Applied Surface Science 39 (1989) 578–596
North-Holland, Amsterdam

THE PHYSICS OF HOT-ELECTRON DEGRADATION OF Si MOSFET's: CAN WE UNDERSTAND IT?

M.V. FISCHETTI, S.E. LAUX and D.J. DIMARIA

IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA

Received 29 March 1989; accepted for publication 31 March 1989

We discuss some of the difficulties we must overcome to analyze the "hot-electron problem" in Si MOSFET's from a first-principle point of view. Using our Monte Carlo–Poisson device simulator DAMOCLES, we find that many physical elements are very important in determining the gate current in short-channel devices. Examples are given for effects whose importance was expected (band-structure and short-range carrier–carrier interaction), and for some whose importance was underestimated (such as the long-range plasma effects at the channel–drain region). Fortunately, recent results show that only electrons reaching the SiO_2-gate interface with energy in excess of about 2 to 3 eV can damage the oxide. We discuss the important technological consequences of this finding.

1. Introduction

In this paper we want to follow a "first-principle" approach to analyze the "hot-electron problem" [1] in small Si MOSFET's. We shall adopt a model-case which we can use to judge the status of our understanding. Restricting our attention to n-channel devices, we shall break the hot-electron problem into four idealized steps: (1) electron heating in the Si channel, (2) their injection into the gate oxide, (3) their transport in the SiO_2 layer, and, finally, (4) the SiO_2 degradation during electron injection. Admittedly, this is an oversimplified picture, which ignores complications such as hot-electron degradation of p-channel devices, or avalanche and "hole-injection" effects, among others. (We shall not touch again on these issues, but we realize their importance.) Furthermore, we shall discuss only the first and last steps of this model-case. Since it is our purpose to show that even the idealized case is beyond our present "first-principle" understanding, at least as far as the crucial step (1) above is concerned, these added complications will simply reinforce our pessimism. Better news comes from our comprehension of the electronic processes occurring in the final step of the process, as we shall see in our concluding section.

0169-4332/89/$03.50

2. Electron transport in Si

Models based on the simple *drift-diffusion* (DD) approximation [2] to solve the Boltzmann transport equation in semiconductors have been generally recognized as unable to provide any information about high-energy carriers in MOSFET's. This is due to their basic approximations (carriers *more or less* at thermal equilibrium and transport parameters which are *local* functions of the electric field) which hold only for small fields and devices large enough so that the electrostatic potential is slowly varying over a distance scale of several carrier-mean-free-paths. Despite these limitations, the use of DD models has been widespread – with some success – even outside their strict limits of validity, thanks to ad hoc improvements to treat typical hot-carrier effects, such as the existence of a saturated velocity (implemented via empirical field-dependent mobility models [3]) or impact-ionization phenomena (usually implemented via the use of models based on the "lucky electron" concept [4,5]). Looking at the "macroscopic" behavior of the devices (by this, we typically mean average quantities like terminal currents) these improvements have allowed predictions which can be surprisingly correct even in very small devices, with the engineer's experience often providing a good guideline to select the input parameters. Quite different is the situation when dealing with hot-electron effects: In this case DD models (and their ad-hoc improvements) begin to yield results which are even qualitatively wrong as we move to the 0.5 μm channel length.

Next in the hierarchy of transport models, are the models based on higher moments of the transport equation (the so-called "hydrodynamic" or "energy-transport" models [6–8]). They extend the validity of the DD models by relaxing the quasi-thermal and local approximations. They do so by making some implicit or explicit assumptions on the carrier distribution function and solving for one extra moment (the average carrier energy). Thus, they are able to account for some nonlocal phenomena, such as velocity overshoot and ionization coefficients [9] dependent on the local electron energy (rather than the local electric field, as done by the "lucky electron" model). However, in the context of hot-carrier effects, the assumptions made on the distribution function to derive the moment equations – and the crucial energy and momentum relaxation times – prevent us from the onset from gaining any accurate knowledge of the very high-energy carriers.

We are left with Monte Carlo (MC) simulations [10] as the only viable tool to address such problems as electron injection into the gate insulator. But also in this case we must face several difficulties, although, admittedly, with MC models we begin to consider "first principle" difficulties. To list a few among the most important ones: (1) Conventional MC simulations employ drastic simplifications in describing the band-structure of the semiconductor (typically, parabolic or first-order nonparabolic bands are used [11–13]). (2)

Confusion dominates the literature about the scattering parameters required to describe high-energy carrier–lattice interactions [14,15]. (3) Impact ionization is a process that has remained largely uncharacterized, despite its importance in controlling the high-energy tails of the carrier distribution functions [14]. (4) The long and short-range Coulomb interaction between carriers is usually ignored. (5) Quantization in the Si inversion layer is thought to be important only at low biases [16] (and even there, no satisfactory explanation exists [17] for the low value of the electron saturated velocity seen experimentally [18,19]), but no proof has been given that this is true. Moreover, the scattering processes in the channel (mainly, interface scattering) are treated somewhat crudely [20]. (6) Finally, at energy and fields large enough, the Boltzmann equation itself may fail and no consensus can be found in the literature about the time and energy scale up to which we can rely on semiclassical transport [21].

