RC23381 (W0410-119) October 20, 2004
Physics

IBM Research Report

Thirty Years of Monte Carlo Simulations of Electronic Transport in Semiconductors: Their Relevance to Science and Mainstream VLSI Technology

M. V. Fischetti, S. E. Laux, P. M. Solomon, A. Kumar
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g., payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home.

Thirty Years of Monte Carlo Simulations of Electronic Transport in Semiconductors: Their Relevance to Science and Mainstream VLSI Technology

M.V. FISCHETTI, S.E. LAUX, P.M. SOLOMON, and A. KUMAR
IBM Semiconductor Research and Development Center, Research Division
Thomas J. Watson Research Center, P. O. Box 218, Yorktown Heights, New York 10598, USA
fischet@watson.ibm.com

Abstract. We review briefly some aspects of the history of Monte Carlo simulations of electronic transport in semiconductors. In the early days their heavy computational cost rendered them suitable only to study problems of pure physics, as simpler models provided the answers necessary to design 'electrostatically good' devices. Now that scaling has taken another meaning (*i.e.*, looking for alternative materials, crystal orientations, device geometries, etc.), Monte Carlo simulations may gain popularity once more, since they allow an efficient and reliable evaluation of speculative ideas. We show examples of both aspects of the results of Monte Carlo work.

Keywords: Electronic transport, Monte Carlo methods, MOSFETs, scaling

1. Introduction

At the 1996 ESSDERC one of us (MVF) gave a review entitled "Monte Carlo simulations of electron transport in Si: The first twenty years"[1]. At the time, emphasis was placed on the relevance of Monte Carlo (MC) simulations to science – such as on the determination of the electron-phonon and impact ionization scattering rates, or on the suitable band-structure model to employ – but much less on their relevance to technology. Perhaps this was because such an impact on VLSI technology was, sadly, absent. Yes: MC solutions of the Boltzmann Transport Equation (BTE) had helped the improvement of 'moments' approximations, such as the drift-and-diffusion, energy-transport, and hydrodynamic models[2]. But not much else... Let us recall that at that time VLSI technology was evolving along the predictable evolutionary path of scaling Si MOSFETs. Electrostatic properties (e.g., threshold-voltage and sub-threshold swing) dictated by processing choices and accurately modeled by 'mundane' Poisson solvers were the dominant effects. After all, device designers where mainly concerned with being able to turn-off the device. Transport characteristics had always been – up to then – secondary worries, as device scaling took care of improved performance.

'Scaling', as viewed then, is now gone, replaced by novel device designs (e.g., FINFETs, Double-gate FETs) and novel active materials (e.g., strained-Si, Ge, Si surfaces of 'unusual' orientation, even the never-dying III-V compound semiconductors) as technologists attempt to bypass the (real or perceived) end of this 'conventional scaling'. We do not have 'pre-canned' drift-and-diffusion models of mobility, saturated velocity, or pair-production-vs.-field expressions for these materials, neither do we know the effect of novel gate-insulator materials on device performance. This quasi-revolution in VLSI technology has now given MC simulations a second wind and a practical importance. As was the case a decade ago, science still benefits from the flexibility that the MC method allows, by decoupling physical models from numerical - mainly convergence - issues. Similarly, the reduced dimensions of the devices present interesting new physical aspects, such as (intentionally ignoring here 'quantum concerns') the role of long-range Coulomb interactions and their effects on device performance. But the novelty of the structures and materials considered now in 'real-world' VLSI technology requires the deep physical foundations of MC models: physically accu-

rate band-structure, scattering matrix elements, and a transport model valid all the way to the ballistic limit.

In this paper, after having reviewed a few chapters of the science-related history of MC simulations, we shall delve into this 'real world'.

2. A Brief History

It was at the 1966 "Hot Carriers in Semiconductors" Conference that MC methods – originating from the work done on neutron transport during World War II – were proposed by Kurosawa to study electronic transport in semiconductors.[3] Compound semiconductors, such as GaAs or InP, were the first obvious applications, since the small effective mass of these materials meant that electrons would heat up beyond the quasi-thermal regime – amenable to drift-and-diffusion simulations – and more accurate solutions of the BTE were required to account for the Gunn effect, the presence of Gunn domains, etc. From these pioneering efforts by the Malvern group[4], the foundations of MC simulations were put on solid foundations by Price[5].

