INT. J. NUM. METH. HEAT FLUID FLOW, VOL. 2, 291–298 (1992)

HEAT SOURCES AND TEMPERATURE DISTRIBUTION IN INSULATED GATE BIPOLAR TRANSISTORS

K. BOARD AND P. A. MAWBY

Department of Electrical & Electronic Engineering, University College of Swansea, Singleton Park, Swansea SA2 8PP, UK

INTRODUCTION

The insulated gate bipolar transistor (IGBT)[1–3] is an increasingly used power transistor which has the advantage over the more conventional DMOS structure of achieving a lower on-resistance through the high-injection of electrons and holes into the drift region thereby causing conductivity modulation and a lowering of the electrical resistance.

The real issue underlying the need for a low on-resistance is, however, the requirement that the device in its on-state does not exceed a specified maximum temperature. These devices are in all cases thermally limited and careful design is required to ensure that this maximum is not exceeded as it has an important bearing on reliability. It is not sufficient, however, simply to overdesign the part, as this involves increasing the silicon area consumed. The latter, of course, is extremely price-sensitive.

The usual way of establishing the temperature rise in these structures is to make the assumption that heat dissipation occurs uniformly over certain volumes within the device. While such an approach gives an indication of average temperature rise it does not take account of local hot-spots that may occur. Thus it is possible for such an approach to indicate that spatially averaged temperatures are below the required limit while locally the temperature may be well in excess of that value.

In this paper the semiconductor equations are solved in two dimensions using a finite element approach and the resulting current density, electric field and recombination process throughout the domain used to calculate the heat dissipation at each position within the device. The heat equation is then solved to provide the temperature distribution accurately. The temperature dependence of the thermal conductivity of silicon is included in the calculation.

THEORY

The defining equations for the semiconductor are:

Poisson's equation

$$\nabla \cdot (\varepsilon_s \nabla \psi) = -q(p - n + N_D^+ - N_A^-) \tag{1}$$

where ψ is the electrostatic potential and N_D^+, N_A^- are the donor and acceptor impurity densities.

0961–5539/92/040291–08$2.00

Received March 1991
Revised February 1992

electron continuity equation:

$$\frac{\partial n}{\partial t} - \frac{1}{q}\nabla\cdot\mathbf{J}_n = G - R \tag{2}$$

hole continuity equation

$$\frac{\partial p}{\partial t} + \frac{1}{q}\nabla\cdot\mathbf{J}_p = G - R \tag{3}$$

Equations (2) and (3) apply onto to semiconductor regions, inside insulator regions only Poisson's equation is solved. The current densities in (2) and (3) are given by:

$$\mathbf{J}_n = -q\mu_n n\nabla(\psi + \theta_n) + qD_n\nabla n \tag{4}$$

$$\mathbf{J}_p = -q\mu_p p\nabla[-(\psi + \theta_p)] + qD_p\nabla p \tag{5}$$

where μ_n and μ_p are the electron and hole mobilities respectively and D_n and D_p are the electron and hole diffusivities which are related to the mobilities by Einstein's relation. The quantities θ_n and θ_p are 'band-parameters' associated with the conduction and valence bands. They represent the shift in the band edges due to either heavy doping effects in silicon, or to compositional variations in non-uniform materials, such as heterojunctions. Boltzmann statistics are used to describe the carrier concentrations:

$$n = n_{i_{ref}}\exp[(\psi + \theta_n - \phi_n)/U_t] \tag{6}$$

$$p = n_{i_{ref}}\exp[(\phi_p - \psi - \theta_p)/U_t] \tag{7}$$

where ϕ_n, ϕ_p are the quasi-Fermi potentials for electrons and holes, U_t is the thermal voltage and $n_{i_{ref}}$ is the intrinsic carrier concentration referred to the main semiconductor region.

