RC 20203 (9/26/95)
Physical Sciences

# IBM Research Report

## Monte Carlo Study of Sub-Band-Gap Impact Ionization in Small Silicon Field-Effect Transistors

M. V. Fischetti and S. E. Laux

IBM Research Division
T.J. Watson Research Center
Yorktown Heights, New York



IBM Research Division
Almaden · T.J. Watson · Tokyo · Zurich

# Monte Carlo Study of Sub-Band-Gap Impact Ionization in Small Silicon Field-Effect Transistors

M. V. Fischetti and S. E. Laux

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

**Abstract**

The observation of impact ionization in small Si FETs at sub-band-gap source-to-drain bias is explained by the presence of high-energy tails in the electron energy distribution around the drain region. These tails are caused by the strong thermalizing effect of dynamically-screened electron-electron interactions. Under these bias conditions, a larger band-gap, enhanced dielectric screening, and reduced high-energy tails in the distribution function at the source cause a reduction of impact ionization at lower lattice temperatures.

## Introduction

The observation of substrate currents at sub-band-gap ($E_g$) source-to-drain bias ($V_{ds}$)[1-4] in small ($< 0.2\mu$m) Si FETs, as well as their 'anomalous' temperature dependence (*i.e.*, reduced ionization at low temperature for $eV_{ds} < E_g$, the so-called 'crossover' effect[1,3-6]), has been considered a puzzling result[3, 4]: Eitan *et al.*[1] speculated that the 'source function' (*i.e.*, the Fermi tail in the source) could be responsible for both observations, while Tam and co-workers[2] invoked the thermalizing effect of the electron-electron interaction in the channel in order to explain the presence of carriers at energies larger than the applied bias. Recently, Esseni *et al.*[4] presented these observations as suggestive that something is amiss in our understanding of electron transport in Si devices. Here we intend to show that there is nothing 'mysterious' about these phenomena. Considering the growing importance of low-bias transport in very small devices, a clarification of these issues is not a matter of simple academic interest. Full-band Monte Carlo simulations of short devices[7] indicate that most of the ionization events occur in the region around the channel-drain junction and even within the drain. In these regions the electron-electron interactions are indeed responsible for 'thermalizing' the distribution function and placing carriers at energies above the maximum supplied bias. Source-function effects[1] and the well-known existence of high-energy tails above the supply voltage[8, 9] can only account for a small fraction of the high-energy tails. Dynamically-screened Coulomb interactions are indeed essential ingredients to account for the observations.

## Transport Model

The results presented here have been obtained using our self-consistent, two-dimensional Poisson/Monte Carlo program DAMOCLES described before[7, 10]. Electron-phonon and impact ionization rates are obviously of utmost importance in determining the distribution functions. Therefore, it is worth stressing how the electron-phonon scattering rates we have used[10], are consistent with empirical[11, 12] as well as theoretical[13, 14] results. Similarly, our impact-ionization rate[15] has been confirmed by recent pseudopotential calculations[16-19]. This growing and long overdue agreement on the strength of the basic scattering processes in Si is what makes it possible to study high-energy processes in Si with some degree of confidence. We show in Fig. 1 the energy dependence of the scattering rates (averaged over equienergy surfaces) we have used, compared with the most recent results reported in the literature.