Work on Si initially dealt with purely 'scientific' problems: The group at the University of Modena took to measuring and explaining velocity-vs.-field characteristics of electrons and holes in homogeneous Si crystals at various temperatures, paying attention to the strength of intra- and inter-valley processes, valley repopulation, and diffusion constants.[6] Applications to device simulations were scant at first, as the state of computational tools at the time did not allow moving much beyond either homogeneous transport or simple idealized one-dimensional problems.

The availability of faster computers in the 1980's was the factor which triggered an 'explosion' of activity. The group at the University of Illinois at Urbana-Champaign 'did the impossible', by employing the full band-structure of semiconductors (GaAs at first[7], followed by Si[8]) to study high-energy transport. Groups in Japan[9] and the US[10] tackled the problem of transport in low-dimensionality situations, and the first flexible, multi-purpose MC device simulators began to be available. Yet, the computational cost of these methods remained too high for daily use in industrial development. The 1980's are probably best characterized by saying that the need to go beyond the 'thermal transport' approximation of drift-and-diffusion was filled by moments methods (energy-transport and hydrodynamic), MC remaining a 'curiosity' in the eye of the device designer, used only for the occasional calibration of relaxation times.

3. Past science

Having tackled successfully the 'warm electron' physics of the 1970's – thanks to the work by the Modena group – electronic transport in small MOSFETs was next on the list. As devices were approaching the 100 nm channel length, MC simulations appeared to be set for center stage. Velocity overshoot, nonlocal effects (as for pair production), and injection into the gate oxide were all problems which required accurate physical models. Curiosity about the experimental confirmation of velocity overshoot was running high around the end of the 1980's. Hydrodynamic simulations seemed to be unable to tackle the problem (recall the infamous 'spurious velocity overshoot problem') because of their strong dependence on energy and momentum relaxation times. However, the current-voltage characteristics of the devices did not seem to be affected by 'overshoot' even down to effective channel lengths of 70 nm.

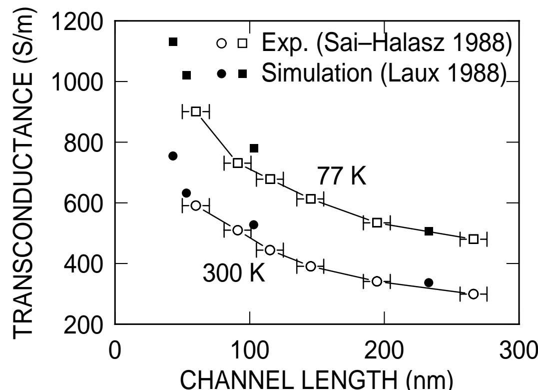

Figure 1: Measured and simulated transconductance vs. metallurgical length for the 'early' sub-100 nm nMOSFETs. Velocity overshoot was looked for, but barely detected in the smaller devices at 77 K.

The early data shown in Fig. 1 – either theoretical[11, 12] or experimental[13] – of the large-signal transconductance of sub-100 nm nMOSFETs did not provide any definitive proof that the electron velocity exceeded, on average across the channel, the saturated velocity. Perhaps a hint was provided by the (electrostatically marginally short-channel) 60 nm devices at 77 K. But it was clear, even then, that the velocity near the source-end of the channel fixed the drain current. Yet, the possibility of 'looking inside the device' provided by MC simulations showed that indeed the saturation velocity was a concept of vanishing usefulness, as was the case for the similar 'quasi-

equilibrium' concept of mobility: Many collisions are required to establish a steady-state drift velocity proportional to the field (as required by 'mobility') or independent of it at sufficiently high fields (as required by the concept of 'saturated velocity'). This large number of collisions is not present in the channel of a sub-250 nm device. Yet, on the grounds that drift-and-diffusion simulations seemed to be able to predict the gross features of the current-voltage characteristics (suitably 'tweaking' a few mobility- or saturated-velocity-related parameters!) many ignored the message that MC simulations were sending about the physics of electronic transport in small devices: Equilibrium and low-energy concepts such as mobility, effective mass, and saturation velocity do not apply anymore. If simpler, near-equilibrium simulations (drift-and-diffusion, energy-transport and/or hydrodynamic models) 'work', it is thanks to the skill of the engineer who tweaks the input parameters and thanks to the fact that near the source things behave 'normally', down to – perhaps – the 100 nm length.