In a device consisting of both semiconductor and oxide as in the present application, the current continuity equations are only solved in the semiconductor, whereas Poisson's equation is applied to the whole domain. A contact existing on the oxide region gives its applied bias to the electrostatic potential and the semiconductor/oxide interface acts as a Neumann boundary condition to the continuity equations in the semiconductor. Ohmic contacts existing on semiconductor regions force the quasi-Fermi potentials to assume the applied bias. This is equivalent to enforcing infinite recombination velocities and charge neutrality.

Recombination in the model is controlled by two mechanisms, namely Shockley–Read–Hall and Auger recombination, important under high injection conditions. Shockley–Read–Hall recombination is expressed by:

$$R = \frac{np - n_{ie}^2}{\tau_p(n + n_{ie}) + \tau_n(p + n_{ie})} \tag{8}$$

Auger recombination is defined by:

$$R = (C_n n + C_p p)(pn - n_{ie}^2) \tag{9}$$

where n_{ie} is the effective intrinsic carrier concentration and τ_n, τ_p the recombination lifetimes for electrons and holes.

Heat generation within the semiconductor is given by[4]:

$$P = \mathbf{J}\cdot\mathbf{E} + (R - G)(E_g + \tfrac{3}{2}kT) - (\mathbf{J}_n + \mathbf{J}_p)(\tfrac{3}{2}k\nabla T + \tfrac{1}{2}\nabla E_g) \tag{10}$$

where the first term represents classical Joule heating due to both hole and electron currents, and the second term accounts for energy exchange with the lattice through recombination or generation of electrons and holes $3/2\,kT$ above and below the respective band edges, in agreement with average thermal values. The third and fourth terms provide corrections for non-uniform temperature distribution, and band gap variation.

In this analysis the effect on the heat dissipation of temperature variations are not taken into account. Also band gap narrowing effects are ignored so that the fourth term in (8) may also be taken as zero. Equation (8) is in agreement with Adler's 1-dimensional formulation[5] apart from the $\pm 3/2\,kT$ energy shift.

Finally, the temperature may be obtained from:

$$\nabla \cdot (k \nabla T) = -P \tag{11}$$

where P is given by (10) and k is the thermal conductivity, given, for silicon by[6]:

$$k(T) = 1.54\left(\frac{T}{300}\right)^{-4/3} \tag{12}$$

METHOD OF SOLUTION

The package developed to solve (1)–(3) in two dimensions in the steady state employs a triangular-element mesh, which can easily be made to conform to any device geometry. Using a triangular mesh has the added advantage that local refinement is strictly confined to regions of rapid change, which is of a very high order in the semiconductor solution.

The adopted discretization scheme employs the box integration method (BIM) otherwise known as the control region approximation (CRA). It relies on the divergence theorem of Gauss to convert the divergence terms in (1), (2), (3) and (11) into line integrals enclosing the Voronoi region surrounding each mesh node. *Figure 1* illustrates a typical node (i), its surrounding neighbours and the Voronoi region associated with that node. Equation (1) can be rewritten:

$$\nabla \cdot \vec{D} = \rho \tag{13}$$

where the right hand side is the charge density (ρ) and then integrated over the Voronoi area (shaded in *Figure 1*). The area integral of the left hand side can then be converted to a line integral by applying the divergence theorem of Gauss to give:

$$\oint_{\Gamma} \hat{n} \cdot \vec{D}\, d\Gamma = \int\int_{\Omega} \rho\, d\Omega \tag{14}$$

