Semicond. Sci. Technol. **23** (2008) 035014 (13pp) doi:10.1088/0268-1242/23/3/035014

Kinetic investigation of electron–electron scattering in nanometer-scale metal-oxide-semiconductor field-effect transistors

D A Fixel[1] **and W N G Hitchon**[2]

[1] Electrical and Microsystem Modeling, Sandia National Laboratories, PO Box 5800, MS 0316, Albuquerque, NM 87185-0316, USA
[2] Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA

E-mail: dafixel@sandia.gov

Received 10 October 2007, in final form 7 January 2008
Published 7 February 2008
Online at stacks.iop.org/SST/23/035014

Abstract
The effects of electron–electron scattering on the electron energy distribution, as well as substrate and gate currents in short channel MOSFETs (metal-oxide-semiconductor field-effect transistors) are explored using the convective scheme, or CS, a method of characteristics. Effects of electron–electron scattering are explored for a MOSFET with uniform doping in the channel as well as for an asymmetric device structure, a focused-ion-beam (FIBMOS) transistor, for both 70 nm and 250 nm channel length devices. Effects of electron–electron scattering on a standard 35 nm channel length MOSFET are also included. The high substrate doping that is required for such short channel length devices leads to large electric fields. The purpose of the FIB implant is to improve hot-carrier reliability by reducing the electric field in the channel. Electron–electron scattering increases the amount of electrons in the tail, despite the fact that the applied potential is significantly below the threshold for injection of electrons into the gate oxide. The ratio of gate-to-substrate current, I_g/I_{sub}, is investigated as an indicator of the level of degradation. At such short channel lengths, there are degrading and non-degrading components of gate and substrate current. The non-degrading components of gate and substrate current correlate strongly, so that the ratio of I_g/I_{sub} is an efficient indicator of device degradation. The energy thresholds for impact ionization and for emission of electrons into the gate oxide are crucial in determining the ratio of these currents. The substrate and gate currents obtained indicate that hot-carrier effects continue to be an issue for device performance, even for nanometer-scale devices. The density of electrons is higher at very short channel lengths due to the need to have shallow junctions and leads to a greater amount of Coulomb collisions. Increased Coulomb collisions may lead to strongly reduced lifetimes in nanometer-scale devices.

1. Introduction

With the advent of nanometer-scale electronics, hundreds of millions of transistors reside on a chip area of roughly a few square centimetres. As the feature sizes of these devices approach 10 nanometers, circuit densities are projected to reach the gigascale [1]. In a short channel metal-oxide-semiconductor field-effect transistor (MOSFET), high electric fields develop at the drain end of the device, which can lead to hot-carrier effects and device degradation. Originally, it had been thought that for drain biases below approximately 2.7 V, electrons would not be able to gain enough energy to cause interface damage or device degradation, no matter how short the channel length is [2]. Tam *et al* directly measured

 © 2008 IOP Publishing Ltd Printed in the UK

gate currents at a drain bias of 2.35 V [3]. While one would have expected that hot-carrier effects would be diminished as the power-supply voltages are scaled, the electron energy distribution is affected at energy values as high as 3 eV, even with a power supply voltage as low as 1.0 volt [4]. Ang *et al* [5] noted that the substrate current at low drain bias (the substrate current is one measure of hot-carrier effects) arises from a large population of electrons with energies between 1.2 eV and qV_d, where V_d is the potential applied to the drain.

In this work a kinetic transport model is implemented that is relevant in cases where the mean free path is comparable to the device dimensions. The method is referred to as the convective scheme, or CS, and has previously been used extensively for plasma simulations [6–9]. The energy distribution of electrons is explored in this work using three velocity dimensions and two space dimensions. In order to find the electron distribution, it is necessary to solve the Boltzmann transport equation, which is given by [10]

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla f + \frac{d\mathbf{k}}{\mathrm{d}t}\cdot\frac{\partial f}{\partial \mathbf{k}} = \left.\frac{\partial f}{\partial t}\right|_{sc} \tag{1}$$

where the second term on the left-hand side of the equation represents the rate of change of the distribution due to the particles' velocity and the third term on the left-hand side represents the rate of change of the distribution due to the particles' experiencing external forces. The term on the right-hand side of the equation is the collision term, i.e., the effect of collisions in altering the distribution function [10]. The Boltzmann transport equation is valid in the semiclassical limit. We want to note that as the device dimensions approach the deBroglie wavelength, the validity of the semiclassical approach begins to come into question. At room temperature, the deBroglie wavelength, λ_{DB} is equal to

$$\lambda_{DB} = \left(\frac{2\pi\hbar^2}{mk_bT}\right)^{1/2} = 8.4\ \text{nm}. \tag{2}$$

The 35 nm channel length device thus approaches this limit and the results should be viewed with the appropriate caution, although Lundstrom and Ren employed a 10 nm MOSFET transistor in their work, citing the observation that MOSFETs operate essentially classically down to roughly 10 nm channel length [11].

The particular implementation of the CS used in this work finds the scattering rate in a phase space cell from a set of probabilities that are found from the convective scheme solution of the Boltzmann equation. This basically consists of finding a Green's function for a differential equation by considering equations with a delta-function source. The probabilities used have to do with the manner in which particles move from one cell to another on the mesh and with scattering events of the particles. The method is referred to as the transition probability matrix (TPM) method, and this method of obtaining the collision rate of particles has been implemented in previous works [12, 13]. More details on this approach are given in the appendix.

Childs and Leung [14] established a technique for solving the one-dimensional spatially dependent BTE with electron–electron interactions included in the scattering model. This was accomplished by solving the BTE over a potential profile typical of that found in the channel of a MOSFET. A comparison was made between the distribution functions obtained when electron–electron interactions are included and excluded from the scattering model. They incorporated a few approximations in order to make the problem more tractable. The first approximation considers the electron–electron interaction to be a localized event. The second assumes that the distribution of the interacting electrons is isotropic.

The gate current is often used as a measure of hot-carrier damage. Gate currents were observed in MOSFETs with drain biases substantially below what is normally required to cause hot-carrier effects. Childs and Leung concluded that in the absence of other energy sources, electron–electron interaction can account for hot-carrier effects with low supply voltages.

Many authors recognize the importance of electron–electron scattering in semiconductor device simulation as it relates to the high-energy tail. Dyke *et al* [15] demonstrated a hybrid Monte Carlo/iterative technique for solving the Boltzmann transport equation. Their results indicate that the method can be used to explain the form of the hot-carrier distribution in MOSFETs operated at low drain bias. Their results indicate that in the absence of electron–electron interactions, the hot-carrier distribution at energies greater than those available from the electric field falls off rapidly. On the other hand, when electron–electron interactions are included, the hot-carrier distribution is enhanced and increases as the drain voltage.

McMahon *et al* [16] also noted the existence of device damage due to hot carriers when the supply voltage has been scaled below the threshold energy for damage. They cite the importance of electron–electron scattering in addition to the high-energy tail, but note a couple of other reasons for such damage. The activation energy may not be a 'hard' barrier; in other words, there may be a distribution of activation energies. Also, the barrier could potentially be surmounted by multiple carrier processes acting together [16, 17].

Hess *et al* [17] noted that if the electron energy distribution decreases exponentially above the operating voltage V, then according to Boltzmann's law, there should be large differences in the amount of degradation as the operating voltage is scaled. A reduction in power supply from 3.3 V to 2.5 V, following this same exponential variation, should lead to a factor of 10^{20} reduction in degradation. This does not happen experimentally, i.e., the reduction in degradation does not scale with the reduction in voltage. Some have suggested that carrier–carrier interactions could explain the degradation at low bias, but it is not clear that the level of carrier–carrier scattering is sufficient to induce the level of damage that is seen. An example from the scattering rates involved in emission for electron–electron scattering should illustrate this point. A scattering rate that gives an energy exchange of 1 eV or greater is roughly $10^7\ \text{s}^{-1}$. The transit time of a high energy electron will typically be less than 10^{-12} s. Hence, less than 1 in 10^5 electrons would reach the energies of a 3.6 eV threshold with a power supply of 2.5 V. They note that the Si–H desorption energy can be significantly reduced for Si–H bonds

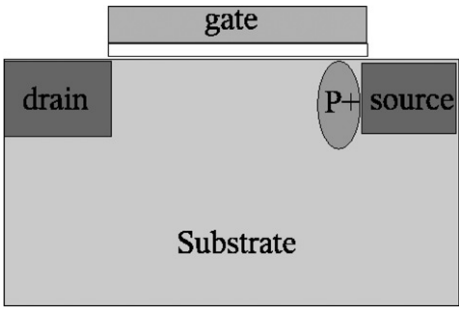

Figure 1. The figure shows a MOSFET with a focused-ion-beam (FIB) implant next to the source region.

at the interface because hydrogen can desorb by entering the silicon first. Also, studies of Si–H desorption with scanning tunnelling microscopy (STM) have shown that lower-energy carriers (less than 3.0 eV) can break bonds. The point is that carrier–carrier scattering does in fact play a role in enhancing the high-energy tail but the actual threshold for damage may be somewhat lower than previously expected [17].