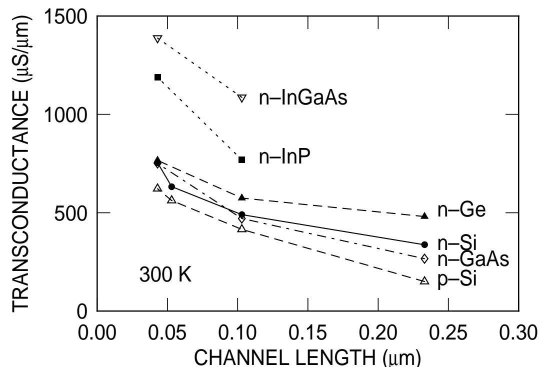

Figure 2: Calculate transconductance vs. metallurgical length for small MOSFETs fabricated on several semiconductors. From these results we argued that low-field mobility did not matter anymore at small length scales.

This concept was reinforced by the results shown in Fig. 2: Despite the huge differences in carrier mobility in GaAs, Ge, and Si, the MC-simulated performance of devices fabricated using these semicondutors do not differ that much. As soon as carriers heat-up above thermal transport, the high-energy, coarse-scale features of the band-structure dominate. The mobility becomes, once more, a meaningless concept. Only In-based compounds, by virtue of the large energy-splitting of their satellite valleys, still exhibit an improved performance.[14] Even this, however, will vanish in the ballistic limit, as we shall see below, if such a limit is at all reachable.

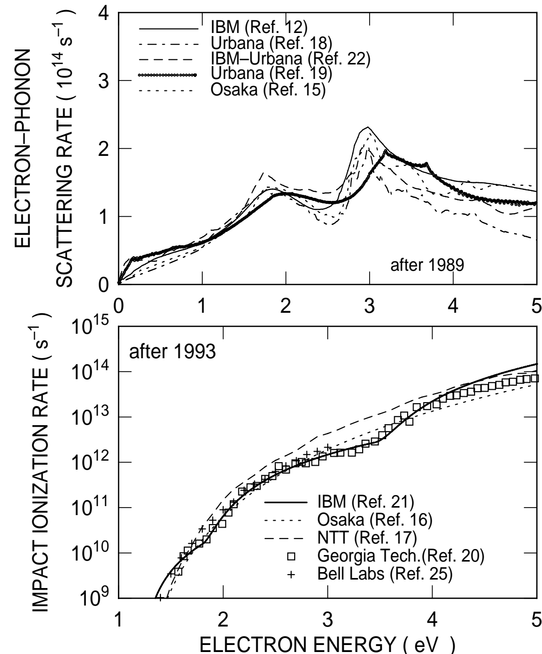

Figure 3: The not-so-trivial agreement among several groups worldwide in assessing the strength of the electron-phonon (top panel) and impact-ionization scattering rates for electrons in Si. The importance of full-band models and of density of final states was the main result of decades of effort required to produce the data plotted here.

What, in our opinion, remains the real triumph of MC method, lies in the role which they have played in helping us understand the strength of electron-phonon and impact-ionization scattering rates in Si. Considering the ubiquitous role played by electron transport in Si in the present technology, the importance of this achievement cannot be overemphasized. Details about this new 'standard model' for electron transport in Si, together with the depressing status of our understanding in the years preceeding 1989 can be found in Ref. [1]. Here it suffices to look at Fig. 3. Electron-phonon and impact-ionization scattering rates are shown as employed/calculated/measured by several groups worldwide[15-23]. It should be noted that the agreement between the rates used by these many groups was not easy to obtain. It resulted from extensive theoretical and experimental work. It should also be noted that the meetings which originated the present series of conferences (IWCE) were mainly aimed at assessing the early discrepancies among groups and, hopefully, at reaching a consensus.