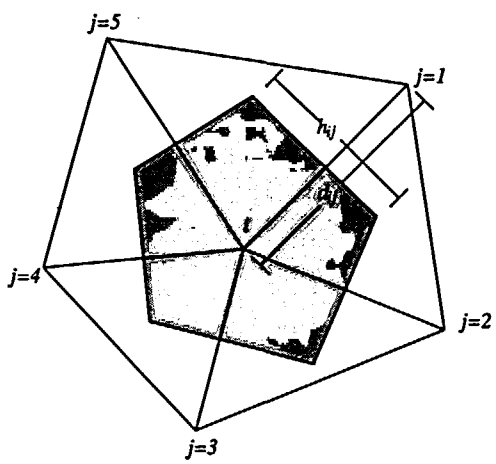

Figure 1 Voronoi region associated with a particular node

This is done by assuming the outward electric flux density D along each facet of the Voronoi area is constant, reducing the line integration for each node to a simple summation as follows:

$$\sum_{j=1}^{M_i} D_{ij} d_{ij} = \rho_i A_i \tag{15}$$

where D_{ij} is the electric flux density between the nodes i and j, and d_{ij} is the width of the connecting pipe, ρ_i is the nodal charge density, A_i is the nodal Voronoi area and M_i is the number of edges in the Voronoi region associated with the node i. The relationship between D and the scalar nodal potentials is given simply by assuming that the field is piecewise constant, thus

$$D_{ij} = \varepsilon \frac{\psi_i - \psi_j}{h_{ij}} \tag{16}$$

Note that the minus sign associated with this expression has been taken account of by transposing the positions of the two nodal potentials ψ_i and ψ_j, h_{ij} is the length of the pipe between nodes and ε is the dielectric constant. Equation (15) can be written as a column vector of rank N, which is equal to the number of nodes in the mesh:

$$\vec{F}_{\psi}(\psi) = 0 \tag{17}$$

where the ith component of F_{ψ} is:

$$\vec{F}_{\psi i} = \sum_{j=1}^{M_i} D_{ij} d_{ij} - \rho_i A_i \tag{18}$$

The steady state current continuity equations (2) and (3) are discretized in the same way so that for each node (i) we end up with the equation:

$$\vec{F}_{ni} = \sum_{j=1}^{M_i} J_{nij} d_{ij} - (G_i - R_i) A_i \tag{19}$$

$$\vec{F}_{pi} = \sum_{j=1}^{M_i} J_{pij} d_{ij} - (G_i - R_i) A_i \tag{20}$$

where the current density terms along edges (J_{nij} and J_{pij}) are evaluated using the standard Scharfetter–Gummel scheme. The discretized heat equation similarly becomes:

$$\vec{F}_{Si} = \sum_{j=1}^{M_i} S_{ij} d_{ij} - P_i A_i \tag{21}$$

where S_{ij} is the heat flux between nodes i and j and is calculated in an analogous way to (16):

$$S_{ij} = k \frac{T_{Li} - T_{Lj}}{h_{ij}} \tag{22}$$

where k is the temperature dependent thermal conductivity and T_{Li} is the lattice temperature at node i.

The resulting non-linear discretized equation set is solved using the Newton–Raphson method. The method of solution is illustrated in *Figure 2*. The user can select which equations are solved from (1)–(3), that is, Poisson's equation only, or Poisson's equation coupled to either or both of the continuity equations. Automatic back-tracking is applied to the solution process should a bias increment prove too large, ensuring that a solution is always reached for a particular bias point.

The underlying linear equation set is solved using the incomplete Choleski conjugate gradient method (ICCG) for symmetric cases and the conjugate gradient squared method (CGS) for non-symmetric equations. The Bank and Rose sparse-storage scheme is used to store all the