An attempt to tackle the hard problem of "predicting" (as opposite to "fitting") gate currents in a real MOSFET was begun by our group a few years ago. The strategy was the step-by-step inclusion of many of the physical ingredients listed above. The underlying hope was that a few of these ingredients could have been easily recognized as "important", while others could have been dismissed as "second-order-effects" equally easily. Our experience to date is that this is not the case. Essentially, the problem stems from the very high accuracy which is needed to obtain information about the high-energy tail of the carrier distribution function. We are forced to look at part-per-million effects or worse. Even ignoring the statistical problems imposed on the MC scheme, it is not surprising that minor perturbation to the physics seem to have dramatic effects on the quantity we would like to compute.

We shall present below two examples of what we have just said: The first example shows how strongly the hot-carrier effects in a 0.25 μm n-channel MOSFET depend on the band-structure of the semiconductor. This is an anticipated result which justifies our choice of employing the full empirical-pseudopotential band-structure of Si in our MC code. The second example deals with an even more dramatic and largely unexpected result: the strong effect of the long-range electron–electron interaction on the high-energy tail of the electron distribution function. A brief outline of our Monte Carlo simulator is in order before we focus on these examples.

2.1. *The DAMOCLES program*

Our self-consistent two-dimensional-Poisson/Monte Carlo simulation program (DAMOCLES: Device Analysis with MOnte CarLo Et poiSson. Obviously the acronym came first, its meaning later) has already been extensively described in the literature [15,22,23]. Here we wish simply to stress that our starting point has been the belief that only by using a better description of the

band structure of the semiconductor (from local empirical pseudopotentials [24] in our code) can we have any hope of describing carriers with as much as 3 eV of kinetic energy (as needed to study electron injection over the $Si-SiO_2$ barrier). This point was pioneered very convincingly by Hess and coworkers [14,25]. We have decided to extend their scheme to study not only the carrier kinematics, but also their dynamics by computing the carrier–phonon scattering rates using the "exact" density-of-states (DOS) obtained from the band-structure. Where phenomenology still enters is in the selection of the coupling constants (the deformation potentials). We have fitted them to available experimental data on bulk and homogeneous situations. Similarly, other scattering processes, such as carrier-ionized impurities and Coulomb carrier-carrier interactions, are also evaluated from the Fermi golden rule and Born approximation respectively. In this case, no phenomenological help is required, but a simple Debye–Hückel (non-degenerate) or Thomas–Fermi (degenerate) screening parameter is used. Better approximations are probably needed. Impact ionization is still treated very crudely with a Keldysh formula [14] and no quantum effects in the inversion layer are included as yet. In view of our lack of two-dimensional transport, the presence of the interface is handled empirically with a mixture of specular and diffuse scattering. The program allows the simulation of electrons and holes (also simultaneously for true bipolar operation) and the analysis of GaAs devices, all at 77 or 300 K.

2.2. Band-structure effects

As a test case to investigate the effect of our choice for the model of the Si conduction band, we have simulated an experimental n-channel Si MOSFET with an effective channel length of about 0.23 μm [23,26]. We show in fig. 1 the effect of three different band-structure models on: (a) the average electron energy at a distance of 1 nm from the $Si-SiO_2$ interface with 2.5 V applied both at the gate and at the drain, (b) the average electron velocity in the source–drain direction along the same "cut" of the device, and (c) the electron energy distribution just inside the highly doped ($\sim 2\times 10^{20}$ cm^{-3}) drain region. The parabolic band model greatly exaggerates both average energy and velocity. The "usual" first-order nonparabolic model [27] behaves more reasonably, but – as seen in (c) – lacks the DOS details that the full-band-structure model reveals. Of course, how different the distribution functions shown in (c) really are is a somewhat subjective matter. As in every comparison between models, similarities and differences depend on the answer we are after. In this case, if all we look for is the average electron energy, the pseudopotential and the nonparabolic model differ by less than a factor 2 along the entire channel. On the contrary, if we look at the substrate currents predicted by the models, the difference grows larger, the nonparabolic model yielding a current larger by almost an order of magnitude. Both direct

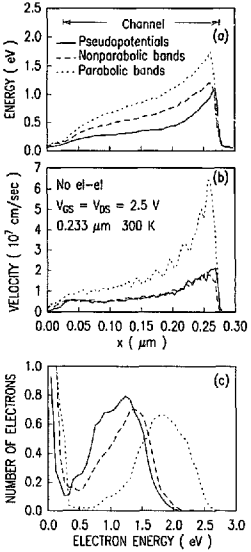

Fig. 1. Different band-structure models are used to simulate the electron average energy (a) and velocity along the channel (b) 1 nm away from the $Si-SiO_2$ interface in an n-channel MOSFET having an effective channel length of 0.233 μm. In (c) we show the electron energy distributions just inside the drain ($x = 0.275$ μm in (a) and (b)).

kinematical effects (the band-structure itself) and indirect dynamical effects (the DOS in the computation of the scattering rates) combine to give these results. It is interesting to note that two L-valley minima at about 1.1 and 2.9 eV (this last one doubly degenerate) appear [15] at energies where the substrate and gate currents have their respective thresholds. The lower velocities of carriers in these minima and the larger scattering rates to the L-valleys both influence the observed distribution function. It is not surprising that simpler band-models miss the "correct" (?) behavior altogether.

A nontrivial result is the dramatic difference between the parabolic and the pseudopotential model which results from these considerations. Almost all models proposed to-date to estimate substrate and gate currents in short-chan-

nel MOSFET's are based on (1) strictly parabolic bands and (2) the "lucky-electron" model [5] (with or without some " local-equilibrium" models [28]). The latter model has been shown not to hold true far away from thermal equilibrium by Hess and coworkers [14]. Here we see how badly the first approximation fails. As much as one could stress the different philosophy behind our approach and a more "realistic" (or "engineering") approach, we should question the validity of those simple models whose outcome is a set of empirical parameters (such as mean-free-paths [5,28] and effective $Si-SiO_2$ barriers [5,28,29]) probably devoid of any physical meaning in this context.