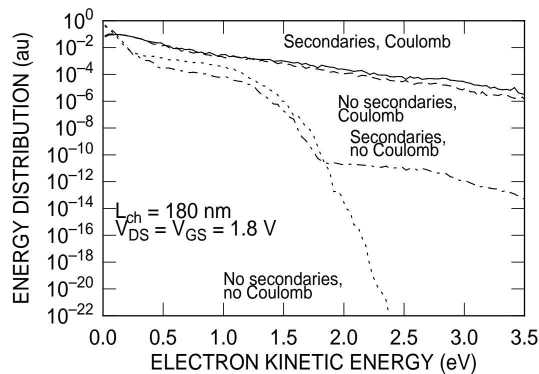

Figure 4: Calculated electron energy distributions in the channel of a 180 nm-long nMOSFET. Interparticle Coulomb interactions are responsible for the dramatic 'thermalization' of the distribution. Even sophisticated effects such as Bude's 'ionization feedback' (secondary ionization) are dwarfed buy electron-electron scattering.

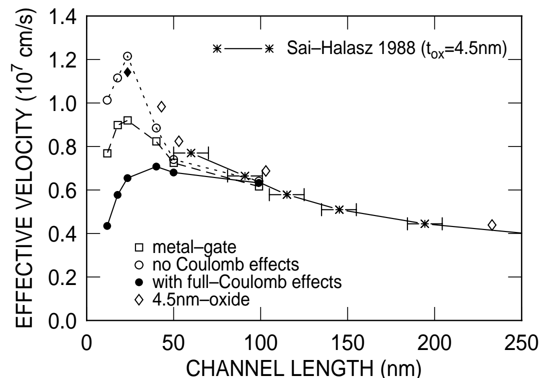

Figure 5: Calculated electron 'effective velocity' in the channel (namely: transconductance divided by gate capacitance) accounting for long-range Coulomb interactions among electrons in the channel and those in the source/drain regions and in the poly-silicon gate. Thermalization of the distribution function due to the former interactions and direct momentum losses via gate Coulomb drag result in a reduced performance at channel lengths shorter than approximately 40 nm.

Finally, MC simulations have opened up the possibility of studying the role of the most important interaction in nature, namely, Coulomb interactions. At first, the 'thermalizing' effect of electron-electron interactions in the channel of nMOSFETs was studied with the goal of understanding why some processes were observed to occur even below their expected 'threshold'.[24] Typically, substrate currents (caused by pair production across the gap) were observed in small devices even with an applied source-to-drain bias below 1.1 V (*i.e.*, the indirect gap of Si). Also, injection into the oxide was observed at biases below the Si-SiO_2 barrier height. Figure 4 shows the dramatic effect of the short-range electron-electron interactions in altering the energy distribution function in the channel: Sophisticated effects such as the 'secondary pair production' studied by Bude[25] appear to be totally negligible compared to the strength of these interparticle Coulomb processes.[26] Sadly, as we have said above, none of these results has had much of an impact on technology. Simple electrostatic simulations provided the answer to the most important question ("Which doping profile should we use to be able to turn the device off?"). Device performance came 'naturally' from its reduced dimensions, no matter whether or not we understood why.

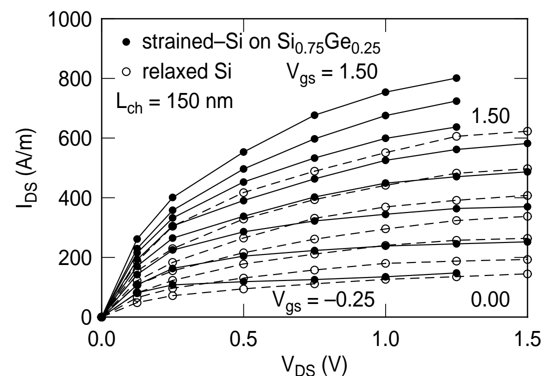

Figure 6: Calculated current-voltage characteristics of a 150 nm-long nMOSFET on bulk Si (open symbols, dashed lines) and on biaxially strained Si on $Si_{0.75}Ge_{0.25}$ (solid symbols, solid lines). Despite an observed 100% mobility boost, the strained-Si device exhibits only an approximately 30% performance advantage. Gate bias parametrizes the curves: From 0.0 to 1.5 V in steps of 0.25 V for the relaxed-Si device, from -0.25 to 1.5 V in steps of 0.25 V for the strained-Si device.