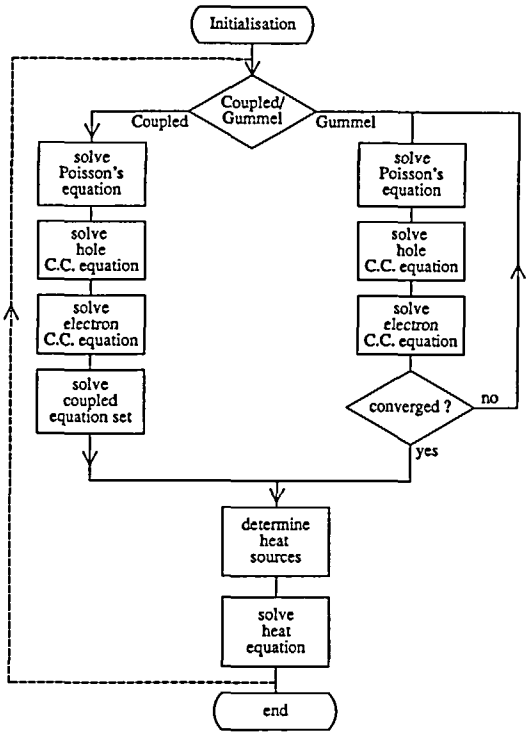

Figure 2 Solution scheme for the calculation of temperature

assembled matrices which makes optimal use of the available memory, thus minimizing the use of virtual memory.

RESULTS AND DISCUSSION

A schematic diagram of the device structure is shown in *Figure 3a.*

The upper N^{++} and P^{++} regions are represented by Gaussian diffusions, with the P^{++} regions forming the transistor channel where it comes to the silicon surface under the oxide. The gate regions extends over the N^- drift region and provides an accumulation layer along which current can spread before flowing towards the drain. Since accumulation layers are typically only tens of Angstroms thick it is essential that the mesh in these regions is very fine in the vertical direction. The mesh used is shown in *Figure 3b.* The log of the impurity density magnitude throughout the device is illustrated in *Figure 4,* and shows clearly the two Gaussian diffusions together with the P^+ drain and N^+ buffer regions. The lowering of on-resistance provided by carrier injection from the P^+ region near the drain is illustrated in *Figure 5* which shows the concentration of holes throughout the device at a drain bias of $+1$ V. The hole concentration in the drift region is 4×10^5 cm^{-3} at zero bias so that the self induced bias on the lower junction has raised the hole concentration to 10^{16} cm^{-3}—over eleven orders of magnitude and over one order of magnitude greater than the impurity density. The electron concentration for this bias was virtually the same confirming the existence of the high injection condition.

For each voltage bias point the solution of the potential and current densities is followed by the calculation of the two relevant heat dissipation terms given in Reference 10.

The Joule heating component of the total heat dissipation is illustrated in *Figure 6a* with a line plot of the same quantity along the top surface of the device shown in *Figure 6b.* Of particular interest are the large localized positive and negative peaks in dissipation near this surface. These