2.3. Intercarrier Coulomb-interaction effects

It has been reported in the literature that substrate and gate currents in short-channel MOSFET's can be measured at biases lower than their respective "intuitive" thresholds (about 1.2 and 3 eV, respectively) [28]. Even ignoring the fact that there is a non-negligible probability of phonon absorption (particularly acoustic phonons at large energies) which can excite carriers to energies above the applied bias, the short-range interparticle Coulomb interaction (i.e. interactions of particles within a screening length of each other) should be very effective in doing so: A "hot" carrier could receive extra energy from a "warm" carrier and find itself at energies higher then the maximum voltage drop in the device. In fig. 2, we show that this is indeed the case: for the same device of fig. 1, this time simulated at 77 K, carriers are found above the 2.5 eV energy threshold, unlike what is found when switching off the short-range interparticle Coulomb interaction. Again, this is an ex-

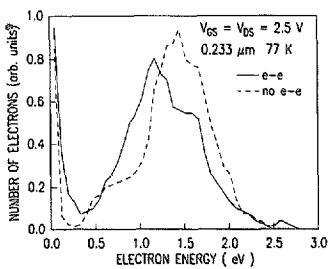

Fig. 2. The short-range Coulomb electron–electron interaction is turned on and off to simulate the electron energy distributions just inside the drain, as in fig. 1. Note the "dip" at about 2.5 eV in the distribution including the short-range interaction, related to a minimum of the DOS in the Si conduction band.

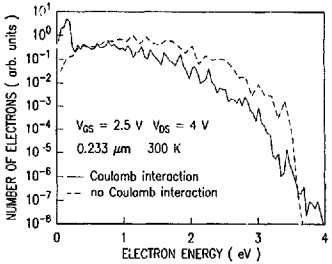

Fig. 3. Electron energy distributions close to the drain region ($x = 0.265$ μm in figs. 1a and 1b) obtained from the simulation of the 0.233 μm long device of figs. 1 and 2 with and without the short- and long-range Coulomb interaction. Note the effect of the plasma losses (shift to lower energy of the "Coulomb" curve) and of the short-range single-particle collisions (higher energy tail and low energy peak).

pected result which confirms the importance of the interparticle coupling when we want to analyze the high-energy tail.

Totally unexpected is the result we show in fig. 3: this time we have biased our 0.25 μm test-device to a higher drain voltage (an overstressing 4 V, just to deal with electrons above 3 eV with better statistics) and we have compared results of two simulations. The first simulation was performed by first reaching steady-state with a MC–Poisson self-consistent run and by turning the short-range interparticle interaction off. After steady-state had been reached, the self-consistent field was "frozen" by moving our Monte Carlo particles, but without solving the Poisson equation. (We tested this mode of operation in order to save the computer-time needed to solve the Poisson equation while implementing a time-consuming scheme to enhance statistically the high-energy tails to get information about very high energy carriers [13,30].) This mode of operation effectively shuts off also the long-range Coulomb coupling between carriers.

This coupling – mediated by the self-consistent Poisson solution we obtain during the simulation – accounts mainly for the potential fluctuations resulting from the random motion of the carriers: In heavily doped regions – where the effect is most notable – carriers subject to the random thermal collisions with phonons redistribute themselves. As some of them statistically accumulate more in one region, they expose more positive charge (the ionized donors) in other areas, generating a field which attracts the carriers back where they came from. The resulting *plasma oscillations* are real physical effects which we tried to treat properly by solving the Poisson equation at

frequencies large enough (typically every 0.2 fs) to avoid undersampling them. (This would result in a catastrophic instability of the electron + field system.) Since these plasma excitations are waves sustained by charge-density oscillations (not the transverse waves sustained by the transverse electric-field/magnetic-field Maxwell "coupling", as in the case of electromagnetic waves), in most cases we can get-by without solving the whole set of Maxwell equations, and use the (non-retarded) electrostatic approximation (the so-called optical surface-plasmons are the most important case which escape this simplifying approximation [31]). Other subtle points must be carefully taken into account to treat correctly in two space dimensions this intrinsically three-dimensional effect, as we have discussed elsewhere [15].

In a second simulation, we have turned the full Coulomb-coupling on, by restoring the self-consistency in the Monte Carlo-Poisson approach and by also including the short-range electron–electron collisions. We can see in fig. 3 that the typical short-range Coulomb-coupling feature appears very strongly: a significant number of electrons are pushed at energies above the applied voltage drop (about 3.5 V below the source Fermi level around the spatial locations at which the distribution function has been plotted) and at very low energies, while carriers are removed from the mid-energy region. Note the low-energy peak in fig. 3 which results from both the excitation of carriers in the drain Fermi-sea by plasma oscillations as well as from "warm" and "hot" carriers which have lost their energy in single-particle collisions. But this is not the most impressive difference: clearly, the entire distribution seems to be down-shifted by about 0.25 eV, roughly equal to the energy of plasma oscillations in the highly doped drain.

