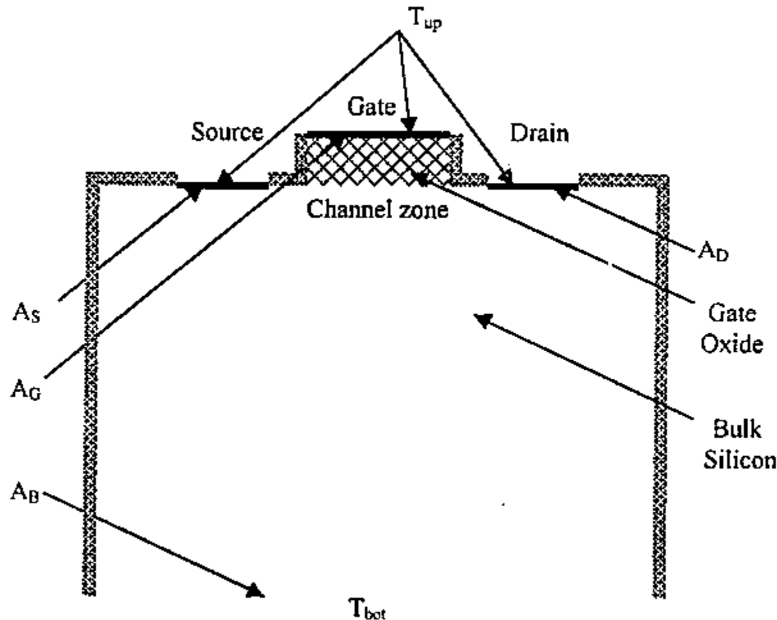

Figure 1- Problem description

2- Compact thermal model

In order to construct the compact thermal model, we need to define the following set of data:

- Governing equation,
- Boundary conditions,
- An adequate topology on which model parameters will be extracted.

The governing equation will be taken here as the heat conduction equation:

$$\rho c \partial T/\partial t = \nabla \cdot (k \nabla T) + q_v \qquad (2.1)$$

where ρ is the material density, c its heat capacity and k its thermal conductivity, T is the temperature, t the time, and q_v the rate of heat generation by Joule effect. The choice of the heat conduction equation should not be considered as evident. Transistor sizes are continuously shrinking, and hence hyperbolic effects will soon be important. The zone affected by heat generation extends over only few microns. At submicron

scales hyperbolic effects start playing a role in conduction (De Cogan 1998). One should be aware of the fact that the use of a parabolic equation is pushed here to its lower limit, and will soon be unjustified for smaller devices. Physical properties will be considered as constant. In fact thermal conductivity is temperature dependant, but temperature variations inside a transistor are not enough to induce large space changes of k. The simplification gained by this assumption is the transformation to a linear problem, which is an advantage that considerably outweighs the resulting small error.

The local heat generation q_v is obtained at each point by post processing results of the device simulator MINIMOS (Selberher et al. 1980, 1990) as follows:

$$q_v = (J_x^2 / \mu_x + J_y^2 / \mu_y) / q N \qquad (2.2)$$

where q is the electron charge, N the minority carrier concentration, μ and J are mobilities and current densities, the index (x or y) indicates the direction of μ and J. Note that MINIMOS calculates electrical properties as a function of a homogeneous temperature all over the transistor. Hence to obtain q_v the following iterative procedure was used:

a- Assume that the transistor operates at ambient temperature,
b- Perform device simulation using MINIMOS at this temperature,
c- Calculate q_v from MINIMOS output using (2.2),
d- Solve the heat conduction equation (2.1),
e- Obtain an estimate of the temperature in the channel zone,
f- Repeat above steps starting from b until convergence.

Current flows mainly in the channel zone, resulting in about half of the overall heat generated in this relatively narrow area (~0.5nm). Hence it is natural to take the channel temperature as representative of the whole transistor temperature in a simulator that depends on a single temperature. In a future extension to this work, full electro-thermal simulation will be performed to calculate simultaneously temperature and electric fields.