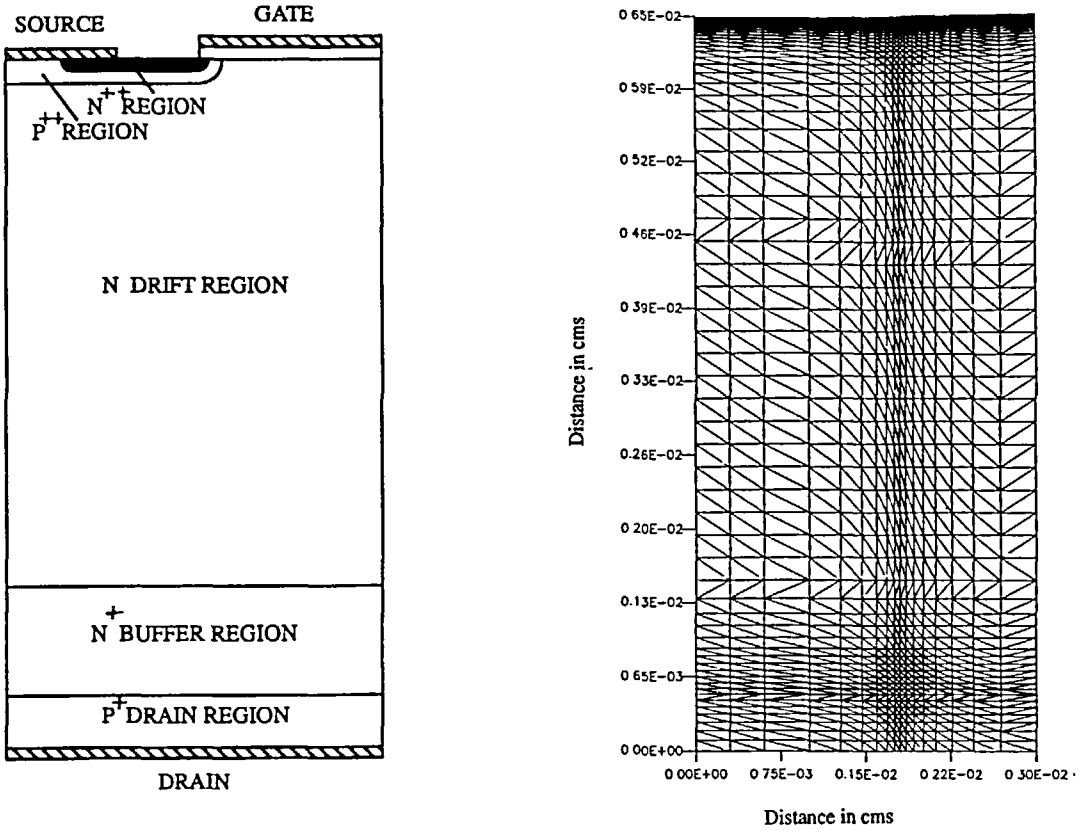

Figure 3 IGBT device configuration: (a) schematic; (b) finite element mesh

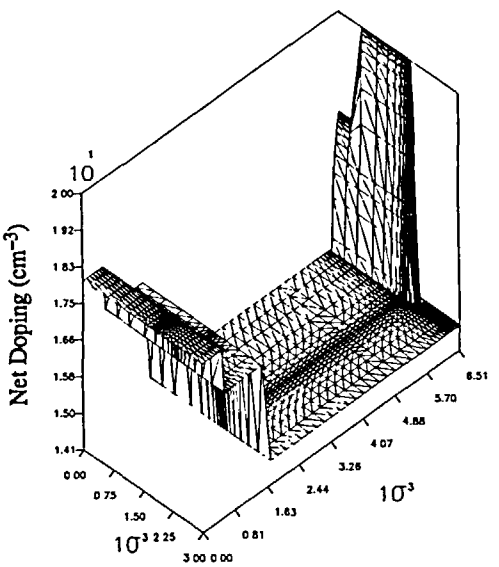

Figure 4 Impurity density throughout device domain

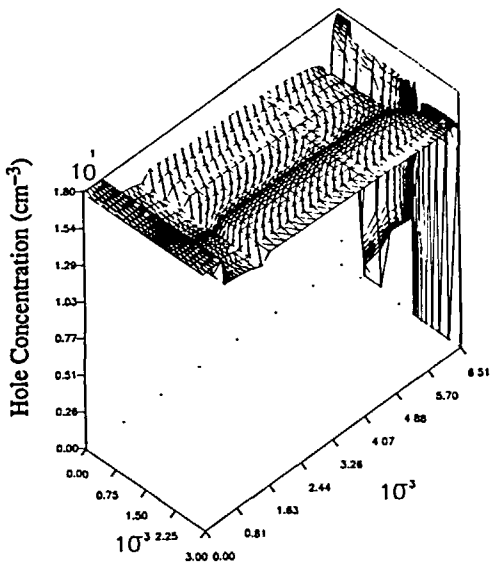

Figure 5 Hole concentration throughout device with $V_D = 1$ V

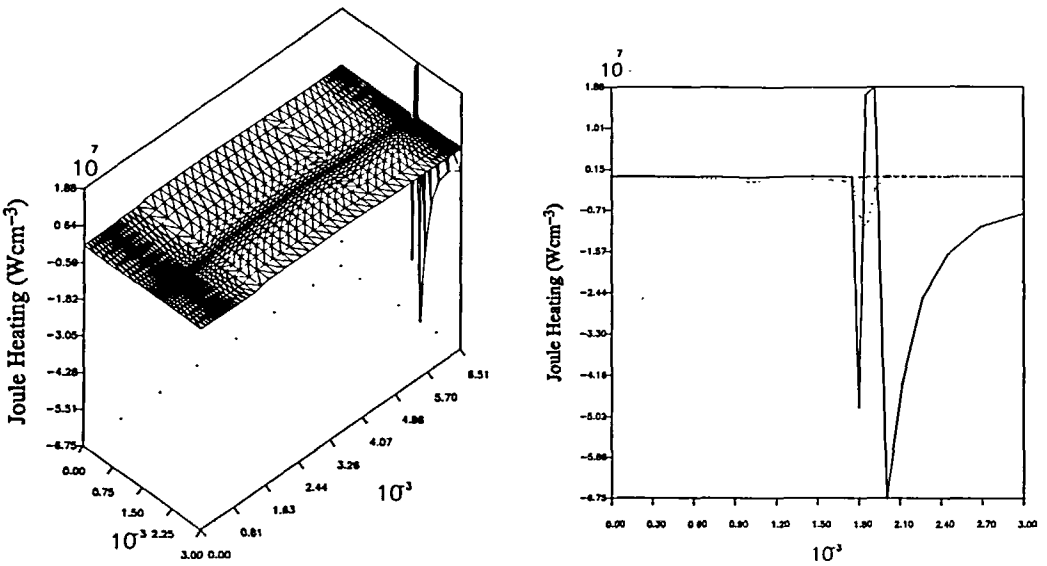

Figure 6 Joule heating within the IGBT: (a) throughout device; (b) line plot along top surface

are explained physically as follows: moving from left to right along the surface and referring to *Figure 6b* a sharp negative peak is first encountered. The negative sign implies cooling and occurs when the current is flowing against the electric field. This first negative peak therefore occurs because electrons entering the channel have to surmount a small potential barrier represented by the difference between the built-in potential of the source-body junction and the induced surface potential at the source end of the channel. Over the length of the channel the electrons flow with the electric field which is considerable because of the space charge penetration of the drain end. Moreover, the current density there is very high because the flow is constricted to the narrow surface channel. This gives rise to large positive peak in the electric field-current density product. On emerging from the channel the electrons enter an even narrower surface channel forward by the accumulation layer under that portion of the gate that lies over the N^- drift region. From this accumulation layer the electrons flow down towards the drain in proportion to the local current density giving rise to the usual exponential