"Smart after the fact", we may now claim that this result should have been expected: It is well known that energy losses to collective excitations (our long-range plasma excitations) dominate over losses to single-particle excitations (our short-range interparticle collisions) in most cases. Pines gives two strong arguments why this should be so [32]: (1) the total density of states available to the collective mode is much larger, and (2) the carrier–plasma coupling is less sensitive to screening. While the first effect probably dominates in metals, the second argument works strongly in our case: As electrons flow along the channel, they are in a relatively low-density configuration, particularly in the pinched-off region (density $= 6 \times 10^{17}\,\mathrm{cm}^{-3}$ in the example of fig. 3). Thus screening is not very effective. As they approach the drain, the large pool of cool carriers there begins the process of screening the incoming hot carrier by polarizing itself. It is during this very process that plasma oscillations arise. The drain electrons induce a fluctuating field in the pinched-off region as they move back-and-forth in the drain. Most of the channel carriers will "ride" the field with random phases and will end up being decelerated even before entering the drain. This corresponds to the emission of a plasmon in the drain, to use a second-quantization jargon.

To draw a few conclusions, we can note that the substrate currents we extract from the results of fig. 3 are about a factor 3 apart, the full-Coulomb interaction showing the smaller value (about 15 and 5 $\mu A/\mu m$, respectively). We carefully avoid any comparison with experimental data, given the uncertainty of our impact-ionization model. If we simply count the number of electrons above a somewhat arbitrary energy threshold of 3.2 eV, we find that the difference now exceeds a factor of 25. On the contrary, more electrons appear to be above 3.5 eV when the full Coulomb interaction is included, despite the smaller values this models yields for substrate and gate currents, and average electron velocity and energy in the pinched off region. If, as we shall argue later, the degradation of very thin oxides (4.5 nm in our case) is triggered by carriers entering it with significant energy, the larger "predicted' damage will result from the model exhibiting the smaller average heating and the smaller substrate and gate currents.

2.1. Quantization effects

It is common wisdom that quantization effects in the silicon inversion layer have little effect on the hot-carrier characteristics. We have no reason to doubt this generally accepted point of view, as long as only terminal currents or average quantities are concerned. However, the lesson learned from the long-range Coulomb coupling has taught us to expect major differences in the hot-carrier tails induced by previously ignored processes. Moreover, the results of section 2.2 indicate that we must preserve the pseudopotential band-structure of Si when quantizing channel-carriers. Indeed, even hot carriers will be quantized in lower-lying subbands when their momentum is mainly directed along a direction parallel to the $Si–SiO_2$ interface, and the necessary band-structure features must be retained in each subband. We shall briefly explain how this can be done (to our knowledge no solution of the problem has been ever given when a pseudopotential band-structure is considered) and speculate about the effects we might expect, in the absence of simulation results.

The "Schrödinger-like"equation we must solve is [33]:

$$[\epsilon(-i\nabla) + eV(z)]\ \psi(r,z) = E\psi(r,z), \tag{1}$$

where z is the coordinate normal to the $Si–SiO_2$ interface, r the coordinate along the plane parallel to the interface, $\epsilon(k)$ is the energy dispersion (so that $\epsilon(-i\nabla) = -(\hbar^2\nabla^2)/(2m_{\text{eff}})$ for a parabolic band approximation, m_{eff} being the carrier effective mass in the quantization direction), $\hbar$ is Planck's constant divided by 2π, e is the electron charge, $V(z)$ is the electrostatic potential in the inversion layer, ψ is the envelope wave function, and E is the eigenvalue. A solution of eq. (1) *accounting for the local pseudopotential band-structure* [24] can be obtained by imposing that the envelope wave function vanishes at the $Si–SiO_2$ interface ($z = 0$) and at a point z_0 sufficiently far from the interface.

Then we can expand ψ around the local conduction-band minima, $k_0 = (K_0, K_{z,0})$, over plane-waves of wavevector K parallel to the interface and a sine-wave Fourier series, $\sin(n\pi z/z_0)$, n being an integer, along the normal direction:

$$\psi(r,z) = \sum_{K,n} \exp[i(K - K_0)\cdot r]\; a_n(K)\sin\left(\frac{n\pi z}{z_0}\right) = \sum_{K,n} \psi_{K,n}(r,z). \tag{2}$$

Eq. (2) originates essentially from an expansion over the pseudo-wave functions obtained computing our band-structure. Thus, we can evaluate the effect of the operator $\epsilon(-i\nabla)$ on each $k = (K, K_{z,0} + n\pi/z_0)$ component of ψ:

$$\begin{aligned}&\epsilon(-i\nabla)\;\psi_{K,n}(r,z)\\ &= \tfrac{1}{2}\left[\epsilon\left(K, K_{z,0} + \frac{n\pi}{z_0}\right) + \epsilon\left(K, K_{z,0} - \frac{n\pi}{z_0}\right)\right]\psi_{K,n}(r,z)\\ &\quad + (\text{orthogonal cosine terms})\end{aligned} \tag{3}$$

Taking the matrix elements of eq. (1), denoting by $V_{m,n}$ the matrix elements of $V(z)$ between two Fourier sine-states, and accounting for the orthogonality of the functions $\psi_{K,n}$, for every K we can find a matrix equation for the coefficients $a_n^{(v)}(K)$:

$$\begin{aligned}&\tfrac{1}{2}\left[\epsilon\left(K, K_{z,0} + \frac{n\pi}{z_0}\right) + \epsilon\left(K, K_{z,0} - \frac{n\pi}{z_0}\right)\right] a_n(K) + \sum_m e V_{n,m}\; a_m(K)\\ &= E(K)\; a_n(K).\end{aligned} \tag{4}$$