Adequate boundary conditions have to be supplied. Their selection is related to the topology of the sought for compact model. Since the present work is mainly interested by self-heating effect, this will allow some simplification of the problem that needs not be considered in its full generality. Source, drain and gate contacts are usually either metallic or poly-silicon, and hence have a high thermal conductivity. Therefore, each contact will be assumed to have a constant temperature along its surface (figure 1). The passivation layer (Silicon dioxide) covering the whole upper surface prevents heat transfer and hence a zero normal temperature gradient can be assumed outside the contacts. The transistor base temperature may vary in space. However, since the main phenomenon studied is self-heating, and since the channel is very far from the base, compared to the base size, the base temperature may also be assumed constant.

Two major assumptions will now be done that will be justified by the great simplifications they induce, while retaining at least first order effects. First, source, gate and drain temperatures will be assumed equal. This will considerably reduce the total number of thermal resistances and capacitances in the final compact model. It will still reflect basic features of heat transfer originating from self-heating. It will only eliminate part of the horizontal temperature gradient that is induced by a

neighboring transistor to another neighboring transistor through the one under study. The second assumption concerns sidewalls. Part of the heat generated in the channel, may flow through sidewalls to affect other transistors, which will be called bulk thermal coupling. Note that heat generated in the channel flows along either source or drain contacts before reaching sidewalls. These contacts have a high thermal conductivity and are at relatively lower temperature than the channel. Since heat is mainly generated in the channel zone near to the upper surface, then most of the heat will be absorbed by the contacts leaving only a small portion to the sidewalls. To validate this assumption two extremes will be investigated. Sidewalls will be assumed as either perfectly insulating, or perfectly conducting with an imposed temperature T_{amb}. Note that the second case extremely exaggerates heat flowing through sidewalls, since it brings points having a temperature equal to ambient temperature very close to the heat source, moreover these points have zero thermal impedance. Sidewalls will be placed at a distance equal to contact (source or drain) length on each side. Heat flowing through sidewalls in the second extreme, as obtained from simulation, is about 19% of the heat generated in the channel. Actual values of heat flowing through sidewalls are far less than this hypothetical upper bound, probably around 10% of the total heat generated. In cases where an oxide trench is used for insulating the transistor from its neighborhood, or in SOI technology, this proportion drops down to nearly zero. Neglecting this heat would result in a considerable simplification since we will not need to bother about bulk thermal coupling requiring heavy 3D transient simulations over the whole chip. Transistors may still be thermally coupled through interconnecting lines. Their thermal resistance can easily be estimated by considering them as fins.

To resume, boundary conditions are:

$$\begin{aligned} T\big|_{A_S+A_D+A_G} &= T_{up} \\ T\big|_{A_B} &= T_{bot} \\ \mathbf{n}\cdot\nabla T\big|_{\text{other walls}} &= 0 \end{aligned} \qquad (2.3)$$

where T_{up} is the temperature at source, drain and gate, T_{bot} is the base temperature, and $\mathbf{n}$ is the unit outward normal vector to the outer surface. Areas A_S, A_D, A_G and A_B are shown in figure 1. The steady state version of equation (2.1):

$$\nabla\cdot(k\nabla T)+q_v=0 \qquad (2.1')$$

is solved together with boundary conditions (2.3) using a finite difference code to give the steady temperature distribution inside the whole transistor. This field will be used in order to build the compact model describing each zone in the transistor.

In order to present the compact thermal model, the transistor will be split into four regions as shown in figure 2. The heat region lies directly below the grid. It extends horizontally to source and drain contacts and vertically down to cover the whole region where temperature field is 2D. Hence, it contains all points where heat is generated. Heat leaves the upper surface of this region to the gate contact, as well as side surfaces to both source and drain contacts, and finally the lower surface to the base. The base region is trapezoidal. Its shape is dictated by heat spreading from the heat zone of limited extent to the base. Remaining regions are named source and drain

regions. This temperature field will be used in the sequel to obtain basic characteristic quantities for each region. This includes:

- The maximum temperature in the heat region,
- The average temperature in each region (by integration over the volume of the required zone),
- The average temperature over each interface between different regions (by integration over the required interface)
- The steady heat flowing through each interface as well as outside surfaces (by integrating the temperature gradient over the required surface).