decay of current along this layer. In flowing out of the accumulation layer the electrons have to mount a small potential barrier—hence the negative sign of $J.E$ over this portion of the surface. The exponential decay of $J.E$ reflects the variation of the current density as described above. Over the bulk of the domain and away from these localized peaks the dissipation density is relatively constant and several orders of magnitude lower than these peak values. Also the dissipation density in the drift region is much lower at $V_D = 1$ V than in low injection because of the effect of conductivity modulation.

The second component of heat dissipation density arises from the recombination process which is fairly pronounced in the IGBT because of the presence of high densities of both carrier types, is illustrated in *Figure 7*. The vertical scale, which represents the number of electron hole pairs recombining per unit volume, is converted to dissipation density by the factor $(Eg + 3kT)$ (see (10)). Thus the peak dissipation density is 6.8×10^3 W/cm^3—much lower than the peak dissipation due to Joule heating (1.88×10^7 W/cm^3). The recombination heating is seen to be positive or near zero throughout the device and the broad peak near the channel reflecting the fact that most of the electron flow is in this vicinity. The sharp peak in the lower drain junction occurs along the whole length of that junction and is centred in the space charge region, as expected. The recombination shown in *Figure 7* corresponds to a carrier lifetime of 10^{-6} sec and the magnitude in this case such as to have a negligible effect on the overall dissipation. In