Inverting eq. (4), we get the eigenvalues $E_v(K)$ (the energy of a carrier with parallel wavevector K in the vth subband) and the eigenvectors (the "envelope" wave function ζ_v in the vth subband in the normal direction):

$$\zeta_v(K,z) = \sum_n a_n^{(v)}(K)\sin\left(\frac{n\pi z}{z_0}\right). \tag{5}$$

In fig. 4, we show the resulting dispersion obtained in the inversion layer of the device and bias conditions of fig. 1 around mid-channel. Fig. 4a shows the situation in a valley having the heavy, longitudinal mass along the normal direction. The opposite situation (transverse mass along the quantization direction) is shown in fig. 4b. Note how the parabolic and pseudopotential minima of the subbands differ, particularly in fig. 4b. Note also that the dispersion in the ground-state subband deviates significantly from the parabolic dispersion at high energies and in the transverse (110) direction away from the minimum at k_0. Finally, we should stress that, unlike what is found in a parabolic band approximation, we cannot simply factor-out the dependence of the envelope wave function $\zeta_v(K,z)$ on K, since we cannot do the same for the dispersion $\epsilon(k)$: every carrier with different parallel component

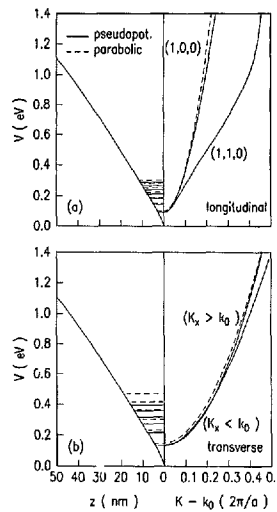

Fig. 4. Bottom of the first 6 subbands (left) and $E_0(K)$ – dispersion of the ground state (right) – obtained at mid-channel in the device of fig. 1 with the empirical-pseudopotential and the parabolic model for the conduction bands of Si. In (a) the quantization is done along the longitudinal-mass direction, in (b) along the transverse-mass direction. In (a) we show the dispersion of the ground state in the (1,0,0) and in the (1,1,0) directions away from the conduction-band minimum at $k_0 = (0,0,0.85)$ $(2\pi/a)$. By definition, the parabolic model yields the same dispersion in every direction. In (b) we show the dispersion when moving along the (1,0,0) direction from the band-minimum $k_0 = (0.85,0,0)$ $(2\pi/a)$ towards the zone center ($K_x < k_0$) and towards and across the zone edge at the symmetry point X into the second conduction band of the next Brillouin zone ($K_x > k_0$).

of momentum will be associated to a different envelope wavefunction $\zeta_v(K, z)$. The effect is not large in most cases, but it represents a significant complication.

Fig. 4 shows that the same dramatic differences between parabolic and empirical-pseudopotential band-structure can be expected also in a quantized situation. We cannot show which effect the quantization has on the high-energy tails of the electron distribution, as we are only beginning to equip DAMOCLES with these features. We can only speculate. Apart from obvious

kinematic differences, deviations (if any) from a three-dimensional situation may stem from the different scattering rates that the quantized carriers will have. This is due to the relaxation of the conservation of normal momentum: the delta-function is replaced by the "fuzzy" conservation represented by an overlap integral between initial and final envelope functions ζ. On the other side, in reasonably long devices, as large as this effect might be, at very high energies the carrier motion is strongly randomized by the large-angle umklapp collisions. We can expect that very few high-energy carriers will have a normal component of k small enough to make quantum effects important. Furthermore, the diffusive motion will push many carriers away from the quantized interfacial region. (In very short devices, when quasi-ballistic transport takes place – such as the experimental 0.07 μm devices operating at 77 K [15,23,26] – scattering is less effective and the kinematics of a two-dimensional electron gas may effectively propagate from source to drain, thus resulting in a significant modification of the transport properties.) This is the common wisdom. But this effect, together with the still largely unexplored problem of matching a three-dimensional gas to a two-dimensional one at the source–channel and channel–drain junctions [34], must be investigated. As usual, we fear that we might find the macroscopic behavior of the device only moderately affected by the new physics, while gate currents may feel the effect strongly.

3. From Si to SiO_2

Let us leave the problem of transport of Si behind and move to the question of what electrons do once in the SiO_2. We skip the problem of how carriers are injected, not because of lack of questions. (What is the role of the image-force barrier lowering? Does tunneling conserve parallel momentum? If not, as it seems, what is the role of surface excitations in breaking the symmetry? What is the role of the amorphous structure of SiO_2 and how to account for it? What is the role of point-like oxide charges in the injection process? What is the role of scattering in the image-force rounded barrier for electrons entering at high energy?...), but because of a lack of answers. The various phenomenological models used to describe the "barrier-height" at the Si–SiO_2 interface to fit observed hot-carrier effects lump these problems (as well as a plethora of unsolved Si-transport problems) into a few fitting parameters and indicate the precarious state of our understanding of this issue.

For once, we are confident both experimentally and theoretically that we understand what happens to electrons in SiO_2. We have published results of our work already many times [35–38]. Here a brief review is necessary in order to introduce the results of the next section.

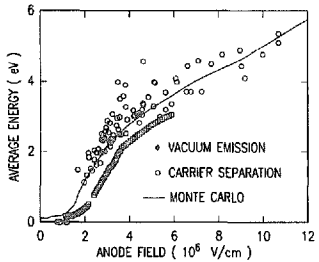

Fig. 5. Average energy versus oxide field of electrons in the SiO_2 conduction band [35–38]. Note the "heating threshold" at about 1.5 to 2 MV/cm.