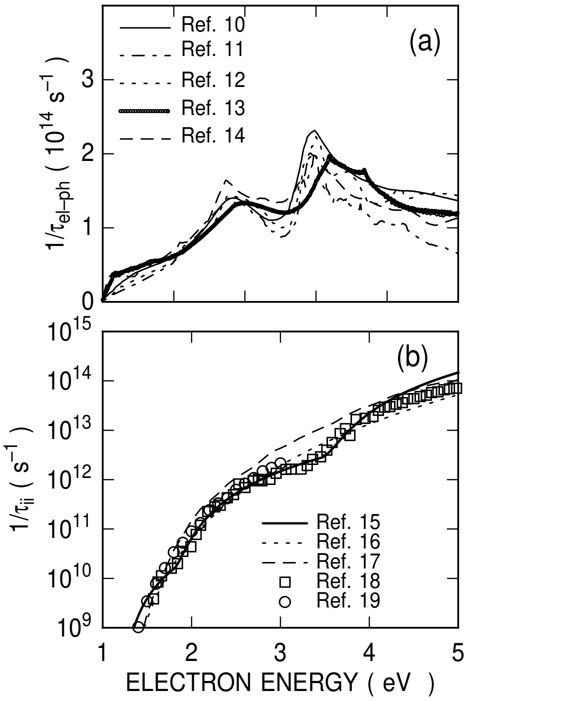


Fig. 1: Electron-phonon (a) and impact-ionization (b) scattering rates in Si versus electron energy at 300 K, averaged over equienergy-surfaces, taken from the references indicated.

The remaining scattering process deserving comments is the short-range Coulomb interaction among electrons. This process is described by a matrix element $M$ which,

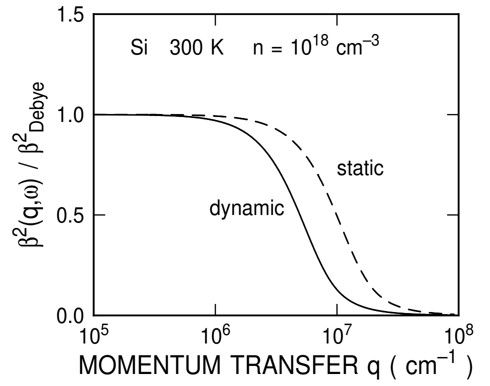


Fig. 2: Screening parameter normalized to its Debye-Hückel value as a function of momentum-transfer in the static (zero-energy transfer, dashed line) and dynamic for the special case of maximum energy-exchange between the two electrons (solid line) computed in the simple case of parabolic valleys and in the high-temperature, quasi-equilibrium approximation of Fetter and Walecka[20].

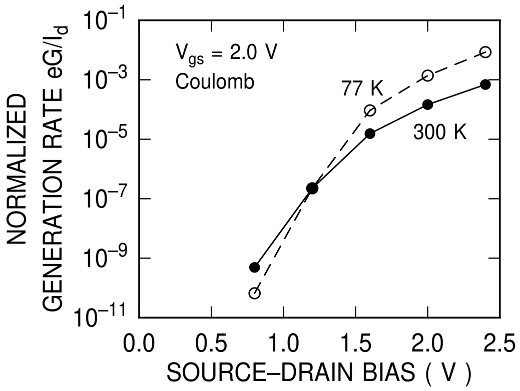


Fig. 3: Generation rate, $G$, normalized to the drain current, $I_d$, as a function of source-drain bias at two temperatures.

in its simplest form (Born approximation, no spin, distinguishable particles, overlap factors between Bloch states set to unity; but see [7] for the expression we have actually used) is:

$$M = \frac{e^2}{\epsilon_s} \frac{1}{\mathbf{q}^2 + \beta(\mathbf{q},\omega)^2}, \quad (1)$$

where $e$ is the magnitude of the electron charge, $\epsilon_s$ is the static, long-wavelength permittivity of Si associated with the response of the valence electrons, $\mathbf{q} = \mathbf{k}_f - \mathbf{k}_i$ is the momentum transfer ($\mathbf{k}_i$ and $\mathbf{k}_f$ are the quasi-wavevectors of the electron before and after the collision), and $\hbar\omega$ is the energy transfer. The dielectric response of the free carriers is represented by the screening parameter $\beta$. Usually, its frequency (*i.e.*, energy transfer) and wavelength (*i.e.*, momentum transfer) dependencies are ignored and the simple static, long-wavelength Debye-Hückel expression is employed. However, in small FETs this might be a very bad approximation. We show in Fig. 2 the ratio of the screening parameter (calculated in the high-temperature approximation of Fetter and Walecka[20]) to its (nondegenerate) Debye value in Si. It can be seen that screening becomes negligible at large wavevectors, even in the static approximation. Physically, this means that perturbing potentials of very short wavelength cannot be screened. The electron gas supports collective oscillations (plasmons) of long wavelength, but as the wavelength of the plasmons shortens, they decay (Landau damping) and the electron gas stops responding. Similarly, Fig. 2 shows that at high frequencies (above the plasma frequency), the electron gas is unable to respond.