Each region will be represented by a lumped set of thermal resistances along heat flow paths, as well as a lumped thermal capacitance. An expression of lumped impedances in each region will be obtained by integrating (2.1) over the region volume. For example in the heat region, we get:

$$\rho C\, d/dt\left[(T_H - T_{up})\int \theta dv\right] + (T_H - T_{up})\oint(-k\, n\cdot\nabla\theta)da = \int q_v dv \qquad (2.4)$$

where θ is a dimensionless temperature defined as:

$$\theta = (T - T_{up}) / (T_H - T_{up}) \qquad (2.5)$$

and T_H is the maximum temperature in the heating region. It depends only on time, and manifests a strong dependence. However, the dimensionless temperature field θ is weakly dependant on time. In fact, it gives only the shape of the field, which is independent of the amount of heat generated. It may only depend on the location of heat generation. This may slightly vary with bias conditions, as well as with time, but as a first approximation it will be assumed that surface and volume integrals of θ are fixed. Hence volume and surface integrals in the Left Hand Side of (2.4) can be evaluated once and for all, using any simple configuration, including the steady case solved above. This is the well-known quasi-static approximation. From the steady temperature distribution obtained above one gets:

$$C_H \frac{d(T_H - T_{up})}{dt} + \frac{(T_H - T_{up})}{R_{HG}} + \frac{(T_H - T_{HS})}{R_{HS}} + \frac{(T_H - T_{HD})}{R_{HD}} + \frac{(T_H - T_{HB})}{R_{HB}} = Q \qquad (2.6)$$

Where:

$$C_H = \rho C \int \theta dv$$

$$R_{Hj} = 1 \Big/ \int_{A_{Hj}} (-k\, n \cdot \nabla\theta)da \qquad (\text{in which } j = G, S, D \text{ or } B)$$

$$Q = \int q_v dv$$

and T_{HS}, T_{HD}, T_{HB} are the average temperatures at the interface between heating region and source, drain and base regions respectively. Other regions are modeled in the same way, except that the node temperature is taken as the average temperature in each zone, not the maximum. This will give the equivalent circuit depicted in figure 2

Each thermal capacitance will induce a time constant. These were found by writing the lumped circuit system of differential equations and calculating the eigen values of the obtained system. Time constants are simply the inverse of these eigen values.

