Physical Models for Semiconductor Device Simulation

Andreas Schenk

Swiss Federal Institute of Technology, Integrated Systems Laboratory, Gloriastrasse 35, CH-8092 Zürich, Switzerland

Summary: Device simulation has two main purposes: to understand and to depict the physical processes in the interior of a device, and to make reliable predictions of the behavior of the next device generation. Towards these goals, the quality of the implemented physical models is decisive, forcing heuristic fit models to be replaced by "first-principle"-based models. Since transport schemes using moments of the Boltzmann equation are still dominant within the simulation community, the challenge in developing models is given by a clever combination of sound physics with functional simplicity. The extent to which this has been possible will be demonstrated by means of two examples: band gap narrowing and band-to-band tunneling. We point out the limitations set by the poor knowledge of certain parameters, by the breakdown of common approximations, and by unresolved principle problems.

1 Introduction

The description of transport in semiconductor devices requires models for both the interaction processes and the embedding system. These models will have different form depending on the transport equations used. However, every transport scheme demands expressions for the scattering of charge carriers with elementary excitations of the crystal as well as with each other, with impurities, device boundaries or interior interfaces, and models of all generation-recombination processes. The embedding system is given by material parameters, e.g. band gap, intrinsic density, and by external quantities like the doping and defect profiles, boundaries, and others.

As there is an upward averaging of the transport equations by considering mean values of the current density operator or the classical momentum, respectively, a corresponding "philosophy" should be applied to the physical models for the transport coefficients arising in these equations. The advantage is obvious: derived quantities and parameters can be traced back to their physical origin, and the performance of the device can be understood on physical grounds. However, because the coarsening procedure itself is difficult, and only in very rare

cases the "first-principle"-based models agree satisfactory with the experimental data, often the preferred way is just the other way around. Simple analytical functions and various adjustable parameters guarantee the required fit and short computation times. This even can lead to the extreme case that a completely wrong model shows a "perfect" agreement with measurements. From these reasons, physical models should be at least consistent and transparent, which e.g. means that changes in one model as response to some observed properties of the device must be inevitably followed by corresponding changes in other models, and that all parameters should allow a physical interpretation. On the other hand, because it is generally necessary to apply certain models to the extraction of experimental data, it seems to be most reliable to use a consistent set of measured parameters from one group instead of mixing the results of different authors without comparing them thoroughly. Although empirical formulas have been applied successfully in numerical device simulation, the trend to miniaturization, vertical integration and higher doping forces the models to become more microscopic and therefore more physics-based. For instance, tunneling phenomena like band-to-band tunneling, defect-assisted tunneling or tunneling across potential- and oxide barriers, require the incorporation of quantum-mechanical models also into the classical equations by means of generation rates or proper boundary conditions.

In this article we will focus on the class of models necessary for all transport schemes that are based on taking moments of the Boltzmann transport equation (BTE). Full-band Monte Carlo (MC) simulations of the BTE are becoming increasingly important, but their application in an industrial environment is still inhibited by very long computation times. To date, the so-called energy balance (EB) model is widely used in simulating sub-micron devices, which extends the familiar drift-diffusion (DD) model by conservation laws for the energy density of the sub-systems (electrons, holes, lattice). In the next section we give a brief summary of physical models needed for the EB equations. In order to keep the advantage of (relatively) short computation times, it is crucial to derive the models in a simple analytical form avoiding numerical integrations or iterations, but at the same time saving as much physical information as possible. In Sections 3 and 4 we will illustrate the extent to which this has been possible by means of two examples: band gap narrowing and band-to-band tunneling. We restrict ourselves to silicon as the technologically most important semiconductor. The discussion in Section 5 will concentrate on the limitations raised by the poor knowledge of certain parameters, by the breakdown of common approximations, and by some principle problems that have been not resolved yet.

2 Physical Models for Energy Balance Equations

Originally, Stratton [1] applied the first three moments of the BTE to formulate equations of flow for charge carriers and energy in semiconductors. Later, Bløtekjær described a model with zeroeth through third order, thus including also the energy flux density as an unknown [2]. The schemes of both authors differed in the treatment of the relaxation time. More recent work by Bringer *et al.* [3] and Azoff [4, 5] has also included the third-order moment. Based on the work of Bløtekjær, Rudan and Odeh [6] gave a detailed derivation of the hydrodynamic (HD) transport model and also proposed a discretization technique for the steady-state case, including appropriate boundary conditions. A discretization of the full time-dependent HD model was presented by Forghieri [7]. By neglecting a term which is quadratic in the current density (the so-called *convective* term) and the drift part of the kinetic energy as compared with the carrier mean thermal energy $\frac{3}{2}k_B T_c$, Cook and Frey [8] proposed a simplified carrier and energy transport model which has come to be known as the energy balance (EB) model (see also Bløtekjær [9]). Fukuma *et al.* [10] and Meinerzhagen [11, 12] have subsequently applied extensions of this model.