Along most of the channel of a very short Si FET the Debye approximation is valid: The quasi-ballistic nature of transport implies very narrow energy-distributions (*i.e.*, low electron temperatures, but high average energies). Thus, the interaction among electrons involve small energy (and, often, quasi-momentum) transfers. The relatively low density of free carriers and the strong static, long-wavelength Debye screening dictate that a small role is played by Coulomb interactions along most of the channel. Thus, the suggestion that Coulomb interactions could be invoked to use a Maxwellian distribution along the channel of a short FET[2] does not appear to be well founded. On the contrary, as we approach the drain we face the possibility that a 'cold' electron in the heavily-doped drain region interacts with a 'hot' electron coming from the channel. In this case, classically the most likely event is a collision in which the hot channel-electron exchanges (loses or, less frequently but very importantly, gains) a small fraction of its kinetic energy with a drain-electron. But since electrons are undistinguishable particles, the 'exchange' collision is obtained by switching the role of the two carriers. This results in a very large energy transfer. The matrix element corresponding to this event is thus essentially unscreened: In simple words, the fast 'channel electron' is moving too fast for the sluggish 'drain electrons' to follow its potential and screen it. As a result, the Coulomb interaction is extremely effective in thermalizing the distribution functions, typical thermalization times being in the 100 fs range.

## Results

SOI devices show dramatically the onset of impact ionization at low bias: their Si layer collects the generated holes, with the threshold-voltage shift acting as an integrator of the generated charge, resulting in the observation of the 'kink effect'[21] at $V_{ds}$ as low as 0.6 V[22]. Here we have considered an ideal 0.25 $\mu$m gate-length, SOI MOSFET having a 7 nm-thick gate oxide, an 80 nm Si layer uniformly doped ($10^{17}$ cm$^{-3}$) sitting on a buried oxide on a Si substrate, and an effective channel length of about 0.18 $\mu$m. Fig. 3 shows the calculated generation rate normalized to the drain current as a function of $V_{ds}$ at a fixed gate bias. Note the persistence of impact ionization below a $V_{ds}$ of 1 V, as well as the 'crossover effect', the integrated generation rate being larger at 300 K than at 77 K below about 1.2 V. In Fig. 4 we show the results of simulations performed under three different conditions: at 300 K and at 77 K with electron-electron interaction included (labeled 'Coulomb' in the figures), and at 300 K without either the long-range (achieved by switching off the Poisson/Monte Carlo self-consistency and using a 'frozen' field obtained from the time-averaged self-consistent solution) or the short-

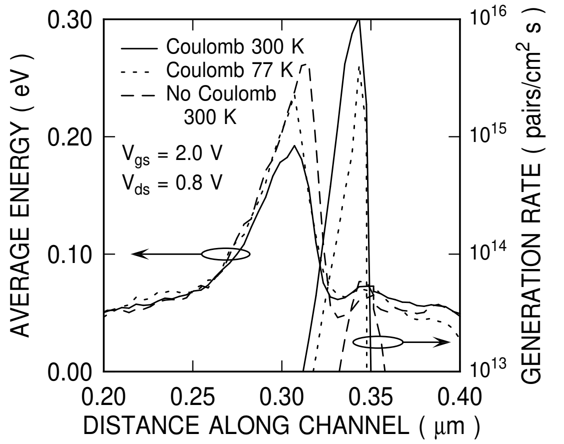


Fig. 4: Calculated average energy along the top silicon-oxide interface and generation rate integrated over the 80 nm Si layer in a 0.25 $\mu$m gate-length SOI MOSFET. The drain contact is at the right of the figure and the metallurgical channel/drain junction is at about 0.29 $\mu$m.