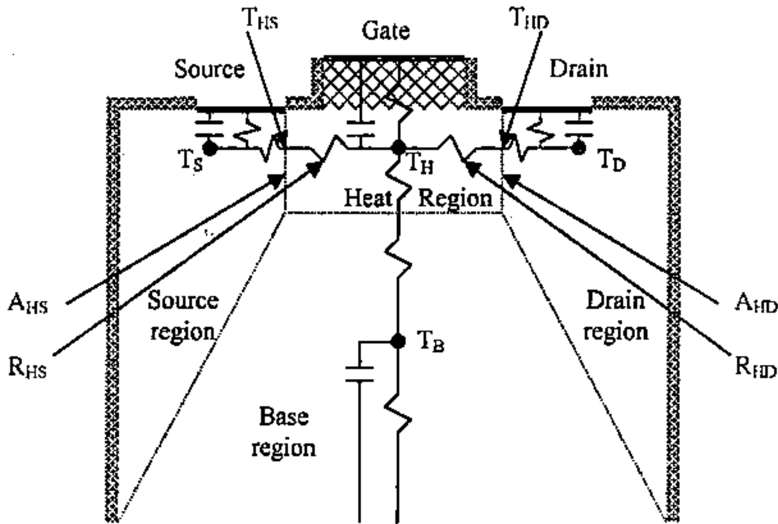

Figure 2 – Regions and model topology

Different authors (Rosten & Lasance 1994, Sabry 1999), have shown that star shaped compact models, such as the one proposed here, may not be an adequate form for a compact model describing the general problem. However, the model proposed here aims at describing self-heating effects at the transistor level. Transverse heat transfer due to different boundary conditions on source and drain for example is certainly less important than the heat transfer due to self-heating. The compact model proposed here can be viewed in fact as a "coarse mesh" finite element approximation.

3- Results and discussion

The lumped model proposed here, will now be compared with experimental results obtained by Mautry & Trager 1990. This work was selected because, to the authors' knowledge, it is the only one that presented thermal transient data for MOSFET devices measured at time scales of the order of a nano-second. They have used a 0.6μ technology, which was advanced by that time. Effects studied in this work, and revealed by their experiments, are even more valid in modern 0.25μ technology.

Unfortunately, they did not report sufficient data on some technological parameters, such as ion implantation and gate oxide thickness. The first step was to select a set of technological parameters that would produce I_d-V_d curves from MINIMOS that are as close as possible to their results at "cold" operation. By cold operation, they meant the case where junction temperature was maintained constant (350K in their experiments). MINIMOS and experimental results at cold operation are compared in figure 3.

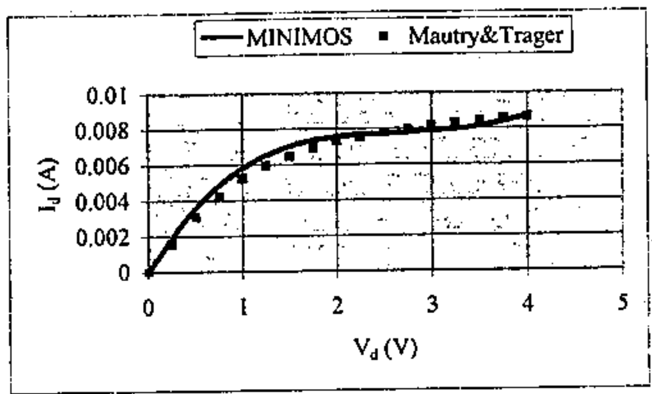

Figure 3- Drain current versus drain voltage
(for constant Gate voltage) at cold operating conditions