4. Present and future technolgy

The very same interparticle Coulomb interactions considered above may be responsible for a puzzling behavior observed in sub 100 nm nMOSFETs: Their unexpected poor performance. Noted first by the MIT group[27], this puzzling trend was attributed by them to increased scattering with the rough Si-SiO_2 surface as the confining field increases with shrinking oxide thickness. An alternative explanation may cause the effect illustrated in Fig. 5: High-density electron gases in the source, drain, and (poly-silicon) gate of an ULSI nMOSFETs are separated by a dis-

tance smaller than any screening-length in the system. Electrons in the channel may lose momentum to electrons in the gate via a Coulomb-drag process, while electrons in the drain may excite plasmons in the drain, the resulting 'thermalization' of the electron energy distribution in the channel causing, in turn, enhanced momentum-dissipating scattering with phonons, impurities, etc. The figure illustrates the effective electron velocity in the channel (*i.e.*, transconductance divided by gate capacitance) as a function of (metallurgical) channel length when accounting for all Coulomb processes, when suppressing the channel-gate drag (labeled 'metal gate'), and when supressing also all remaining Coulomb scattering. Note that below 40 nm or so the performance begins to drop as the source/drain and channel/gate distances decrease.[28]

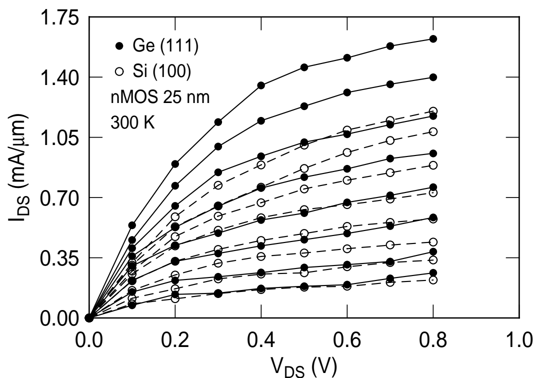

Figure 7: As in Fig. 6, but comparing a 25 nm-long bulk (100)-Si with a bulk (111)-Ge nMOSFET. Once more, the 3X mobility advantage of Ge over Si is largely negated – Ge outperforming Si by about 30% – at the small length scales of present interest. Gate bias ranges from 0.0 to 0.7 V (in steps of 0.1V) for both devices. No corrections has been applied for a difference in threshold voltages.

This last observation brings us to the present and near-future. Because of increasing technological difficulties, diminishing returns and physical limits, 'scaling' is being perceived – justifiably or not – as a thing of the past. Performance should be sought by looking for alternatives to evolutionary down-scaling. MC simulations – with their ability to handle physically accurate models – can finally help us, since the vast number and variety of possible alternatives requires some early guidance. Simpler simulation models simply cannot handle transport in other semiconductors and crystal orientations, strained materials and alternative gate dielectrics, ballistic or quasi-ballistic transport. Thus, MC simulations have recently been applied to 'real world' problems: Transport in strained Si, showing, as illustrated in Fig. 6, that a significant performance gain (up to 40%) can be obtained, the boost of the on-current dropping only marginally as the channel length shrinks down to 20 nm. Figure 7 illustrates a similar situation in which a Ge nMOSFET (fabricated on the (111) surface) outperforms a Si (100) nMOSFET by almost 30%. Even allowing for intrinsic 'calibration' uncertainties in the MC simulations (deformation potentials and scattering rates for many semiconductors are not as well known as for electrons in Si), this information provides some guidance as to which semiconductors, strain configuration, crystal orientation, device design (*i.e.*, planar double-gate, FINFET, ground-plane devices, fully-depleted SOI, etc.) should be considered with the highest priority.