In previous work, the importance of electron–electron, or Coulomb, collisions was recognized [18] as they affect the energy transfer between 'cold' and 'hot' electron components, which are found in the negative glow of dc discharges and in the 'bulk' plasma of rf discharges. A technique often used in gaseous electronics was employed wherein the electron distribution function can be approximated by the sum of two or more Maxwellians and whereby a formula from Longmire [19] gives the rate of energy transfer between two Maxwellians and is used to produce the effect of electron–electron scattering. It is evident that whether one is referring to the fluid state or the solid state of matter that Coulomb collisions play a significant role in calculating the high-energy tail of the electron energy distribution.

A number of asymmetric channel-engineered MOSFET structures have been developed in order to satisfy two competing requirements: the reduction of short-channel effects and an improvement in hot-carrier reliability. One very promising structure that improves upon both is the structure made with the focused-ion-beam (FIB) implantation. The FIB region is a p+ region that is located near the source end of the channel [20]. We explore here the effects of electron–electron scattering on a standard MOSFET and on a MOSFET including a FIB implant. Figure 1 shows a MOSFET with a FIB implant.

In this work, the ratio of gate-to-substrate current, I_g/I_{sub}, is investigated as an indicator of device degradation. For ultra-small channel lengths, there are degrading and non-degrading components in the gate and substrate currents. The non-degrading components have a strong correlation, so that the ratio I_g/I_{sub} is an accurate indicator of the device degradation. In the conclusion, we discuss reasons for the apparent short device lifetimes found. Section 2 reviews the theory of electron–electron scattering as it is implemented in this work. Section 3 describes plasmon scattering. Then section 4 discusses the calculation of gate and substrate currents. Section 5 goes into the simulation results and section 6 gives the conclusions. Following the conclusions, the appendix explains the kinetic simulation technique used in this work.

2. Theory of electron–electron scattering

In this work, the effects of electron–electron scattering on the high-energy tail electrons are investigated. Scattering mechanisms used in the simulations include acoustic phonon scattering, optical deformation potential scattering, ionized impurity scattering, impact ionization, and electron–electron scattering. This section discusses some of the relevant theory for electron–electron scattering and the details of implementation based on Ferry [21, 22], who followed the work of Takenaka *et al* [23]. They had implemented inter-carrier scattering to calculate the distribution function for GaAs. In this work we employ a single-particle electron–electron scattering approach which uses energy-dependent frequencies as derived by Ferry [22].

Since the literature indicates that electron–electron scattering can account for hot-carrier effects at low supply voltages, it has been included in the simulations. Although the version of the code which implements the electron–electron scattering takes longer to run than the version without it, its inclusion is necessary in order to explore the effects on the tail, especially at low supply voltages.

The difficulty in dealing with the electron–electron interaction arises from the nonlinear behavior of the interaction potential and the long range of this Coulomb potential. In devices of interest today, the Debye length is smaller than the interatomic distance [21]. Goodnick and Lugli performed a study of non-equilibrium transport in a quantum well in which they incorporated electron–electron scattering explicitly. They monitored the energy exchanged during the simulation and found that it was generally quite small [24].

The method employed here uses the Fermi golden rule approach in order to find the scattering rate. The Coulomb scattering potential is treated in the static screening approximation, which means that the scattering matrix element is

$$M(q) = \frac{e^2}{\epsilon_s\left(q^2 + q_D^2\right)} \tag{3}$$

where ϵ_s is the static dielectric function [21, 22].

The expression for the electron–electron scattering rate is given by

$$\Gamma(\mathbf{k}) = \frac{m^* e^4}{8\pi\hbar^3 \epsilon_s^2 q_D^2} \sum_{\mathbf{k_2}} f(\mathbf{k_2}) \left(\frac{|\mathbf{k_2} - \mathbf{k}|}{(|\mathbf{k_2} - \mathbf{k}|)^2 + q_D^2} \right) \tag{4}$$

where $\mathbf{k}$ and $\mathbf{k_2}$ are the momentum wave vectors of the initial particle and the particle with which it interacts, respectively. The wave vectors $\mathbf{k}'$ and $\mathbf{k}_2'$ represent the corresponding quantities after the scattering event.

This form of the scattering rate expression does not distinguish between emission and absorption of energy by the incident electron. By reformulating the energy-conserving delta function in the original integration, the integration over $\mathbf{k}'$ can be changed to an integration over $\mathbf{q} = \mathbf{k} - \mathbf{k}'$. One can

then arrive at expressions that explicitly treat the energy loss or gain of the incident electron during the collision.

This results in an emission rate of

$$\Gamma_{ee}^{em}(\mathbf{k}) = \frac{nm^*e^4}{4\pi\epsilon_s^2\hbar^3 k}\left(\frac{m^*}{2\pi k_B T_e}\right)^{1/2} \times \int_0^{E(k)/\hbar} d\omega \int_{q-}^{q+} \frac{dq}{\left(q^2+q_D^2\right)^2} \times \exp\left[-\frac{\hbar^2}{8m^*k_B T_e}\left(q-\frac{2m^*\omega}{\hbar q}\right)^2\right] \quad (5)$$

where the limits of integration on q are

$$k - \sqrt{k^2 - \frac{2m^*\omega}{\hbar}} = q_- < q < q_+ = k + \sqrt{k^2 - \frac{2m^*\omega}{\hbar}} \quad (6)$$

and an absorption rate of

$$\Gamma_{ee}^{ab}(\mathbf{k}) = \frac{nm^*e^4}{4\pi\epsilon_s^2\hbar^3 k}\left(\frac{m^*}{2\pi k_B T_e}\right)^{1/2} \int_0^{\infty} d\omega \int_{q-}^{q+} \frac{dq}{\left(q^2+q_D^2\right)^2} \times \exp\left[-\frac{\hbar^2}{8m^*k_B T_e}\left(q+\frac{2m^*\omega}{\hbar q}\right)^2\right] \quad (7)$$

with limits of integration on q of

$$\sqrt{k^2 + \frac{2m^*\omega}{\hbar}} - k = q_- < q < q_+ = k + \sqrt{k^2 + \frac{2m^*\omega}{\hbar}}. \quad (8)$$

where $\hbar\omega$ represents the energy gain or loss of the incident electron and m^* is the appropriate effective mass. For the portion of the expression that precedes the first integral sign, the density-of-state effective mass ($1.084m_0$) is used, where m_0 is the electron rest mass, given by 9.1095×10^{-31} kg. For the portion of the electron–electron scattering rate expression following the first integral sign, which involves a change in energy, the conductivity effective mass is used ($0.26m_0$). n is the electron density, T_e is the electron temperature and q_D is the Debye screening wave vector, which is given by

$$q_D = \sqrt{\frac{ne^2}{\epsilon k_B T_e}}. \quad (9)$$

The value of electron temperature used in the expression for the electron–electron scattering rate as well as for the Debye screening wave vector used in this same expression corresponds to the value given in Ferry [22], as this is the value ($T_e = 2500$ K) found in modern MOSFETs by characteristic light emission.

In the expression for the scattering rates for both emission and absorption, one of the integrations is performed over ω. This means that for carriers at a given energy, they will go to various values of energy by losing or gaining an energy of $\hbar\omega$ during the collision. In the implementation of electron–electron scattering in the code, thirty-two values of ω are used for each value of energy. This was to ensure that the scattering would not produce results that had electrons clumping in certain values of energy within the distribution. In the expressions given above for electron–electron scattering, if the outer integral is not performed, we have the scattering rate at energy E for a change in energy of $\hbar\omega$.

The convective scheme tracks particles by following cells. Each particle in a given cell is at the same velocity. The cells are triangles and quadrilaterals that move in the plane of the simulation region. The corners of these cells get moved according to the equations of motion. The cells originate from fixed cells on the mesh and are launched as sheets, or sets of contiguous cells, at the beginning of an iteration. In this case, the iteration lasts for 25 timesteps, after which the result of the previous iteration is used as an input for the next iteration. The reason for using an iteration of 25 timesteps is that the iteration needs to last for 5 collision times, and 5 timesteps represent one collision time.