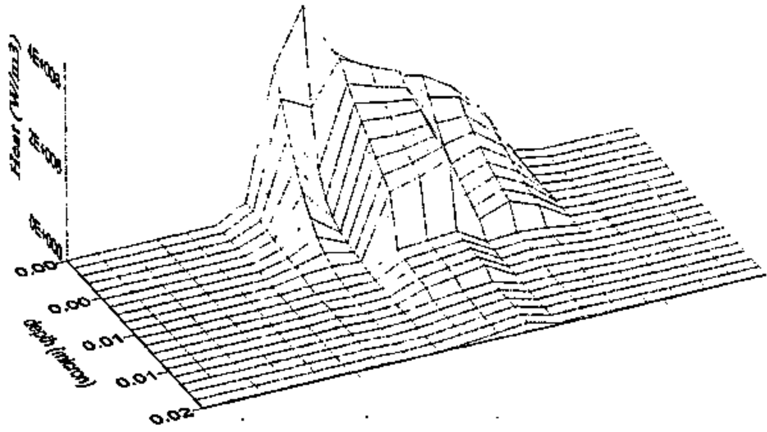

Figure 4- Typical distribution of heat generation

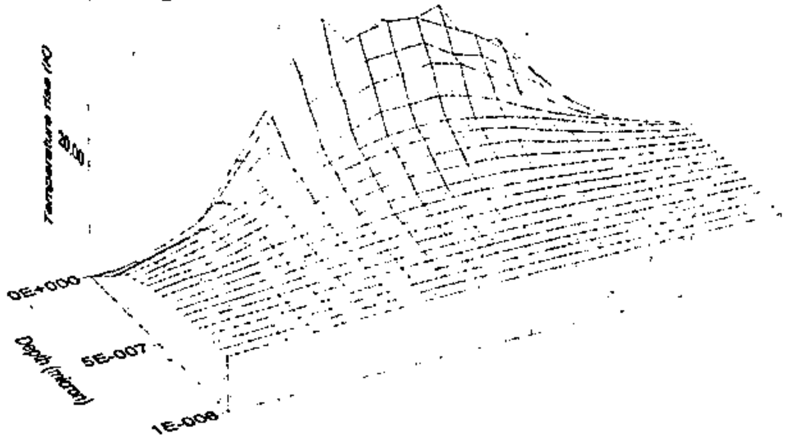

Figure 5- Typical temperature field

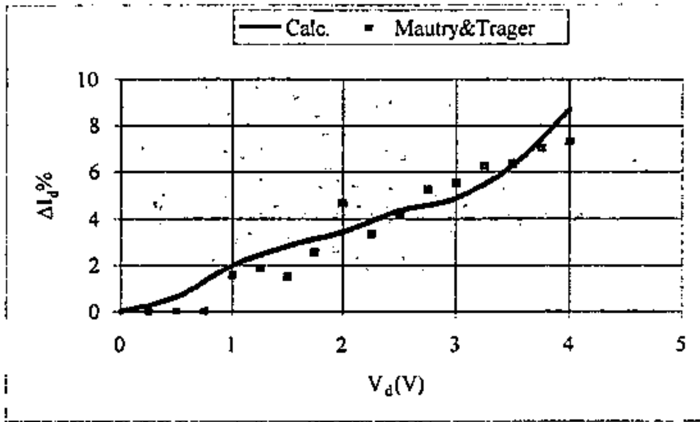

Figure 6- Percentage drop in drain current due to self heating

Typical distributions of heat generation and temperature fields are depicted in figures 4 and 5 respectively.

The second step was to validate the steady state resistive part, at "hot" operation. By hot operation they meant that only ambient temperature was maintained constant at 350K, but transistor temperature was higher due to self-heating. As a consequence of self-heating, electron mobility decreases, and hence I_d current at a given V_d is lower than the corresponding current at cold operation. No information whatsoever was given to estimate the thermal resistance between transistor upper surface and ambient air. Hence, its value was estimated from one operating point, and assumed constant in all other operating points. Also, since it was not possible to obtain identical values of I_d at cold operation, it would not be appropriate to compare theoretical and experimental I_d curves for hot operation, but rather the percentage reduction in I_d due to self-heating. Results are given in figure 6, showing a good agreement. The scatter in experimental results is mainly due to the difficulty in measuring small differences (between I_d hot and I_d cold curves) from the published figure of Mautry & Trager (1990). Finally, to study thermal transient effects, they have presented transient results in response to a V_d pulse of rise time 2ns. They claimed a measurement resolution of 3ns. As expected, results fit very well with an error function curve, especially in the early stages just after the pulse starts. Error function can describe distributed system behavior with a very good accuracy. Other methods have been proposed for modeling transient effects in a distributed form, among them the structure function (Szekely et al. 1999) and the S-parameter method (Sabry 1999). However, our interest here is in extracting a lumped model, which may be less accurate, but significantly simpler than the distributed one. Hence, another approach is needed in order to extract lumped compact model parameters. The results of Mautry & Trager (1990) are reproduced in figure 7, in a slightly modified way. The ordinate here is: $-\log((I_d - I_{d,steady}) / I_{d,steady})$, while the abscissa is still time. Hence the slope of the curve should directly give the inverse of the time constant.