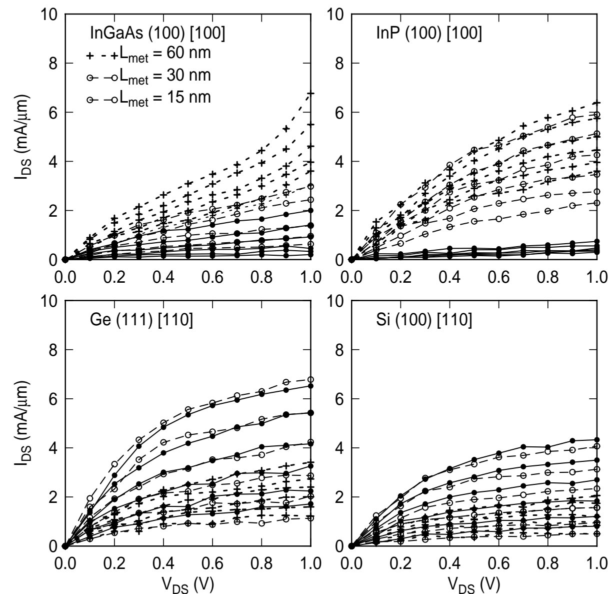

Figure 8: Calculated current-voltage characteristics of nMOSFETs of 60, 30, and 15 nm metallurgical channel lengths fabricated on (100) $InGa_{0.47}As_{0.53}$, InP, Si and (111) Ge. Channels are oriented along the [100] directions in the case of the In-based FETs, along the [110] direction for Si and Ge. As the quasi-ballistic regime is approached, the small density-of-states effective mass results in the inability of the high-speed materials to carry current.

These activities leave us with a few basic considerations which, eventually, acquire a 'fundamental' interest. Figures 6 and 7 show that even when boosting the low-field mobility by 100% (as for strained Si on SiGe) or even a factor of 3 or more (as when moving from Si to Ge), we should not expect a similarly large performance boost in small devices. Simply put, the low-field mobility, as we had emphsized above, is not

a useful concept any longer. Perhaps, the following results, once more obtained by full-band MC simulations, will illustrate the point more dramatically.

As shown in Fig. 8, scaled MOSFETs with 60, 30, and 15 nm effective channel lengths using Si, Ge, InP and $In_{0.53}Ga_{0.47}As$ have been simulated. While at (relatively) large channel lengths the small transport mass of III-V compound semiconductors gives them some performance advantage, at the shortest channel-length their small density-of-states effective mass makes them 'choke' (too few 'channels' – *à la* Landauer - to carry current). Ge appears superior, but only on the chosen (111) surface orientations and ignoring band-to-band leakage issues. (To be pedantic, (111)-Ge is to be preferred in the scattering-dominated limit, (110)-Ge in the ballistic limit... but we shall ignore these tiny details.) Note that the self-consistency between MC and the Poisson equation is a crucial ingredient in this analysis. Any model which bypasses this self-consistent step is bound to mistreat the basic physics of electron transport at small length scales. To put it in more general terms, in the scattering-dominated regime we prefer materials with a small conductivity effective mass – in order to boost carrier velocity – and with a small density-of-states effective mass – in order to reduce scattering (through a reduction of the density of available final states after collision). In the ballistic regime, instead, we still require a small conductivity effective mass, but we now require a *large* density-of-states effective mass, in order to have many Landauer channels open for conduction. This latter requirement is actually the dominant one. Ge is indeed the 'winner', as its small conductivity mass is accompanied by a large density of states resulting from the many valleys entering the picture. Whether the ballistic limit will ever be reached remains an open question (in view of the strength and unavoidable presence of interparticle Coulomb interactions we are inclined to believe that ballistic transport is a 'pipedream'). The fact remains that what's good for mobility in near equilibrium, long-channel transport, is bad for performance in ultra-short devices. In our opinion the search for 'good materials' (where 'good' means 'with high ohmic mobility') should be carried out with an open mind: Mobility and performance depend on the strength of scattering. The fact that they appear to be correlated is a tautology, both being effects of the same cause. Yet, their correlation becomes increasingly weaker at increasingly short dimensions. In the limit of ballistic transport, where scattering ceases to matter, they may become 'anticorrelated'.

References

Elements of this work have been summarized in an extended abstract published by IEEE.