Electron–electron scattering depends upon the density of electrons in the spatial cell, and so the scattering frequency for the cell was made to depend on the electron density in the given cell. However, the scattering frequency used to determine the timestep for each sheet was maximum for that sheet and so it was calculated based on the maximum concentration of electrons on the mesh. The sheet samples the entire mesh, hence the need to check the density on the entire spatial mesh. The scattering frequency for each individual cell did, however, depend on the actual concentration in the given spatial cell.

At the end of a timestep, each cell is in a position of overlap with one or more cells of the original mesh. Particles get mapped back to the fixed mesh according to the fractional portion of overlap with the underlying cells as well as the scattering frequency. The product of the scattering frequency and the timestep for each value of velocity on the mesh are equal to 0.2. The timestep for a given sheet in the simulation was determined by $\nu(E)\,dt(E) = 0.2$, where $\nu(E)$ is the scattering frequency. This gives a different value for the timestep for each value of velocity. This is different than some previous implementations of the convective scheme, in which the same timestep was used for all of the sheets.

3. Plasmon scattering

The previous section was concerned with the screened Coulomb interaction for single-particle scattering. In this section, we discuss the long-range part of the Coulomb interaction, which is responsible for scattering by collective oscillations of the electron gas, or plasmons. The summation over $\mathbf{q} = \mathbf{k} - \mathbf{k}'$ that appears in a Fourier transform of potential can be split into a short-range part, for which $q > q_c$, and a long-range part, for which $q < q_c$, where q_c is a cut-off wave vector defining the boundary between these two types of scattering [21].

The scattering rate for plasmon scattering becomes

$$\Gamma_{\text{e-pl}}(\mathbf{k}) = \frac{m^*e^2\omega_p}{4\pi\epsilon(0)\hbar^2 k}\left[N_q \ln\frac{q_D/k}{\sqrt{1+\hbar\omega_p/E_k}-1} + (N_q+1)\ln\frac{q_D/k}{1-\sqrt{1-\hbar\omega_p/E_k}}\right] \quad (10)$$

where m^* is the electron effective mass, ω_p is the plasmon frequency, q_D is the Debye screening wave vector and N_q is a function of the carrier temperature, given by

$$N_q = \left[\exp\left(\frac{\hbar\omega_p}{k_B T_e} - 1\right) - 1\right]^{-1} \quad (11)$$

In this work, plasmon scattering is included in some of the runs in order to take into account the long-range component of the electron–electron scattering. In the following section, the method of calculating substrate and gate currents is discussed.

4. Substrate and gate currents

In this section, the calculation of gate and substrate currents is presented. It begins with a discussion of the applicability of the power-law model to transistor devices for device lifetime prediction, proceeds with a discussion of the manifestation of substrate currents that results in a parasitic bipolar transistor and next talks about a recently derived empirical model for calculating device lifetime based on the ratio of gate-to-substrate current. Following this, it explains the calculation of gate and substrate currents from the electron energy distribution, including an expression used for the oxide injection probability and another for the oxide barrier height.

Takeda and Suzuki [25] developed an empirical model for device degradation for NMOS transistors with channel lengths ranging from 0.35–2 μm. In Takeda's empirical model [25], experimental data were fit to a power law of the form

$$\Delta V_{\text{th}}(\text{or } \Delta G_m/G_{m0}) = At^n. \tag{12}$$

The conventional power-law empirical model is valid only for relatively long channel length devices and/or low stress conditions. Such models underestimate the lifetime for deep-submicron MOS transistors. Cui *et al* [26] derived a relatively simple empirical expression which correlates the MOS lifetime with the ratio of gate-to-substrate current. This is an effective alternative to the power-law models for the lifetime prediction of modern devices.

In this work, gate and substrate currents are obtained from the electron energy distribution. The substrate current, I_{sub}, can be used as a predictor for device lifetime, since the hot-electron effects are driven by the channel electric field, which is maximum at the drain end of the channel. The substrate current, if not controlled, can lead to electron injection into the substrate and potential fluctuations in the substrate. These in turn induce snap-back breakdown and CMOS latchup which is an effect whereby a parasitic p-n-p-n device is created. This parasitic structure with a bipolar action consists of the p+-region of the PMOS device, the n-type substrate, the p-well of the NMOS device and the n+-region of the NMOS device. The point is that a low impedance path is created from the power supply to ground, which leads to failure within the device. [27]

In the longer channel devices, substrate current is an accurate representation of the degradation in the device. In a deep-submicron MOSFET, however, the base width of the parasitic n+-p-n+ transistor (formed by source, substrate and drain) is quite narrow. This means that the base current from the parasitic transistor becomes a significant part of the substrate current. In the longer channel device, all of the substrate current is due to impact ionization so that all of the substrate current contributes to device degradation.

The gate current in deep-submicron MOSFETs also has a component that does not contribute to device degradation. In these devices, thin gate oxides are used and a significant portion of the gate current is due to direct tunnelling. Only a portion of the gate current arises from trapping, detrapping and interface trap generation. So, both the gate and substrate currents have what can be referred to as 'non-degrading' current components. There is no known experimental procedure by which the non-degrading currents can be separated from the total current [26].

It was found that the non-degrading gate current and the non-degrading substrate current correlate strongly. Hence, Cui *et al* [26] developed an expression for the device lifetime using the ratio of gate current to substrate current, given by

$$\tau = \tau_0 \exp(\alpha I_g/I_{\text{sub}}) \tag{13}$$

where τ_0 and α are the empirically derived constants. The ratio of I_g/I_{sub} is a measure of the significance of the degrading component. The device lifetime improves with an increase of this ratio. It has been shown experimentally that the substrate current is six to nine orders of magnitude higher than the gate current [28].

In the kinetic simulation, we integrate across the distribution in the channel of the MOSFET to calculate the substrate and gate currents. The expressions used to obtain the substrate and gate currents from the electron energy distribution are given by

$$I_{\text{sub}} = q \int_0^L \int_{E_{\text{th}_{ii}}}^{\infty} f_\sigma(E, x) W_{ii}(E)\, \mathrm{d}E\, \mathrm{d}x \tag{14}$$

and

$$I_{\text{gate}} = q \int_0^L \int_{E_{\text{th}_{ox}}}^{\infty} v(E) f_\sigma(E, x) P_{inj} P_2\, \mathrm{d}E\, \mathrm{d}x \tag{15}$$

where q is the magnitude of the electron charge, x is the x-coordinate along the channel, L is the channel length, $E_{\text{th}_{ii}}$ is the threshold energy for impact ionization, $E_{\text{th}_{ox}}$ is the threshold energy for injection of electrons into the oxide and $v(E)$ is the electron velocity, where only electrons with momentum directed toward the interface are considered, $f_\sigma(E, x)$ is the energy distribution integrated over the channel cross section, $W_{ii}(E)$ is the impact ionization scattering frequency, P_{inj} is the probability of injection into the oxide and P_2 is the probability that an electron does not scatter in the image potential well in the oxide. Impact ionization creates electron–hole pairs, and the electrons leave through the drain, while the holes diffuse from the drain into the substrate. We now turn to a description of the probability of emission into the oxide.

4.1. Probability of electron emission into oxide

Hot electrons can be emitted into the oxide by either tunnelling or by overcoming the Schottky-lowered barrier. The P_{inj} used in the simulations represents the probability of an electron tunnelling into the oxide. In this work, the transmission probability is found by using the WKB approximation as given by Huang *et al* [29]. The electrons in the channel encounter a trapezoidal barrier when their energies are lower than the barrier minimum, ϕ_{min}. They encounter a triangular barrier when the electron energy is higher than the barrier minimum

$\phi_{\min}$ but lower than the barrier maximum $\phi_{\max}$. The tunnelling probability across a trapezoidal barrier is given by

$$P_{inj} = \exp\left\{-\frac{4\pi}{h}(2m^*)\left[\frac{2}{3q|E_{ox}|}\right] \times [(\phi_{\max} - E_y)^{3/2} - (\phi_{\min} - E_y)^{3/2}]\right\}. \quad (16)$$

The tunnelling probability across a triangular barrier is given by

$$P_{inj} = \exp\left\{-\frac{4\pi}{h}(2m^*)\left[\frac{2}{3q\,|E_{ox}|}\right](\phi_{\max} - E_y)^{3/2}\right\} \quad (17)$$

where $m^* = 0.5m_0$ is the electron effective mass in the oxide layer, E_{ox} is the normal electric field in the oxide and E_y is the energy of an electron moving in the y-direction.