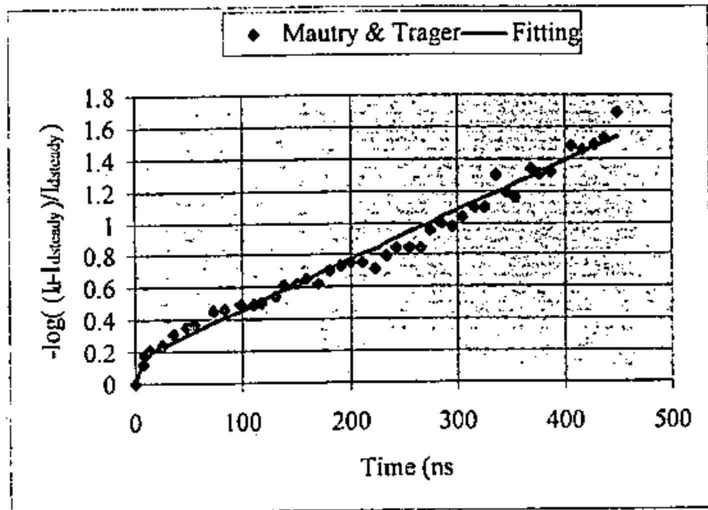

Figure 7- Measured thermal transient variations of drain current

It is clear from the figure, that there is a constant slope that dominates most of measured data corresponding to a time constant of 322ns. However, the straight line obtained by regression does not pass through the origin, indicating that another smaller time constant prevails at the early phase. The separation between early and late phases is not well defined in the obtained figure. The late phase contains a large number of data points. Hence fitting gave results of good quality independent of the selected starting point. The early phase, on the other hand, contains very few data points: only 2 to 3 points (depending on where we put the separation) plus the origin. Taking into consideration the high resolution error (3ns) compared to the whole range of the early phase (9 to 13ns), regression would give uncertain results. In fact, regression gave a time constant that is much higher than 13 ns, which is not realistic. All what can be said is that the early phase time constant lies somewhere between 0 and 13ns. These time constants were compared with those of the simplified compact model proposed in this work in table 1. The smallest one corresponds to the capacitance in the tiny heat region. It lies in fact within the expected range for the early phase. The following two time constants correspond to the drain and source regions respectively. They compare very well with the time constant of the late phase extracted from measurements. The last time constant corresponds to the rather huge base region. Measurements ended at 500ns, and hence the effects of this time constant were not detected in experimental data. It is clear that the proposed model was successful in obtaining at least the correct order of magnitude for both early and late phase time constants.

The main outcome of this work is an adequate methodology as well as a compact model topology to study self-heating effects in MOSFETs. Results can be refined further either by extensive electro-thermal simulation and/or by extensive testing. But both approaches require an adequate equivalent thermal network to extract results, which can thus be that proposed in the present work (figure 2).

Table 1- Comparison of measured and predicted time constants

	Early phase	Late Phase
Measured	0 – 13 ns	322 ns

	Heat region	Drain region	Source region	Base region
Predicted	5.3 ns	284.7 ns	354.1 ns	3.01 μs

4- Conclusion

A modeling methodology was proposed, that yielded a simplified compact thermal model of self-heating effects in MOSFETs. The upper and lower bound analysis has shown that thermal coupling between different transistors occurs mainly through interconnects and not through the bulk (or substrate). This greatly facilitates the construction of simplified electro-thermal models of different devices. The proposed modeling methodology was applied to a particular case where static and transient measurements were available. It yielded a model that satisfactorily described static (steady state current reduction due to self-heating) as well as dynamic (time constants) behavior of the device under test. The model also showed that transient effects inside a single transistor could be modeled by a small set of time constants. Some of them are indeed comparable to electrical time constants, contrary to the common belief. Compact model parameters could be further refined using adequate electro-thermal device simulators and/or extensive experiments, using the same extraction methodology proposed here. The simplicity of the proposed model will make electro-thermal simulations at the circuit level accessible, and hence will considerably improve circuit design.