Electrons in SiO_2 suffer two major types of phonon collisions: small-angle polar collisions with longitudinal-optical (LO) phonons of large energy (63 and 153 meV) and large-angle (particularly in umklapp processes) nonpolar collisions with acoustic, optical, and mixed zone-edge modes at large phonon wavevectors. The former collisions dominate the low-energy and/or low field transport. The latter dominate at electron energies above about 2 eV, thus at high fields. In fig. 5, we show the experimental and simulated average energy versus electric field characteristics. At low fields, electron transport is dominated by the polar collisions. At a critical field of about 1.5 to 2 MV/cm (the *heating threshold*), polar scattering cannot keep electrons at low energy any more. This is the well-known polar *runaway* accounting for the breakdown processes in alkali-halides. At higher fields, the task of preventing the carriers from reaching band-gap energies is picked up by nonpolar scattering. Transport now is stabilized at significantly larger energies (2 eV and up). Because of the large-angle nature of nonpolar collisions, transport is now very dispersive, unlike the streaming-like behavior occurring in the low-field, polar-dominated regime.

One major point should be kept in mind: since it takes the electrons 3 to 4 nm to "thermalize" (i.e., to reach the steady-state situation at a given oxide field), thick oxides will not feel the effect of how and with which energy-distribution electrons are injected. Not so for thin oxides (less than 6 nm): in these thin films, most of the electrons will not be thermalized by the time they reach the gate-SiO_2 interface and remnants of the energy distribution of the injected carriers can be seen in the distribution of carriers entering the gate. As an example, the Fermi–Dirac distribution of carriers in Si accumulation layers has been seen during tunneling and transport in 6 nm thin oxides [38].

4. SiO_2 degradation

Finally, what sort of damage do the electrons generate in the SiO_2? Even ignoring impurity-related effects (such as trapping in process dependent centers), this question has been left unanswered for almost two decades [39]. We shall attempt neither a review on the subject nor any discussion of the microscopic nature of those defects which appear electrically as positive charges, slow states, fast interface traps, or newly created electron traps [40]. The main conclusive results we shall present is that *hot electrons reaching the anode–SiO_2 interface with kinetic energy in excess of about 2 to 3 eV are responsible for the SiO_2 degradation* [41,42]. The idea that hot electrons are indeed responsible for all types of oxide damage is quite old and very intuitive [39]. However, not until recently has this simple fact been demonstrated [42].

Gate oxides of poly-Si-gate n-channel MOSFET's have been stressed with two different techniques (injection of hot electrons optically generated in the substrate –HE – and Fowler–Nordheim tunnel injection – FN) over a wide range of fields (0.3 to 12 MV/cm) and oxide thickness (3.7 to 96 nm). By monitoring various voltage shifts (flat-band and threshold voltages) and using the photo-I–V charge-sensing technique, we have found that oxide degradation appears in the form of newly generated electron traps, positive charge, and interface traps only above some critical threshold: In thick oxides (≥ 10 nm), independent of the injection technique, degradation occurs only for fields above about 1.5 MV/cm, while in thin oxides (≤ 5 nm), using the FN injection, only for applied voltages larger than about 5 V. Considering the results of the previous section and shown in fig. 5, a direct correlation is seen between the onset of electron heating and degradation. Fig. 6 shows the generation rate of bulk electron traps versus average electron energy in the

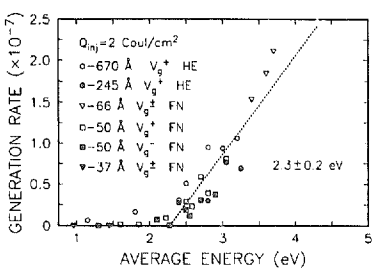

Fig. 6. Generation rate for the creation of bulk electron traps by hot carriers as a function of average energy of the electrons [41,42]. Note the "degradation threshold" at about 2 to 3 eV.

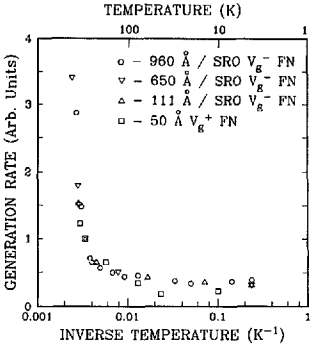

Fig. 7. Temperature dependence of the trap-generation rate of fig. 6 in Arrhenius form. Note the "freeze-out" of the process below about 150 K [42].

oxide. (Similar results are obtained for the generation rate of interface traps and positive charge, although these appear with smaller concentrations.) The degradation process proceeds slowly at lower temperatures and almost disappears below about 150 K [42], as shown in fig. 7. Hydrogen incorporation in the SiO_2 enhances the process [43], but the qualitative trends remain unchanged.

Recently these results have been reproduced by other groups [44,45] and lead to a clear picture of the process: looking at fig. 8 as a guideline, electrons entering the oxide do not produce any damage at the Si–SiO_2 interface *in a direct way*. This can be inferred from the insensitivity of the degradation mechanism to the injection technique used: when HE injection is used, many carriers are allowed to enter the oxide with a significant kinetic energy. Not so when the FN injection is used. Still, no significant differences are seen in the degradation. Electrons are then transported across the insulator. The fact that the field at the anode–SiO_2 interface controls both the generation rate [39] and the amount of electron heating [35–38] indicates that *it is the energy lost by the electrons at the anode–SiO_2 interface which triggers the generation of the defects in the oxide*. Some species generated at that interface must then travel to the Si–SiO_2 interface across the oxide in order to generate the well-known interfacial defects. The combination of these processes (migration + interface reactions) is strongly temperature dependent.