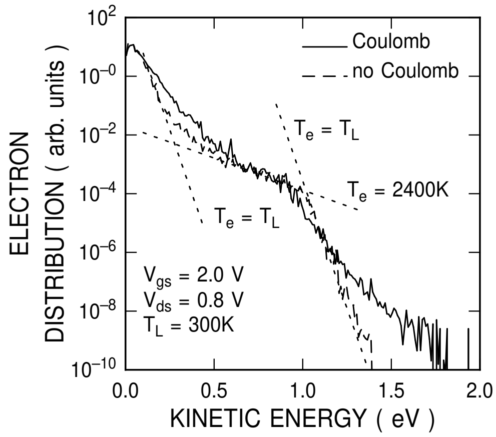


Fig. 5: Electron distribution just inside the drain (at x = 0.34 $\mu$m in Fig. 4), where the generation rate peaks, as obtained from simulations performed under the conditions of Fig. 4.

range Coulomb interactions. The figure, relative to results obtained at a gate bias of 2.0 V and 'sub-band-gap' source-to-drain bias of 0.8 V, shows the generation-rate along the Si layer, integrated over the layer. The average kinetic energies along the top Si-SiO$_2$ interface are also shown: At 300 K the Coulomb interactions reduce the maximum energy attained by the electrons: The long-range interactions are directly responsible for inelastic collisions with plasmons, and the short-range interactions, while not directly affecting the energy-loss rate of the electron ensemble, increase the momentum relaxation rate, thus enhancing the energy-loss rate caused by inelastic processes. At 77 K, higher energies are observed, as expected. On the contrary, little correlation (actually: an anti-correlation) is seen between maximum average energies and generation rates: The total generation rate, integrated over the device, is of $1.1 \times 10^{10}$ pairs/(cm s) for the Coulomb case at 300 K, and only $1.0 \times 10^8$ pairs/(cm s) when the Coulomb interaction is turned off. When accounting for the Coulomb interactions, at low temperature the total integrated generation rate drops by almost a factor of 4, to $2.7 \times 10^9$ pairs/(cm s). This contrasts the 'usual' high-bias case for which the much higher generation rate is depressed at higher temperatures.

The cause of this 'anomalous' low-bias behavior is seen in Fig. 5. Here we show distributions of the electrons after they have just entered the field-free region in the drain. In the absence of Coulomb interactions (dashed line) the distribution function exhibits three distinct regions: 1) The low-energy, quasi-thermal contribution of the drain-electrons, 2) the middle-energy 'hump' associated with the hot carriers coming from the channel, at a temperature, $T_e$, fixed by the accelerating field in the channel, and 3) a high-energy tail, above the supply-voltage (0.8 V plus the the Fermi-level in the source), at a temperature which approaches the lattice temperature, $T_L$, at large enough energies[8, 9]. In the presence of Coulomb interactions (solid line), the distribution is 'thermalized' and approaches a Maxwellian, despite the fact that the 'hot' channel electrons have barely entered the drain. The strong influence of the short-range Coulomb interaction on the high-energy tails at very low bias has been independently observed also by Abramo *et al.*[23], even for the case of static Debye-Hückel screening. But, as we stated above, dynamic (and exchange) corrections render this effect even more dramatic. At 77 K the distribution (not shown) is very similar to the 300 K case, with a marginally lower high-energy tail. This is caused in part by 1) enhanced screening, resulting in reduced Coulomb scattering, 2) reduced phonon scattering, resulting in narrower quasi-ballistic distributions with reduced high-energy tails because of reduced phonon absorptions, and, 3) to a much smaller extent and in extremely small devices, a smaller 'above supply voltage tail' in the 'source function' (*i.e.*, there are fewer electrons above the Fermi level in the source of the device). What really reduces the strength of the impact-ionization process at lower temperatures is the wider gap: It is only a 60 meV effect, or so, but it matters almost exponentially at biases below band-gap. Therefore, this is an effect of utmost importance at $V_{ds}$ near or below the ionization threshold, but negligible at large bias. Reduced phonon absorption may become a very significant factor explaining this 'anomalous' temperature dependence of impact ionization in even smaller devices as we move to a still small but above band-gap bias, as shown recently by Mastrapasqua and Bude[24] and by Sano and Yoshii[25].