For electrons with energy above $\phi_{\max}$, thermionic emission occurs at the silicon-oxide interface, i.e., electrons are able to surmount the barrier into the oxide without tunnelling. Not all electrons that are injected into the oxide make it to the gate electrode to contribute to the gate current. The probability P_2, or the probability of not scattering in the oxide image-potential well [30] is given by

$$P_2 = \exp(-X_{ox}/\lambda_{ox}) \quad (18)$$

where X_{ox} is the oxide thickness and λ_{ox} is the mean free path in the oxide. Fischetti *et al* [31] explored ballistic transport in thin silicon dioxide films. Their results indicated an average value of 1 nm for the mean free path in the oxide. This agrees well with experiment [31], and so this is the value applied in the expression for P_2 in this work. Although P_2 does not depend on energy in the expression used in this work, the gate current depends on the energy of the electrons through P_{inj}.

The lucky-electron model has found widespread application in the calculation of substrate and gate currents. Due to the finite mass of an electron and the relaxation time effect, there is a finite time involved in the carriers reaching a steady-state condition. It has been found that models based on the local field cause substrate and gate currents to be overestimated by several orders of magnitude, although the trend is still captured. This phenomenon refers to a fluid code, or drift–diffusion simulation. The reason that this type of simulation overestimates the actual currents is that it assumes implicitly that carriers instantaneously reach a steady-state energy corresponding to the local field [32]. A kinetic simulation, such as the one used in this work, integrates the distribution along the characteristic curves. This allows the carriers to respond to the variation of the field in a manner that approximates reality much more closely. In some instances, in the lucky-electron model, the mean free path used is energy independent. The mean free path is clearly energy dependent, so the use of the correct energy-dependent mean free path in this work improves the accuracy of the model [33].

The distribution that is used to calculate the currents is found by the kinetic simulation, which employs energy-dependent scattering rates. In each spatial cell on the mesh, there exists a set of cells of various energies. The gate and substrate currents are calculated from the distributions obtained from the converged convective scheme simulations, which use energy-dependent scattering rates, and hence energy-dependent mean free paths for each scattering mechanism. The CS implemented in this work uses a uniform density of electrons in a cell, so that the distance from the silicon-oxide interface is determined by the centroid of the cell. The cells are irregularly shaped triangles and quadrilaterals which are 'launched' at the beginning of an iteration.

4.2. Oxide barrier height

In this section the oxide barrier height with corrections for Schottky-barrier lowering and tunnelling is discussed. At large gate voltages where the surface is strongly inverted, the average oxide field is determined by $E_{ox} = V_G/t_{ox}$, where t_{ox} is the oxide thickness. The oxide thickness for the 35 nm channel length device is 2.5 nm, and the oxide thickness for the 70 nm channel length device is 5 nm. The total barrier height including corrections for Schottky-barrier lowering was determined to be

$$qV = 3.1\ \text{eV} - \beta E_{ox}^{1/2} \quad (19)$$

where 3.1 eV is the interfacial barrier (Si–SiO_2 interface) and $\beta E_{ox}^{1/2}$ is the term for Schottky-barrier lowering. The value for β was experimentally determined to be $2.59 \times 10^{-4} q\,(\text{V cm})^{1/2}$ [34], where q is the magnitude of the charge on an electron. In the case of the 35 nm channel-length device, with a 2.5 nm oxide and 0.9 V applied on the gate, the barrier height corrected for Schottky-barrier lowering becomes 2.61 eV. In the case of the 70 nm channel-length device, with a 5.0 nm oxide and 1.5 V applied on the gate, the barrier height corrected for Schottky-barrier lowering becomes 2.651 eV. Tunnelling across the barrier is handled by implementing the WKB approximation as discussed in the previous section.

In this section, we have discussed the substrate and gate currents as well as a couple of quantities that are necessary in order to calculate the gate current, namely the probability of injection into the oxide and the barrier height for injection into the oxide. In the following section, we will see some results from the simulation with the CS.

5. Results

In this section we observe results from the simulation of 35 nm and 70 nm channel length MOSFETs. The 2D device simulation tool SGFramework [37] is run before the kinetic simulation to obtain the electric potential and electron density at each node on the mesh. This in turn is used in the initial guess for the CS simulation. We next look at simulation results from a 35 nm channel length transistor.

5.1. 35 nm channel length transistor

In this section, we observe results from the 35 nm MOSFET. In this work we have implemented an NMOS transistor (p-type substrate with n-type source and drain regions). The channel length is 35 nm and the device length, including source and drain regions, is 50 nm; two cases are considered. The first case has a substrate doping concentration of 8.0×10^{17} cm^{-3} and a

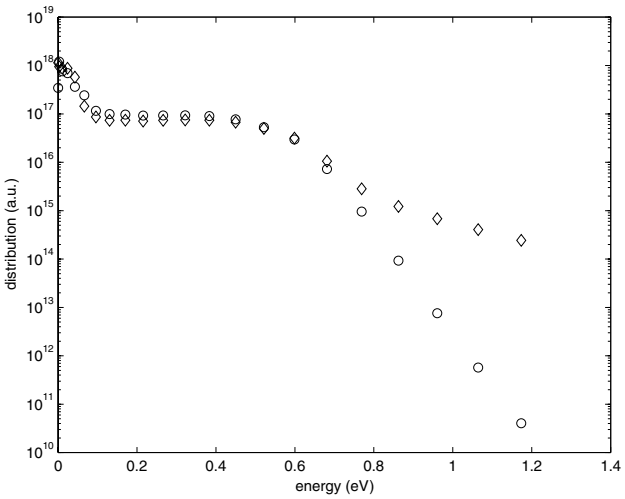

Figure 2. Distribution of electrons in a cell in channel for 35 nm channel length device with 0.9 V applied on gate and drain with and without electron–electron scattering included. The line indicated by (○) represents the distribution without electron–electron scattering and the line indicated by (◊) represents the distribution with electron–electron scattering. The cell is in the channel approximately at the midpoint between the source and the drain. The potential in the cell is 1.32 V, the substrate doping concentration is 8.0×10^{17} cm^{-3}, and the source and drain doping concentration is 8.0×10^{19} cm^{-3}.

source and drain doping concentration of 8.0×10^{19} cm^{-3}. The second case has a substrate doping concentration of 1.0×10^{18} cm^{-3} and a source and drain doping concentration of 2.0×10^{20} cm^{-3}. The minority carriers in the device are electrons, and during device operation the strong inversion in the channel of the device gives rise to a high density of electrons in the channel. A potential of 0.9 V was applied on the gate and drain of the MOSFET. The reason for applying 0.9 V on the gate and drain is that device reliability experiments are often performed under these conditions in order to 'stress' the transistor device and to observe hot-carrier effects [36]. The condition $V_g \simeq V_D$ is characteristically associated with maximum gate current, and the condition $V_g \simeq V_D/2$ is characteristically associated with maximum substrate current. Note, however, that the condition $V_g \simeq V_D$ is the worst case degradation bias condition for MOSFETs with channel lengths below 110 nm [38].

Figure 2 shows the electron energy distribution in the cell with 0.9 V applied on both the gate and the drain for the device with a substrate doping concentration of 8.0×10^{17} cm^{-3} and a source and drain doping concentration of 8.0×10^{19} cm^{-3}. Figure 3 shows the electron energy distribution in the cell with 0.9 V applied on the gate and drain for the 35 nm device with substrate doping of 1.0×10^{18} cm^{-3} and source/drain doping of 2.0×10^{20} cm^{-3}. The device shown in figure 3 has a higher doping concentration than the device shown in figure 2 which is representative of the values indicated in the ITRS (International Technology Roadmap for Semiconductors). Each plot exhibits a curve of the electron energy distribution excluding electron–electron scattering and a second curve that includes electron–electron scattering. It is apparent from the plots that there is one peak at low energy and a second peak at roughly the value of energy that the electrons can acquire as they traverse the channel. It can also be noted that the second peak occurs at a lower energy as the gate voltage is decreased. Beyond this value of energy the two curves are different in that the electron–electron scattering, or Coulomb scattering, has increased the distribution of electrons in the tail. As the channel length becomes shorter, the gate voltage has more of an effect on the transistor because of charge sharing [39]. This means that depletion charge underneath the gate is shared with depletion charge in the source/drain regions. The gate charge encroaches on the source/drain charge with a resulting triangular region of charge shared between the gate and source or drain region.