This clear (finally!) picture is open to various speculations as soon as we look for a possible model: For example, one may look at the 2 3 eV threshold

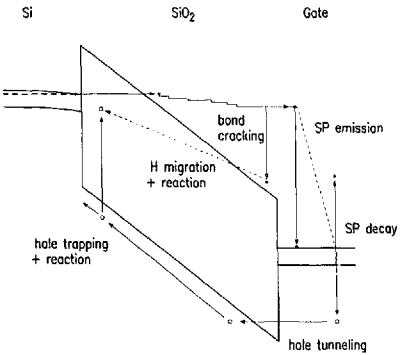

Fig. 8. Schematic diagram illustrating the process of SiO_2 degradation caused by hot carriers. Independent of the injection conditions (FN injection is illustrated for clarity only), carriers reaching the SiO_2-gate interface with energy in excess of 2 to 3 eV generate damage. Two speculative models (liberation of H after a bond-cracking process or hole-emission from the anode) are also shown.

as the critical energy needed to crack weak bonds (e.g., Si–H bonds) in the oxide close to the gate–SiO_2 interface. The species released by this cracking process (H, in our example) then migrates via a temperature activated process to the opposite interface where more defects are generated. Alternatively, the 2–3 eV threshold may be viewed as a 5–6 eV threshold for electrons entering the poly-Si gate. Here, the loss of this energy may result in the back-injection into the SiO_2 of another species (e.g., holes generated by the decay of the surface-plasmons emitted by the hot carriers [46]). The temperature dependence may result from either hole-transport or from the thermal activation of further interfacial processes (chemical and/or purely electronic). The effect of hydrogen is obvious in the first model, while it may be needed in the context of interfacial chemistry by the second model. Since obvious disagreement exists even among two of the present authors (D.J.D. and M.V.F.) on which model should be preferred, we could safely conclude that more evidence is needed to make an intelligent choice.

5. Conclusions

Bad news and good news emerge from our discussion. On the bad side, we think we have justified the "halo" of pessimism surrounding the title of this

paper (as well as the name DAMOCLES given to our simulation program) addressing the problem of "predicting" gate currents. The answer to our title question will ultimately be a "yes", since as physicists we should not care how long this might take us. The real question is then: Can we understand it in time to help device engineers design the 16 or 64 Mbyte DRAM's? Here we can only hope. But it is clear that empirical approaches both technologically (such as the fluoridation of SiO_2 presented by Nissan-Cohen at this workshop [43]) and in the modeling arena (i.e. more parameter fitting) might do the job

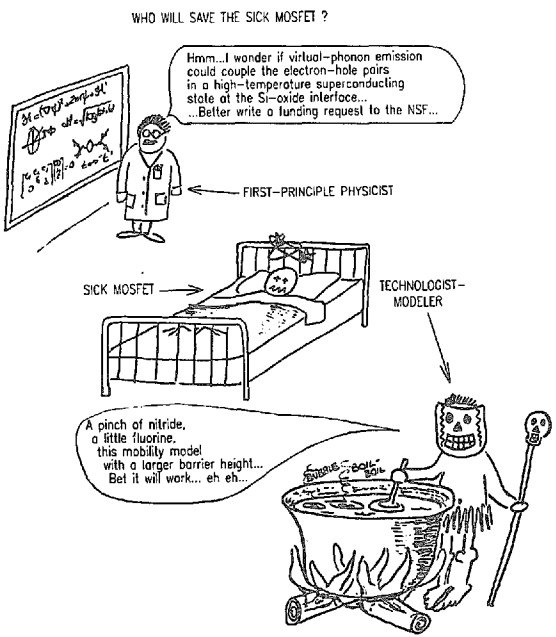

Fig. 9. Schematic diagram illustrating the two possible approaches we can take to tackle the "hot-electron problem". We may recall that, had we waited for "first-principle physicists" to understand the oxidation of Si, we would still be using vacuum tubes.

long before a "first-principle" approach, at the expense of understanding what is really happening in our devices. Thus, there is little doubt about witch (sic!) of the two characters of fig. 9 will cure the "sick MOSFET".

On the good side, the existence of a clear threshold for the degradation of SiO_2 brings hope for future devices working at reduced biases. If the voltage drop across the gate-oxide is kept below 2 V, even the existence of substrate and gate currents below the 1.2 and 3 V thresholds poses no danger to the oxide reliability. Moreover, this conclusion is reached by virtue of a "first-principle" understanding of electron transport in SiO_2 and we may regard this as a really remarkable success of our approach. Perhaps, the lesson we should learn is better explained by a closely related example: in the past, simple parameter-fitting has indeed claimed the ability to "explain" the so-called "intrinsic breakdown" of SiO_2 (see the references quoted in ref. [39]). But it has fallen far short in predicting oxide degradation and has been largerly abandoned. Empirical modeling approaches might find the same end not too far down the line, in what has been called in a picturesque way the "deep submicron" range, where a "first-principle" understanding is really required.