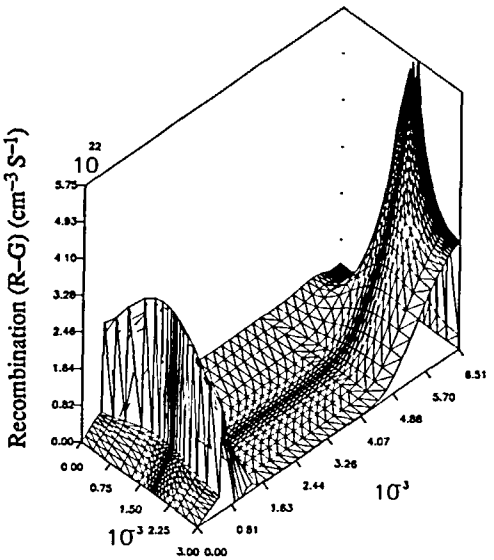

Figure 7 Recombination heating within the IGBT

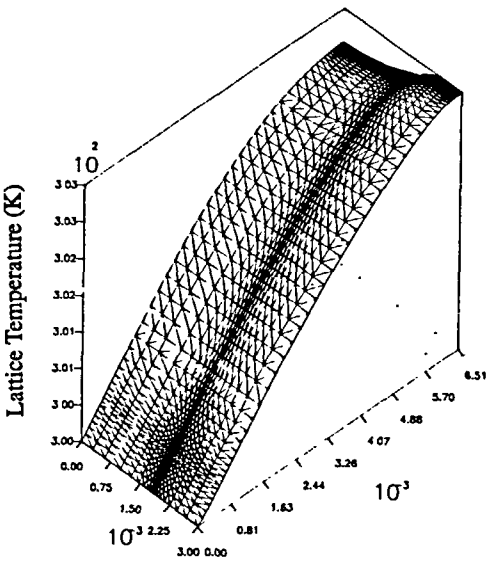

Figure 8 Temperature distribution within the IGBT

practical IGBT structures the lifetime is likely to be very much shorter so that contribution to total dissipation from the recombination process is likely to be more dominant. *Figure 8* shows the resultant temperature distribution obtained by solving (11) using (10) and taking account of the temperature dependence of the thermal conductivity using (12). A surprising feature is the absence of local hot spots arising from the highly localized power dissipation shown in *Figure 5*.

CONCLUSIONS

The Joule heating throughout the IGBT under high injection conditions is highly non-uniform over the device domain, showing large positive and negative peaks in the vicinity of the MOS channel and extended-gate accumulation regions.

The corresponding temperature distribution calculated using this dissipation distribution and taking the temperature dependence of the thermal conductivity into account shows no corresponding peak. Negative heat dissipation is observed in the accumulation region which could account for the absence of any pronounced temperature peaks at the location of the high, localized dissipation peak.

The dissipation throughout the drift region was found to be relatively uniform and of a low value compared with the low injection condition. This was attributed to the role of conductivity modulation in reducing the electrical resistance of this region.

REFERENCES

1 Baliga, B. J. The insulated-gate rectifier in a new power switching device, *IEDM Tech. Dig.*, pp. 264–267 (1982)
2 Russel, V. J. P. The COMFET—a new high-conductance MOS-gated device, *IEEE Electron. Dev. Lett.*, **EDL-4**, 63–65 (1983)
3 Darwish, M. and Board, K. Lateral resurfed COMFET, *Electron. Lett.*, **20**, 519–520 (1984)
4 Wachutka, G. Rigorous thermodynamic treatment of heat generation and conduction in semiconductor device modelling, *Simulation of Semiconductor Devices and Processes*, Vol. 3, pp. 83–96, Alma Mater Studorium, Bologna (1988)
5 Adler, M. S. Accurate calculations of the forward drop and power dissipation in thyristors, *Proc. IEEE Trans. Electron Dev.*, **ED-25**, 16–22 (1978)
6 Gaur, P. and Navon, D. H. Empirical law for the temperature dependence of the thermal conductivity of silicon, *IEEE Trans. Electron Dev.*, **ED-23**, 50–57 (1976)