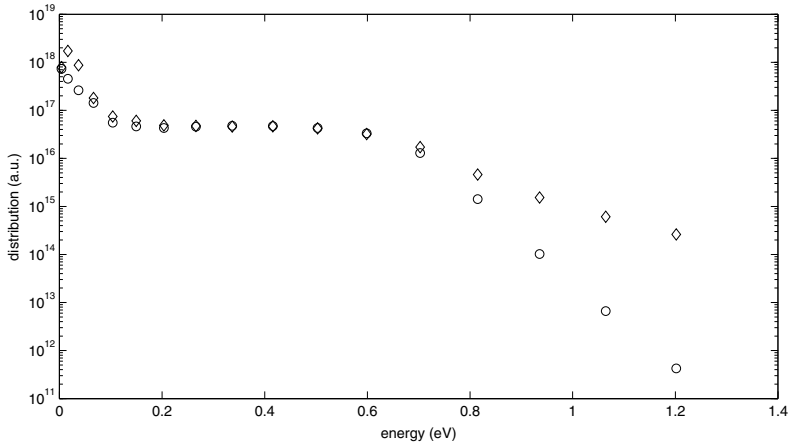

Figure 3. Distribution of electrons in cell in channel for 35 nm channel length device with 0.9 V applied on gate and drain with and without electron–electron scattering included. The line indicated by (○) represents the distribution without electron–electron scattering, and the line indicated by (◊) represents the distribution with electron–electron scattering. The cell is in the channel approximately at the midpoint between the source and the drain. The potential in the cell is 1.28 V, the substrate doping concentration is 1.0×10^{18} cm^{-3}, and the source and drain doping concentration is 2.0×10^{20} cm^{-3}.

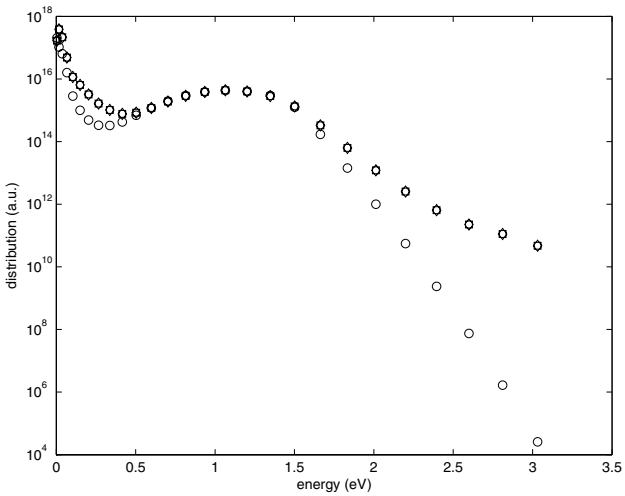

Figure 4. Distribution of electrons in cell at the drain end of the channel for a 70 nm channel length device with 1.5 V applied on gate and 1.5 V applied on the drain. The oxide thickness is 5 nm. The symbol (○) represents the distribution without electron–electron scattering, the symbol (◊) represents the distribution with electron–electron scattering, and the symbol (□) represents the distribution with both electron–electron scattering and plasmon scattering. The potential in the cell is 1.08 volt, the substrate doping concentration is 8.0×10^{17} cm^{-3}, and the source and drain doping concentration is 8.0×10^{19} cm^{-3}.

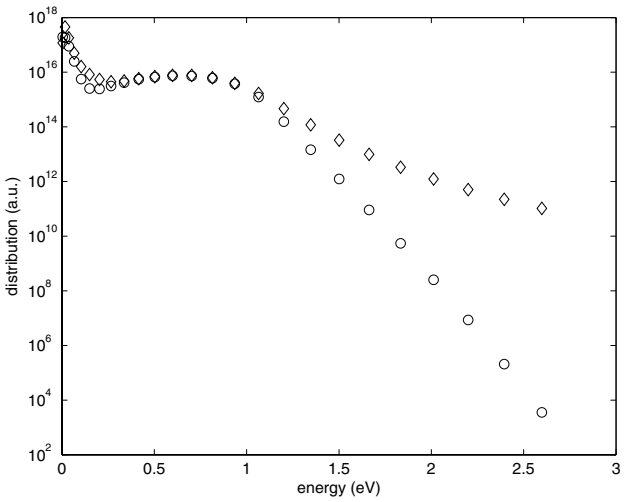

Figure 5. Distribution of electrons in the cell in the center of the channel for a 70 nm channel length device with 1.5 V applied on the gate and 1.0 V applied on the drain. The oxide thickness is 5 nm. The symbol (○) represents the distribution without electron–electron scattering and the symbol (◊) represents the distribution with electron–electron scattering included. The potential in the cell is 1.05 volt, the substrate doping concentration is 8.0×10^{17} cm^{-3}, and the source and drain doping concentration is 8.0×10^{19} cm^{-3}.

The effect of the Coulomb collisions is to increase the distribution in the tail. We do not attempt to apply the model for device lifetime prediction to the 35 nm device. For an oxide thickness less than 5 nm hot electrons are not the driving mechanism for gate current, but rather direct tunnelling occurs [40]. It would be necessary to solve the Schrodinger equation self-consistently with Poisson's equation in order to obtain the gate current for such thin oxides. In the following section, results from the 70 nm channel length transistor are presented.

5.2. 70 nm channel length transistor

In this section, results from the 70 nm channel length transistor are considered. The channel length is 70 nm and the device length, including source and drain regions, is 100 nm. Two cases are considered, both of which have an oxide thickness of 5 nm. The first case has a substrate doping concentration of 8.0×10^{17} cm^{-3} and a source and drain doping concentration of 8.0×10^{19} cm^{-3}. The second case has a substrate doping concentration of 8.0×10^{17} cm^{-3}, a source and drain doping concentration of 8.0×10^{19} cm^{-3} and a FIB implantation region which is 20 nm wide with a doping concentration of 1.0×10^{19} cm^{-3}. Figure 4 shows the case for 1.5 V applied to the gate and drain of the 70 nm device with a 5 nm oxide thickness, a substrate doping concentration of 8.0×10^{17} cm^{-3} and a source and drain doping concentration of 8.0×10^{19} cm^{-3}. In this case, plasmon scattering was also included in the simulation. As can be seen from the plot, there is no difference between the case that includes electron–electron scattering and the case in which both electron–electron scattering and plasmon scattering are included.

As the drain bias is increased, the second peak of the curve is moved to a higher value of energy. This occurs because the drain bias can contribute to the energy an electron acquires in traversing the channel. An increase in the gate voltage appears to increase the amount of electrons in the tail region. The effect of decreasing the drain voltage while leaving the gate voltage at 1.5 V moves the second peak to a lower value of energy while the electron–electron scattering still increases the distribution in the tail by roughly the same amount. This demonstrates the influence of the gate voltage on the gate and substrate currents. Note that when the drain voltage is higher than the gate voltage the configuration of the electric field is such that the carriers have a greater tendency to stay within the channel. Figure 5 shows the distribution of a cell in the channel for the same doping concentrations as figure 4. Figure 6 shows the distributions for the 70 nm MOSFET with and without the FIB implant. Figure 7 shows the gate and substrate current values as the gate voltage varies from 0.5 volt to 1.5 V for the 70 nm transistor without the FIB implant. Figure 8 shows the gate and substrate current values as the drain voltage is varied for the 70 nm device with a FIB implant. Channel hot-electron (CHE) injection is the mechanism primarily responsible for gate current that occurs from electrons surmounting the barrier height into the gate oxide. The escape of 'lucky electrons' from the channel results in significant degradation of the oxide and of the Si/SiO_2 interface. In a nanoscale device, tunnelling current also gives a significant contribution to the gate current. Due to avalanche multiplication at the drain (which contributes to the substrate current), the gate current is usually several orders of magnitude smaller than the substrate current. Gate current, which has a peak value at $V_g \simeq V_d$ characteristically results from channel hot-electron injection. Ning *et al* [34, 36]

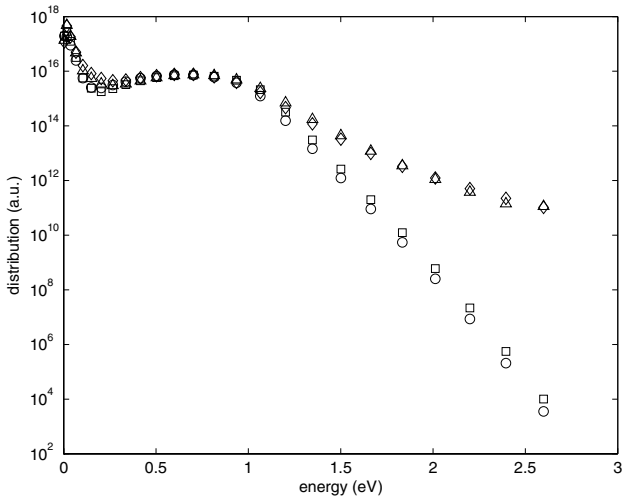

Figure 6. This figure compares the electron energy distributions for 70 nm MOSFETs with and without the FIB implant. 1.5 V is applied on the gate and 1.0 V is applied on the drain. The oxide thickness is 5 nm. The symbol (○) represents the distribution without electron–electron scattering and without the FIB implant. The symbol (◊) represents the distribution with electron–electron scattering and without the FIB implant. The symbol (□) represents the distribution without electron–electron scattering and with the FIB implant. The symbol (△) represents the distribution with electron–electron scattering and the FIB implant. Both devices have a substrate doping concentration of 8.0×10^{17} cm^{-3} and a source and drain doping concentration of 8.0×10^{19} cm^{-3}. The MOSFET with the FIB implant also has a 20 nm wide FIB implant with a doping concentration of 1.0×10^{19} cm^{-3}.