References

[1] Just to quote some early references on the subject:
P.E. Cottrel, R.R. Troutman and T.H. Ning, IEEE Trans. Electron Devices ED-26 (1979) 520;
B. Eitan and D. Frohman-Bentchkowsky, IEEE Trans. Electron Devices ED-28 (1981) 328.
[2] S.M. Sze, Physics of Semiconductor Devices, 2nd ed. (Wiley, New York, 1981) p. 50.
[3] See, for example, G. Baccarani and M.R. Wordeman, Solid-State Electron. 28 (1985) 407.
[4] W. Shockley, Solid-State Electron. 2 (1961) 35.
[5] S. Tam, P.-K. Ko and C. Hu, IEEE Trans. Electron Devices ED-31 (1984) 1116.
[6] K. Blotekjaer, IEEE Trans. Electron Devices ED-17 (1970) 38.
[7] R.K. Cook and J. Frey, IEEE Trans. Electron Devices ED-29 (1982) 970.
[8] W. Hänsch and M. Miura-Mattausch, J. Appl. Phys. 60 (1986) 950.
[9] E. Scholl and W. Quade, J. Phys. C 20 (1987) L861, and references therein.
[10] C. Moglestue, IEEE Trans. Computer-Aided-Design CAD-5 (1986) 326, and references therein.
[11] K. Throngnumchai, K. Asada and T. Sugano, IEEE Trans. Electron Devices ED-33 (1987) 1005.
[12] M. Tomizawa, K. Yokoyama and A. Yoshii, IEEE Trans. Computer-Aided Design CAD-7 (1988) 254.
[13] E. Sangiorgi, B. Riccò and F. Venturi, IEEE Trans. Computer-Aided Design CAD-7 (1988) 259.
[14] J.Y. Tang and K. Hess, J. Appl. Phys. 54 (1983) 5139.
[15] M.V. Fischetti and S.E. Laux, Phys. Rev. B 38 (1988) 9721.
[16] J. Zimmermann, R. Fauquembergue, M. Charef and E. Constant, Electron. Letters 16 (1980) 665.
[17] P.K. Basu, Solid State Commun. 27 (1978) 657.

[18] K. Hess, Solid-State Electron. 21 (1978) 123.
[19] S. Manzini and A. Modelli, Solid-State Electron. 31 (1988) 99.
[20] Y. Park, T. Tang and D.H. Navon, IEEE Trans. Electron Devices ED-30 (1983) 1110.
[21] See the discussion in ref. [15].
[22] M.V. Fischetti and S.E. Laux, in: Simulation of Semiconductor Devices and Processes, Eds. G. Baccarani and M. Rudan (Tecnoprint, Bologna, 1988) p. 349.
[23] S.E. Laux and M.V. Fischetti, IEEE Electron Device Letters EDL-9 (1988) 467.
[24] M.L. Cohen and T.K. Bergstresser, Phys. Rev. 141 (1966) 786.
[25] H. Shichijo and K. Hess, Phys. Rev. B 23 (1981) 4197.
[26] G.A. Sai-Halasz et al., IEEE Electron Device Letters EDL-8 (1987) 463.
[27] C. Jacoboni, R. Minder and G. Maini, J. Phys. Chem. Solids 36 (1975) 1129.
[28] S. Tam, F.-C. Hsu, R.S. Muller and P.K. Ko, IEEE Electron Device Letters EDL-4 (1983) 249.
[29] E. Sangiorgi, M.R. Pinto, F. Venturi and W. Fichtner, IEEE Electron Device Letters EDL-9 (1988) 13.
[30] A. Phillips and P.J. Price, Appl. Phys. Letters 30 (1977) 528.
[31] E.N. Economou, Phys. Rev. 182 (1969) 538.
[32] D. Pines, Rev. Mod. Phys. 28 (1956) 184.
[33] J.C. Slater, Phys Rev. 76 (1949) 1592.
[34] A.M. Kriman and P.P. Ruden, Phys. Rev. B 32 (1985) 8013.
[35] M.V. Fischetti, D.J. DiMaria, S.D. Brorson, T.N. Theis and J.R. Kirtley, Phys. Rev. B 31 (1985) 8124.
[36] M.V. Fischetti, D.J. DiMaria, L. Dori, J. Batey, E. Tierney and J. Stasiak, Phys. Rev. B 35 (1987) 4404.
[37] M.V. Fischetti and D.J. DiMaria, Solid-State Electron. 31 (1988) 629.
[38] D.J. DiMaria and M.V. Fischetti, J. Appl. Phys. 64 (1988) 4683.
[39] This issue was discussed at INFOS 85 by one of us. See the many references in M.V. Fischetti, in: Insulating Films on Semiconductors, Eds. J.J. Simonne and J. Buxo (North-Holland, Amsterdam, 1986) p. 181.
[40] The work by S.A. Lyon in this volume addresses some of these issues: S.A. Lyon, Appl. Surface Sci. 39 (1989) 552.
[41] D.J. DiMaria, Appl. Phys. Letters 51 (1987) 665.
[42] D.J. DiMaria and J.W. Stasiak, J. Appl. Phys. 65 (1989) 2342.
[43] Y. Nissan-Cohen, Appl. Surface Sci. 39 (1989) 511.
[44] C.C.H. Hsu, T. Nishida and C.T. Sah, J. Appl. Phys. 63 (1988) 5882.
[45] M.M. Heyns, D.K. Rao and R.F. De Keersmaecker, Appl. Surface Sci. 39 (1989) 327.
[46] M.V. Fischetti, Phys. Rev. B 31 (1985) 2099.