## Conclusions

In conclusion, we have shown that there is nothing mysterious about the observation of impact ionization processes at sub-band-gap source-to-drain bias: Very seldom one does see an 'abrupt' threshold in Nature, broadening and thermalizing effects almost invariably smoothing-out a sharp threshold expected on the basis of oversimplified models. Having realized the importance of the dynamically-screened Coulomb interactions among electrons, the presence of very soft thresholds must be expected. By the same token, the '3-volt threshold' expected for the onset of electron injection into the gate oxide is as fictitious as the '1-volt threshold' we have just discussed. As a significant example, Fig. 6 shows the distribution function at the drain-end of a 0.15$\mu$m Si n-MOSFET described elsewhere[7]: Even with only 2.5 V applied to the drain, electrons exist in relatively large numbers up to 4.5 eV or so. Thus, gate currents will be observed well below the 3-volt threshold[2, 4, 26], without invoking weak Auger processes[26] or other unknown processes. The only difference we might expect is the absence of the 'crossover effect', as indeed observed experimentally[4]: The reduced role played by the short-range interparticle collisions in the channel of the device (where gate injection is most likely), and the weaker temperature dependence of the threshold (*i.e.*, the Si-$SiO_2$ barrier) will cause a 'normal' behavior, as far as the temperature of the process is concerned. Yet, the basic idea (a smooth, soft threshold induced by the thermalizing effect of Coulomb scattering) controls this process as well.

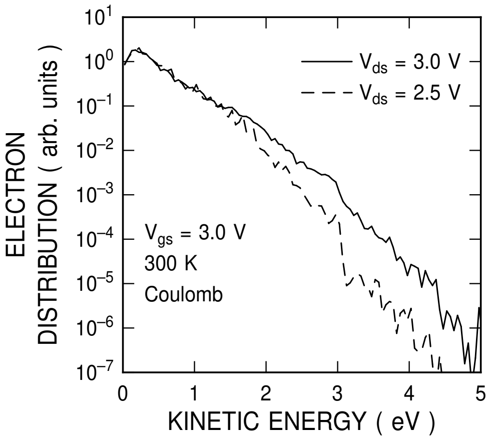


Fig. 6: Electron distribution just inside the drain of a 0.15$\mu$m n-channel Si MOSFET showing the presence of electrons above 3.2 eV even at a $V_{ds}$ of only 2.5 V.


## Aknowledgements

We are pleased to thank A. Abramo, J. Bude, and N. Sano for having informed us of their work prior to publication, and N. Sano for invaluable discussions.


## References


[1] B. Eitan, D. Frohman-Bentchkowsky, and J. Shappir, "Impact ionization at very low voltages in silicon", *J. Appl. Phys.*, vol. 53, pp. 1244-1247, 1982.

[2] S. Tam, F.-C. Hsu, C. Hu, R.S. Muller, and P.K. Ko, "Hot-electron currents in very short channel MOSFETs", *IEEE Electron Device Lett.*, vol. EDL-4, pp. 249-251, 1983.

[3] L. Machanda, R.H. Storz, R.H. Yan, K.F. Lee, and E.H. Westerwick, "Clear observation of Sub-band gap impact ionization at room temperature and below in 0.1 $\mu$m Si MOSFETs", *IEDM Tech. Dig.*, pp. 994-996, 1992.

[4] D. Esseni, L. Selmi, R. Bez, E. Sangiorgi, and B. Riccó, "Bias and temperature dependence of gate and substrate currents in n-MOSFETs at low drain voltage", *IEDM Tech. Dig.*, pp. 307-310, 1994.

[5] A.K. Henning, N.N. Chan, J.T. Watt, and J.D. Plummer, "Substrate current at cryogenic temperatures: Measurements and a two-dimensional model for CMOS technology", *IEEE Trans. Elec. Dev.*, vol. ED-34, pp. 64-74, 1987.

[6] M. Koyanagi, T. Matsumoto, M. Tsuno, T. Shimatani, Y. Yoshida, and H. Watanabe, "Impact ionization phenomenon in 0.1 $\mu$m MOSFET at low temperature and low voltage", *IEDM Tech. Dig.*, pp. 341-344, 1993.

[7] M.V. Fischetti, S. E. Laux, and E. Crabbé, "Understanding hot-electron transport in silicon devices: Is there a short-cut?", *J. Appl. Phys.*, vol. 78, pp. 1058-1087, 1995.

[8] A. Lacaita, "Effective temperature of hot electrons: why it approaches the lattice temperature", *Semicond. Sci. Technol.*, vol. 7, pp. B590-B592, 1992.