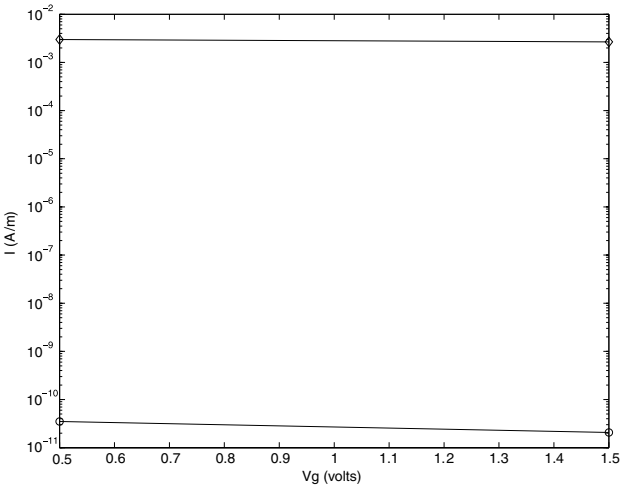

Figure 7. This figure shows the gate and substrate current values for the 70 nm MOSFET without FIB implant when electron–electron scattering is included in the simulation. The drain voltage in each case is 1.5 V and the gate voltage is varied from 0.5 V to 1.5 V. The symbol (○) represents the gate current and the symbol (◊) represents the substrate current.

reported that the conditions are optimum for CHE injection of 'lucky electrons' if an n-channel MOSFET operates under this biasing condition. The substrate currents are larger than the gate currents, since the threshold for impact ionization is lower than the threshold for injection into the oxide. Both

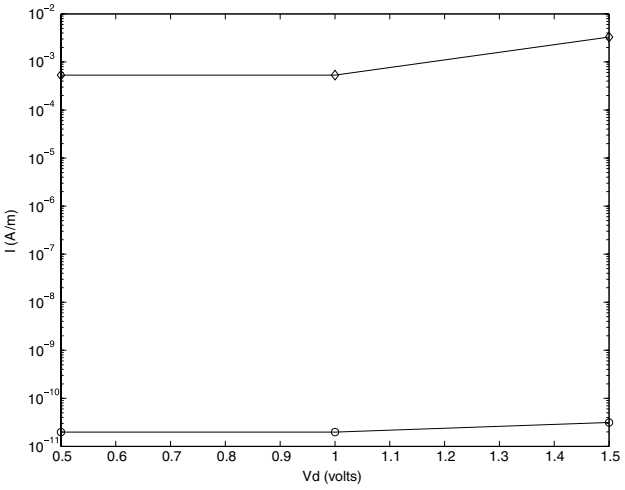

Figure 8. This figure shows the gate and substrate currents for the 70 nm MOSFET with a FIB implant. The gate voltage is 1.5 V in each case, and the drain voltage is varied from 0.5 V to 1.5 V. The symbol (○) represents the gate current, and the symbol (◊) represents the substrate current when electron–electron scattering is included in the simulation.

Table 1. Parameters for device lifetime model.

Channel length	τ_0	α
0.25 μm	1.862×10^3	1.494×10^6
0.18 μm	6.026×10^3	1.128×10^6
70 nm	1.60×10^5	5.188×10^5

currents are increased due to electron–electron scattering, as can be expected from the increase in the number of electrons in the high energy tail.

The values found for the lifetime prediction model of Cui *et al* [26] for the 0.25 μm device were $\tau_0 = 1.862 \times 10^3$ and $\alpha = 1.494 \times 10^6$. The values found for the 0.18 μm device were $\tau_0 = 6.026 \times 10^3$ and $\alpha = 1.128 \times 10^6$. Performing a linear interpolation on these values to determine values for a 70 nm channel length, we obtain values of $\tau_0 = 1.26 \times 10^4$ and $\alpha = 5.53 \times 10^5$. Performing a log($x$) interpolation, we obtain values of $\tau_0 = 1.60 \times 10^5$ and $\alpha = 5.188 \times 10^5$. Since the values were obtained by extrapolation, they are viewed with the appropriate caution. However, we still see the trend of a reduction in the lifetime as the ratio I_g/I_{sub} diminishes. Sometimes lifetime models are based on the time to breakdown from charge-pumping experiments, rather than the actual time that a device lasts in a circuit. Table 1 shows the parameters used in the lifetime model for the various channel lengths.

Applying the parameters obtained by Cui *et al* to the 70 nm MOSFET, we obtain for the MOSFET without the FIB implant a lifetime of 6.08×10^3 s (for a ratio of I_g/I_{sub} of 7.75×10^{-9}). For the MOSFET with the FIB implant, we obtain a lifetime of 6.09×10^3 s (for a ratio of I_g/I_{sub} of 9.51×10^{-9}). Using the values for the lifetime model by linear interpolation, a lifetime of 1.265×10^4 s is found for the MOSFET without the FIB implant and a lifetime of 1.267×10^4 s is found for the MOSFET with a FIB implant. By contrast, using the values obtained by log(x) interpolation, a lifetime of 1.606×10^5 s

is obtained for the MOSFET without the FIB implant, and a lifetime of 1.608×10^5 s is obtained for the device with a FIB implant. We note that the FIB implant causes no significant change in the lifetime when using the same substrate doping in both the MOSFET with and without the FIB implant. In a later section, results are presented for a device with and without a FIB implant where the threshold voltage is the same for both devices, which bears out the point of a decreased lifetime due to an increase in doping concentration more strongly. Again, the results indicate that the hot-carrier degradation does not subside simply by reducing channel length and biases on the device. The I_g/I_{sub} ratio is in the 10^{-8} range, which points toward increased degradation effects. We note that this ratio is comparable to published data for devices of the same approximate channel lengths [4].

It is beyond the scope of this paper to provide a first-principles analysis of the damage caused by the tail, or to justify the scaling of damage with the ratio of the currents. We shall thus avoid the temptation to speculate as to the exact mechanisms involved. We do note, however, that as determined by uniform-substrate hot-carrier injection experiments, probabilities for electron trapping, electron trap generation and interface trap generation are generally 10^{-7} or less [41]. We do observe that electron–electron collisions greatly enhance the tail of the distribution, and so they profoundly affect these currents. The reasons for the strong electron–electron collisions stem from the high carrier density, discussed elsewhere in this paper. We note that the calculated gate and substrate currents do depend strongly on the threshold energies. The threshold energy for impact ionization used in this work is 1.65 eV, which represents 1.5 times the bandgap of silicon. If one were to use a threshold energy of 2 times the bandgap, which is the upper limit of accepted values, the substrate current would be smaller, resulting in a higher I_g/I_{sub} ratio. As mentioned previously, the threshold for the gate current depends upon the gate voltage and the oxide thickness, which are both essential in determining the electric field in the oxide.

Chung *et al* [2] obtained experimental evidence that hot-electron degradation worsens as the effective channel length is reduced due to the nonscalability of the degraded region of the channel. Even at the same value of substrate current, the hot-electron degradation was worsened with a shorter effective channel length [42]. Also, they found that for a given value of $I_{\text{sub}}/W_{\text{eff}}$ less device degradation occurred as the oxide thickness was reduced. It is expected in that case that less trapping and interface-state generation occur due to the reduced time that the hot electrons spend in the oxide. Hot-carrier effects continue to be an issue at biases which are substantially below the barrier height for injection into the oxide and the threshold for impact ionization. The occurrence of these effects can be attributed to electron–electron scattering. Electron–electron scattering most likely has a greater effect at such short channel lengths. This is because in scaled devices, the doping concentration is higher than in long channel devices, and the rate of electron–electron scattering depends on the density of electrons in the device. The results yield lifetimes that are quite short. This can be ascribed to an increase in the amount of Coulomb collisions at these short channel lengths. The high doping levels (2.0×10^{20} cm^{-3}) for the 70 nm channel length device translate into high electron densities and a large amount of Coulomb scattering. Hence, there is a strong effect which points toward decreased lifetimes for ultra-short channel devices.