Nomenclature

A	Area, m^2
a	Area element, m^2
C	Thermal capacitance, J/K
c	Heat capacity, J/kg K
I	Current, A
J	Current density, A/m^2
k	Thermal conductivity, W/m K
N	Carrier density, m^{-3}
n	Unit outward normal to the surface
Q	Total heat generated, W
q	Electron charge
q_v	Volumetric heat generation rate, W/m^3
R	Thermal Resistance, K/W
T	Temperature, K
t	Time
V	Potential, V
v	Volume element, m^3

Greek symbols:

ρ Density

θ Dimensionless temperature

μ Mobility,

Subscripts

amb Ambiant

B Base

bot Bottom

D,d Drain

G Grid

H Heating zone

j Junction

S Source

x,,y Coordinate directions

References

Bar-Cohen A., Elperin T. and Eliasi R., "Θ_{jc} characterisation of chip packages-Justification, limitations and future", IEEE Trans. CHMT, vol. 12, pp. 724-731, 1989

Digele D., Lindenkreuz S. and Casper E., "Fully coupled dynamic electro-thermal simulation", Proceedings of the 2nd THERMINIC Workshop, Sept. 25-27, Budapest, pp. 73-77, 1996.

Chen Y. et al., "An analytical drain current model considering both electron and lattice temperature simultaneously for deep submicron ultra-thin SOI NMOS devices with self-heating", IEEE Trans. Electron Devices Letters, vol. 42, pp 899-906, 1995.

De Cogan D., "The relationship between parabolic and hyperbolic TLM models for heat flow", THERMINIC IV, Cannes, France, September 27-29, pp. 197-202, 1998.

Mautry P.G. and Trager J., "Investigation of self-heating in VLSI and ULSI MOSFETs", Proc. IEEE Int. Conf. Microelectronic Test Structures, Vol. 3, pp21-226, March 1990.

Rosten H. and Lasance C., "DELPHI: The development of libraries of physical models of electronic components for an integrated design environment", Proc. Conf. Elec. Pack. Soc., Atlanta, GA, 1994.

Sabry M.N., Bontmps A., Aubert V. and Vahrmann R., "Realistic and efficient simulation of electro-thermal effects in VLSI circuits", IEEE Trans. VLSI systems, Vol. 5, No 3, 283-289, 1996.

Sabry M.N., "Static and dynamic modeling of ICs", Microelectronics Journal, 30, pp. 1085-1091, 1999.

Selberher S., Schutz A. and Potzl H., "MINIMOS a 2D MOS transistor analyzer", IEEE Trans. Electron Devices, vol Ed. 27, pp 1540, 1980.

Selberher S. et al. "The evaluation of MINIMOS mobility model", Solid State Elctronics, vol 33, No 11, p 1425, 1990.

Su L.T. et al., "SPICE model and parameters for fully depleted SOI MOSFETs including self-heating", IEEE Electron Device Letters, vol. 15, pp. 374-376, 1994.

Szekely V., Poppe A., Rencz M., Csendes A. and Pahi A., "Self consistent electro-thermal simulation: fundamentals and practice", Proceedings of the 1st THERMINIC Workshop, Sept. 25-2, Grenoble, pp.89-93, 1995.

Szekely V., "THERMODEL: A tool for dynamic thermal model generation", Proceedings of the 2nd THERMINIC Workshop, Budapest Sept. 25-27, pp. 21-26, 1996.

Szekely V. et al., "Transient thermal measurements for dynamic package modeling: new approaches", THERMINIC 5, Rome, Italy, October 3-9, pp. 7-11, 1999.

Takacs D. and Trager J., "Temperature increases by self-heating in VLSI CMOS", ESSDERC 1987, Bologna, pp. 59-62, 1987.

Workman G.O., Fossum J.G., Krishnan S. and Pelella M., "Physical modeling of temperature dependence of SOI CMOS devices and circuits including self-heating", IEEE Trans. Electron Devices, vol 45, pp. 125, 1998.