[9] C.C.C. Leung and P.A. Child, "On the above supply voltage hot carrier distribution in semiconductor devices", *Appl. Phys. Lett.*, vol. 66, pp. 162-164, 1995.

[10] M.V. Fischetti and S.E. Laux, "Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects", *Phys. Rev. B*, vol. 38, pp. 9721-9745, 1988.

[11] H. Shichijo, J.Y. Tang, J. Bude, and P.D. Yoder, "Full band Monte Carlo program for electrons in silicon", in *Monte Carlo Device Simulation: Full Band and Beyond*, K. Hess Ed., Boston, MA: Kluwer Academic Press, 1991, pp. 285-307.

[12] T. Kunikiyo, M. Takenaka, Y. Kamakura, M. Yamaji, H. Mizuno, M. Morifuji, K. Taniguchi, and C. Hamaguchi, "A Monte Carlo simulation of anisotropic electron transport in silicon including full band structure and anisotropic impact-ionization model", *J. Appl. Phys.*, vol. 75, pp. 297-312, 1994.

[13] P.D. Yoder, J.M. Higman, J. Bude, and K. Hess, "Monte Carlo simulation of hot electron transport in Si using a unified pseudopotential description of the crystal", *Semicond. Sci. Technol.*, vol. 7, pp. B357-B359, 1992.

[14] M.V. Fischetti and J.M. Higman, "Theory and calculation of the deformation potential electron-phonon scattering rates in semiconductors", in *Monte Carlo Device Simulation: Full Band and Beyond*, K. Hess Ed., Boston, MA: Kluwer Academic Press, 1991, pp. 123-160.

[15] E.A. Cartier, M.V. Fischetti, E.A. Eklund, and F.R. McFeely, "Impact ionization in silicon", *Appl. Phys. Lett.*, vol. 62, pp. 3339-3341, 1993.

[16] Y. Kamakura, H. Mizuno, M. Yamaji, M. Morifuji, K. Taniguchi, C. Hamaguchi, T. Kunikiyo, and M. Takenaka, "Impact ionization model for full band Monte Carlo simulation", *J. Appl. Phys.*, vol. 75, pp. 3500-3506, 1994.

[17] N. Sano and A. Yoshii, "Impact ionization rate near thresholds in Si", *J. Appl. Phys.*, vol. 75, pp. 5102-5105, 1994.

[18] J. Kolnik, Y. Wang, I.H. Oguzman, and K.F. Brennan, "Theoretical investigation of wave-vector-dependent analytical and numerical formulations of the interband impact-ionization transition rate for electrons in bulk silicon and GaAs", *J. Appl. Phys.*, vol. 76, pp. 3542-3551, 1994.

[19] J. Bude and M. Mastrapasqua, "Impact ionization in sub-micron nMOSFET technologies", *IEEE Electron Device Lett.*, in press.

[20] A.L. Fetter and J.D. Walecka, *Quantum Theory of Many-Particle Systems*, New York: McGraw-Hill, 1971, p. 305.

[21] B. Dierickx, L. Warmerdam, E. Simoen, J. Vermeiren, and C. Claeys, "Model for hysteresis and kink behavior of MOS transistors operating at 4.2 K", *IEEE Trans. Electron Devices*, vol. 35, pp. 1120-1125, 1988.

[22] G. Shahidi, private communication.

[23] A. Abramo, C. Fiegna, and F. Venturi, "Hot carrier effects in short MOSFETs at low applied voltages", this digest.

[24] M. Mastrapasqua and J.D. Bude, "Electron and hole impact ionization in deep sub-micron MOSFETs", *Microelectronic Eng.*, vol. 28, pp. 293-300, 1995.

[25] N. Sano and A. Yoshii, "Impact ionization in sub-micron and sub-0.1 $\mu$m Si MOSFETs", presented at the IX Hot Carrier in Semiconductor International Conference, HCIS IX, Chicago, Illinois, July 31-August 4, 1995.

[26] B. Riccó, E. Sangiorgi and D. Cantarelli, "Low voltage hot-electron effects in short channel MOSFETs", *IEDM Tech. Dig.*, pp. 92-95, 1984.