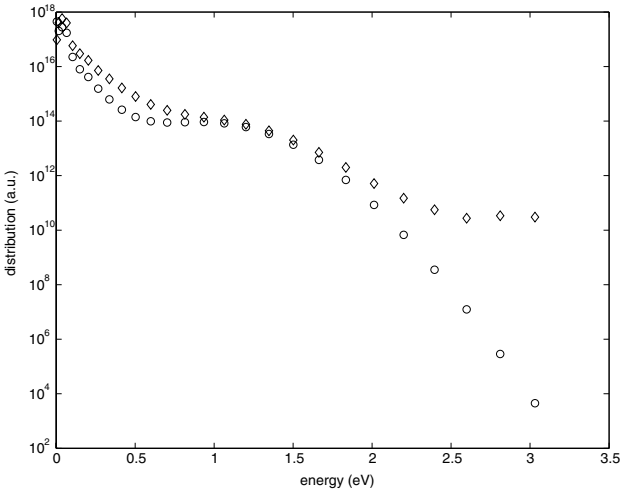

Figure 9. This figure shows the electron energy distribution for the 250 nm MOSFET with a FIB implant with and without electron–electron scattering included in the simulation. The cell is located in the channel at the midpoint between source and drain. The substrate doping concentration is 1.0×10^{16} cm^{-3}, the source and drain doping concentration is 1.0×10^{19} cm^{-3}, and the 70 nanometer-wide FIB region has a doping concentration of 1.8×10^{18} cm^{-3}. 3.0 V is applied on both the gate and the drain. The symbol (○) represents the distribution without electron–electron scattering, and the symbol (◊) represents the distribution when electron–electron scattering is included. The potential in the cell is 0.81 volt.

5.3. 250 nm transistor

Kang *et al* [43] developed an optimization technique for threshold voltage characterization based on 2D device simulation and 3D V_T contour mapping. Based on their results of a constant threshold voltage for a device with and without a FIB implant, we include results for a 250 nm transistor. At this channel length, the standard MOSFET has a substrate doping concentration of 1.0×10^{18} cm^{-3} and a source/drain doping concentration of 1.0×10^{19} cm^{-3}. The MOSFET with the FIB implant region has a substrate doping concentration of 1.0×10^{16} cm^{-3}, source and drain doping concentration of 1.0×10^{19} cm^{-3}, and a 70 nm wide FIB region with a doping concentration of 1.8×10^{18} cm^{-3}. This yields a threshold voltage of 0.68 V for both cases. Figure 9 shows the energy distribution with and without electron–electron scattering for the 250 nm MOSFET with a FIB implant. For the 250 nm MOSFET with the FIB implant, the ratio of gate-to-substrate current is 8.32×10^{-7}, which yields a lifetime of 6.45×10^3 s, using the model of Cui *et al* [26]. The 250 nm MOSFET without the FIB implant, with the same threshold voltage as the device with the FIB implant, has a ratio of gate-to-substrate current of 3.8×10^{-9}, which yields a lifetime of

1.87×10^3 s. Since the substrate doping is very high for the standard MOSFET (1.0×10^{18} cm^{-3}), the lifetime for the device is significantly lower than for the MOSFET that includes the FIB implant. Yet, the comparison of the lifetimes shows the improvements in reliability provided by the FIB implant. The use of the FIB implant in the channel allows the substrate doping to be lower, resulting in less electron–electron scattering. Thus, the lifetime model predicts a longer lifetime for the device with the FIB implant as expected.

6. Conclusions

Kinetic simulation by the transition probability matrix (TPM) method of the CS was used in this work to calculate the electron energy distribution in nanometer-scale nMOSFETs. In order to reconcile short channel effects and hot-carrier reliability, an asymmetric device structure, the FIBMOS (focused-ion-beam MOSFET) was used in the simulations in addition to a MOSFET with uniform doping in the channel. High doping is used in the channel of ultra-small devices in order to prevent punchthrough, although this leads to high electric fields in the device. The FIB implant helps to reduce the electric field in the channel region of the nanometer-scale device and allows a lower substrate doping concentration than would be required without the FIB implant to achieve the same threshold voltage. The potential and electron density at each node of a two-dimensional mesh were obtained from the drift–diffusion simulation tool SGFramework and used as input for the kinetic simulation. A single-particle electron–electron scattering approach was used which enabled the use of energy-dependent emission and absorption scattering rates for electron–electron scattering. Worst-case stress conditions $V_g \simeq V_d$ were used in the simulations. The Coulomb scattering was seen to enhance the high energy tail of the electron distribution in the 35 nm and 70 nm channel length transistors. The electron–electron scattering is seen to play a significant role in the heating of electrons in the MOSFET channel for nanometer-scale devices. The ratio of gate-to-substrate current is calculated, which is an indicator of the severity of degradation effects in the device. The ratio of I_g/I_{sub} is observed to depend strongly on the threshold energies used in the calculation of the currents. At nanometer-scale channel lengths, there are degrading and non-degrading components of substrate and gate current. The non-degrading components correlate strongly, so that the ratio I_g/I_{sub} is an accurate indicator of the level of device degradation. We have assumed that the model for device lifetimes based on the ratio of gate-to-substrate currents is an efficient method for predicting the lifetimes and also that extrapolation of the model down to shorter channel length devices is an accurate assessment of the lifetimes. Additionally, the effects of electron–electron scattering are increased for nanometer-scale channel length devices, because the electron–electron scattering rate depends on the density of electrons in the channel. Degradation effects are likely to have an effect on the transistor performance even at such short channel lengths, as evidenced by the existence of gate and substrate currents. Devices at this scale are fabricated with higher doping concentrations, due to the requirement for shallow junctions, which implies greater electron densities in the channel. This leads to a greater amount of Coulomb collisions and adds significantly to the amount of electrons in the tail of the distribution. Such an increase in the amount of Coulomb collisions may lead to significantly reduced lifetimes in nanometer-scale devices.

Appendix. Overview of kinetic model

In this work the CS is used to find a set of probabilities for use in an iterative scheme which iterates in order to find collision rates in cells. The CS tracks a group of particles which just had a collision in a given initial cell, in order to determine where they have their next collision. Depending on their velocity after the collision, they are divided into different groups. Each group is followed as they move 'ballistically' and when they have their next collision, their location and energy are recorded. The fractions of the particles going to each final phase space cell then yield the probabilities needed by the iterative calculation of the rates.

The CS consists of a ballistic move, which is then followed by collisions. The ballistic move portion can be represented by the Vlasov equation, also referred to as the collisionless Boltzmann equation, which is given by

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla f + \frac{d\mathbf{k}}{\mathrm{d}t}\cdot\frac{\partial f}{\partial \mathbf{k}} = 0. \tag{A.1}$$

The Vlasov equation is integrated along the characteristic curves, which are given by $\mathrm{d}\mathbf{x}/\mathrm{d}t = \mathbf{v}$ and $\hbar\mathrm{d}\mathbf{k}/\mathrm{d}t = \mathbf{F}$. Phase space cells, containing a density of electrons, follow the trajectory of the characteristic curves. This implements Liouville's theorem, which states that along the trajectory of any phase point the probability density in the neighborhood of the point remains constant in time [44]. Collisions are included at the end of a timestep. This method of characteristics, the CS, operates as follows [7]. The electron (or hole) distribution function $f(x, \mathbf{v})$ is advanced in time according to a propagator $p(x, \mathbf{v}, x'', \mathbf{v}'', \Delta t)$ such that

$$f(x, \mathbf{v}, t + \Delta t) = \int f(x'', \mathbf{v}, t)\cdot p(x, \mathbf{v}, x'', \mathbf{v}'', \Delta t)\,\mathrm{d}x''\,\mathrm{d}\mathbf{v}''. \tag{A.2}$$

The propagator, or Green's function, $p(x, \mathbf{v}, x'', \mathbf{v}'', \Delta t)$, determines what fraction of particles move from cell $(x'', \mathbf{v}'')$ to cell $(x, \mathbf{v})$ in the time step Δt. The propagators allow Boltzmann's equation to be solved by successive convolutions of the 'old' distribution function with the propagator.