1. M. V. Fischetti and S. E. Laux, in *Proceedings of the European Solid State Device Research Conference (ESSDERC'96)*, G. Baccarani and M. Rudan Eds. (Éditions Frontières, 1996), p. 813.
2. R. Stratton, *Phys. Rev.* **126**, 2002 (1965); K. Blotekjaer, *IEEE Trans. Electron Devices* **ED-17**, 38 (1970); W. Hänsch, in *Proc. Two-Dimensional Systems: Physics and Devices* (Springer, Berlin, 1986), p. 296. M. Rudan and F. Odeh, *COMPEL* **5**, 149 (1986).
3. T. Kurosawa, in *Proc. Int. Conf. Phys. Semicond. Kyoto, J. Phys. Soc. Jpn.* Suppl. **21**, 424 (1966).
4. W. Fawcett, D. A. Boardman, and S. Swain, *J. Phys. Chem. Solids* **31**, 1963 (1970); W. Fawcett, C. Hilsum, and H. D. Rees, *Solid State Comm.* **7**, 1257 (1969).
5. P. J. Price, in *Semiconductors and Semimetals* (Academic Press, New York, 1979), p. 249.
6. C. Jacoboni and L. Reggiani, *Rev. Mod. Phys.* **55**, 645 (1983) and references therein.
7. H. Shichijo and K. Hess, *Phys. Rev. B* **23**, 4197 (1981).
8. J. Y. Tang and K. Hess, *J. Appl. Phys.* **54**, 5139 (1983).
9. K. Masaki *et al.*, *Jpn. J. Appl. Phys.* **28**, 1856 (1989).
10. K. Yokoyama and K. Hess, *Phys. Rev. B* **33**, 5595 (1986).
11. S.E. Laux and M.V. Fischetti, *IEEE Electron Device Lett.* **EDL-9**, 467 (1988).
12. M. V. Fischetti and S. E. Laux, *Phys. Rev. B* **38**, 9721 (1988).
13. G. Sai-Halasz *et al.*, *IEEE Electron Device Lett.* **EDL-9**, 464 (1988).
14. M. V. Fischetti and S. E. Laux, *IEEE Trans. Electron Devices* **ED-38**, 650 (1991).
15. T. Kunikiyo *et al.*, *J. Appl. Phys.* **75**, 297 (1994).
16. Y. Kamakura *et al.*, *J. Appl. Phys.* **75**, 3500 (1994).
17. N. Sano and A. Yoshii, *J. Appl. Phys.* **75**, 5102 (1994).
18. H. Shichijo *et al.*, in *Monte Carlo Device Simulation: Full Band and Beyond*, K. Hess ed., Kluwer Academic (Norwell, MA, 1991), p. 285.
19. P. D. Yoder, V. D. Natoli, and R. M. Martin, *J. Appl. Phys.* **73**, 4378 (1993).
20. J. Kolnik *et al.*, *J. Appl. Phys.* **76**, 3542 (1994).
21. E. A. Cartier, M. V. Fischetti, E. A. Eklund, and F. R. McFeely, *Appl. Phys. Lett.* **62**, 3339 (1993).
22. M. V. Fischetti and J. M. Higman, in *Monte Carlo Device Simulation: Full Band and Beyond*, K. Hess ed., Kluwer Academic (Norwell, MA, 1991), p. 123.
23. J. D. Bude and M. Mastrapasqua, *IEEE Electron Device Lett.* **16**, 439 (1995).
24. M. V. Fischetti and S. E. Laux, *IEDM Tech. Dig.*, 305 (1995).
25. J. D. Bude, *Sym. on VLSI Tech. Tech. Dig.*, 101 (1995).
26. M. V. Fischetti, S. E. Laux, and E. Crabbé, *J. Appl. Phys.* **78**, 1058 (1995).
27. A. Lochtefeld and D. A. Antoniadis, *IEEE Electron Device Lett.* **EDL 22**, 95 (2001).
28. M. V. Fischetti and S. E. Laux, *J. Appl. Phys.* **89**, 1205 (2001); M. V. Fischetti, *J. Appl. Phys.* **89**, 1232 (2001).