The CS updates the distribution each iteration by calculating successive scattering rates. Figure A1 demonstrates an example of calculating the scattering rates in a cell. Here we describe more specifically how the CS computes the probabilities of particles moving from one cell to a subsequent cell and the probabilities of moving from one energy to another due to collisions in that spatial cell. Using the TPM, or propagator, the subsequent iteration takes the distribution of scattered particles and advances it to the next cells where scattering takes place [12]. The quantity of interest is $R(c, E)$, the collision rate of particles in cell c at energy E. There are two portions to the TPM method. There

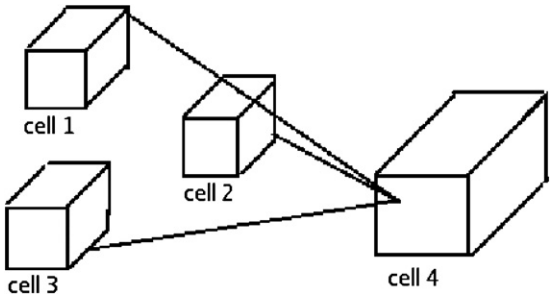

Figure A1. Calculation of the total scattering rate in a cell. The scattering at rate R_1 in cell 1 contributes a scattering rate $P_1^4 R_1$ in cell 4, as P_1^4 is the probability that a particle which just scattered in cell 1 will scatter next in cell 4. Similarly, $P_2^4 R_2$ is the scattering rate contributed by cell 2 and $P_3^4 R_3$ is the scattering rate contributed by cell 3.

is a ballistic, or collisionless portion, followed by the collision operator phase.

The first transition probability matrix computes the number of particles per second that collide in cell c at energy E'',

$$R(c, E'') = \sum_{c'} R(c', E') T_{\text{bal}}(c, E'' : c', E'), \qquad \text{(A.3)}$$

where $R(c, E'')$ is the number rate of particles that collide in cell c at energy E'', $R(c', E')$ is the number rate of particles that collided in cell c' and were redistributed with energy E' in the previous iteration. The way that the rate $R(c', E')$ is iterated on the mesh occurs as follows: consider a spatial cell c' and a group of electrons at energy E'. This group of electrons scatters at a rate $R(c', E')$ and has an angular distribution $f(\theta, \phi, c', E')$. $f(\theta, \phi, c', E')$ is the probability that an electron with energy E' that last scattered in cell c' is moving in a direction within the range ϕ to $\phi + \Delta\phi$ and θ to $\theta + \Delta\theta$. Also, $T_{\text{bal}}(c, E'' : c', E')$ is the probability that a particle having started in cell c' at energy E' will have its next collision in cell c at energy $E'' = E' - q\Delta\Phi$, where E' is the kinetic energy, $\Delta\Phi$ is the change in potential and q is the magnitude of the charge on an electron. The sum is over all mesh cells c' at energy E'.

The second transition probability matrix redistributes the particles after a collision,

$$R(c, E) = \sum_{E''} R(c, E'') T_{\text{col}}(E : E''), \qquad \text{(A.4)}$$

where $T_{\text{col}}(E : E'')$ is the probability that a particle, having previously collided in cell c at energy E'' will be redistributed with energy E' within the same spatial cell c. Equation (A.4) refers to particles which had energy E'' before collision and end up with energy E after the collision. They stay in the same spatial cell during the collision, so the spatial label c is unchanged. The rate of particles colliding in cell c with initial energy E'' is $R(c, E'')$. The fraction of these which have final energy E is $T_{\text{col}}(E : E'')$. Thus particles coming in with energy E'' contribute a rate of particles arriving at energy E in cell c of $R(c, E) = R(c, E'')T_{\text{col}}(E : E'')$. The total $R(c, E)$ must be found by summing over all initial energies E''.

References

[1] Pop E, Sinha S and Goodson K E 2006 *Proc. IEEE* **94** 1587–601
[2] Chung J E, Jeng M-C, Moon J, Ko P-K and Hu C 1990 *IEEE Trans. Electron Devices* **37** 1651
[3] Tam S, Hsu F-C, Muller R S and Ko P-K 1983 *IEEE Electron Device Lett.* **4** 249
[4] Mietzner T, Jakumeit J and Ravaioli U 2001 *IEEE Trans. Electron Devices* **48** 2323
[5] Ang D S, Phua T W H, Liao H and Ling C H 2003 *IEEE Electron Device Lett.* **24** 469–71
[6] Hitchon W N G, Parker G J and Lawler J E 1991 *IEEE Trans. Plasma Sci.* **19** 113–21
[7] Hitchon W N G, Parker G J and Lawler J E 1993 *IEEE Trans. Plasma Sci.* **21** 228–38
[8] Parker G J, Hitchon W N G and Lawler J E 1994 *Phys. Rev.* E **50** 3210–9
[9] Christlieb A J, Hitchon W N G and Keiter E R 2000 *IEEE Trans. Plasma Sci.* **28** 2214–31
[10] Ferry D K 1991 *Semiconductors* (New York: Macmillan)
[11] Lundstrom M and Ren Z 2002 *IEEE Trans. Electron Devices* **49** 133–41
[12] Parker G J, Hitchon W N G and Keiter E R 1996 *Phys. Rev.* E **54** 938
[13] Christlieb A J and Hitchon W N G 2002 *Phys. Rev.* E **65** 056708–1
[14] Childs P A and Leung C C C 1996 *J. Appl. Phys.* **79** 222
[15] Dyke D W *et al* 1996 *Proc. SISPAD Conf. (Tokyo)* pp 101
[16] McMahon W, Haggag A and Hess K 2003 *IEEE Trans. Nanotechnol.* **2** 33
[17] Hess K, Tuttle B, Register F and Ferry D K 1999 *Appl. Phys. Lett.* **75** 3147
[18] Hitchon W N G, Parker G J and Lawler J E 1994 *IEEE Trans. Plasma Sci.* **22** 267
[19] Longmire C L 1963 *Elementary Plasma Physics* (New York: Wiley)
[20] Knezevic I, Vasileska D Z and Ferry D K 2002 *IEEE Trans. Electron Devices* **49** 1019–26
[21] Ferry D K 2000 *Semiconductor Transport* (London: Taylor and Francis)
[22] Ferry D K, Goodnick S M and Hess K 1999 *Physica* B **272** 538
[23] Takenaka N, Inoue M and Inuishi Y 1979 *J. Phys. Soc. Japan* **47** 861
[24] Goodnick S M and Lugli P 1988 *Solid State Electron.* **31** 463
[25] Takeda E and Suzuki N 1983 *IEEE Electron Device Lett.* **4** 111–3
[26] Cui Z, Liou J J, Yue Y and Wong H 2005 *Solid-State Electron.* **49** 505–11
[27] Ker M-D and Wu C-Y 1995 *IEEE Trans. Electron Devices* **42** 1141
[28] Tam S, Ko P-K and Muller R S 1982 *IEEE Trans. Electron Devices* **29** 1740–4
[29] Huang C, Wang T, Chen C N, Chang M C and Fu J 1992 *IEEE Trans. Electron Devices* **39** 2562–8
[30] Tam S, Ko P-K and Hu C 1984 *IEEE Trans. Electron Devices* **31** 1116–25
[31] Fischetti M V, DiMaria D J, Dori L, Batey J, Tierney E and Stasiak J 1987 *Phys. Rev.* B **35** 4404–15
[32] Agostinelli V M (Jr), Bordelon T J, Wang X, Hasnat K, Yeap C-F, Lemersal D B Jr, Tasch A F and Maziar C M 1994 *IEEE Trans. Electron Devices* **41** 1784–95
[33] Hasnat K, Yeap C-F, Jallepalli S, Shih W-K, Hareland S A, Agostinelli V M, Tasch A F and Maziar C M 1996 *IEEE Trans. Electron Devices* **43** 1264–73
[34] Ning T H, Osburn C M and Yu H N 1977 *J. Appl. Phys.* **48** 286–93

[35] Chen Y-Z and Tang T-W 1988 *IEEE Trans. Electron Devices* **35** 2180–8
[36] Takeda E, Yang C Y and Miura-Hamada A 1995 *Hot-Carrier Effects in MOS Devices* (San Diego: Academic)
[37] Kramer K M and Hitchon W N G 1997 *Semiconductor Devices: a Simulation Approach* (Englewood Cliffs, NJ: Prentice-Hall)
[38] Hokazono A, Balasubramanian S, Ishimaru K, Ishiuchi H, Hu C and Liu T-J K 2006 *IEEE Electron Device Lett.* **27** 605–8
[39] Streetman B G and Banerjee S K 2006 *Solid State Electronic Devices* (Englewood Cliffs, NJ: Prentice-Hall) p 322
[40] Sofield C J and Stoneham A M 1995 *Semicond. Sci. Technol.* **10** 215–44
[41] Groeseneken G, Bellens R, Van den bosch G and Maes H E 1995 *Semicond. Sci. Technol.* **10** 1208–20
[42] Takeda E, Jones G A C and Ahmed H 1985 *IEEE Trans. Electron Devices* **32** 322–7
[43] Kang J, He X, Vasileska D and Schroder D K 2001 *VLSI Des.* **13** 251–6
[44] Harris S 1971 *An Introduction to the Theory of the Boltzmann Equation* (New York: Holt, Rinehart and Winston)