Fig. 1 shows a sketch of the EB model for the electronic sub-system. The first three BTE moments were closed with the phenomenological constitutive relation $\boldsymbol{Q}_n = -\kappa_n \nabla_{\boldsymbol{r}} T_n$ between the conductive heat flow vector $\boldsymbol{Q}_n$, the electron thermal conductivity κ_n and the electron temperature T_n according to Bløtekjær [9]. The choice of this closure has recently been cast in doubt by Stettler *et al.* [13]. Further symbols, not explained explicitly in Fig. 1, have the following meaning: $\boldsymbol{j}_n$ – electric current density of electrons, $\boldsymbol{F} = -\nabla_{\boldsymbol{r}}\psi$ – electric field, $\sigma_n = q\mu_n n$ – electric conductivity, $\boldsymbol{S}_n$ – electron energy flux density, $G - R$ – net generation rate, T_L – lattice temperature, q – elementary charge, k_B – Boltzmann constant. The keywords in solid frames represent the models that have to be provided for the physical description of transport coefficients (upper part) and device phenomena (lower part). The latter were attached to different regions of an erasable programmable read only memory (EPROM), which can be regarded as a metal-oxide-semiconductor field effect transistor (MOSFET) containing a floating gate for charge storage. Besides various mechanisms of charge transport across interfaces and the non-ideal behavior of metal-semiconductor contacts, we emphasized the so-called MOSFET degradation, which is one of the most severe problems in modern microelectronics. It is produced by a long-term shift of the threshold voltage due to charge trapping and eventual defect generation at the interface caused by hot carriers in the channel.

We will not attempt here to give an overview of all the models cited in Fig. 1. Taking only the mobility in silicon, there are dozens of published models (or even hundreds when counting all the slight modifications). Excellent summaries of the most widely used models can be found in the work of Selberherr [14, 15]

and Baccarani *et al.* [16].

$$\frac{3}{2}k_B\frac{\partial}{\partial t}(n\,T_n)+\nabla_r\cdot S_n = j_n\cdot F-\frac{3}{2}k_BT_n(G-R)-\frac{3}{2}k_B\,n\,\frac{T_n-T_L}{\tau_{E,n}}$$

energy relaxation time

$$\frac{\partial}{\partial t}n-\frac{1}{q}\nabla_r\cdot j_n = G-R$$

intrinsic density
gap (T)

SRH lifetimes (T, N_{dop})
defect-assisted tunneling
Auger (b2b, defect-assisted)
impact ionization
band2band tunneling
optical generation
alpha particles

thermal conductivity

$$S_n = -\kappa_n\nabla_r T_n-\frac{5}{2}\frac{k_BT_n}{q}j_n$$

band gap narrowing

$$j_n=\sigma_n\left[-\nabla_r\left(\psi+\frac{\Delta E_g}{2q}\right)+\frac{k_BT_n}{q}\nabla_r\ln(n)+\nabla_r\frac{k_BT_n}{q}\right]$$

mobility (bulk, channel)

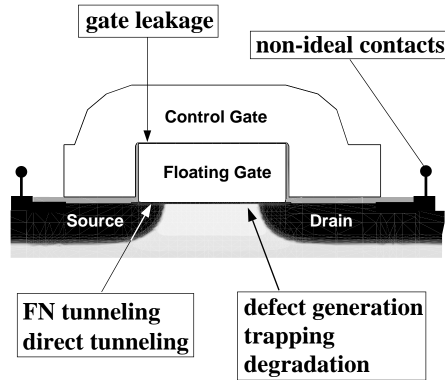

Figure 1 Schematic of important transport parameters and related physical models for energy balance equations (upper part). Important device phenomena in different regions of an EPROM (lower part).

3 Band Gap Narrowing

Heavy doping of certain device regions results in a shrunk band gap, thus the effective intrinsic density can increase by orders of magnitude there. Band gap narrowing (BGN) has a strong impact on device operation, in particular on the current gain of bipolar transistors as shown in Fig. 2.

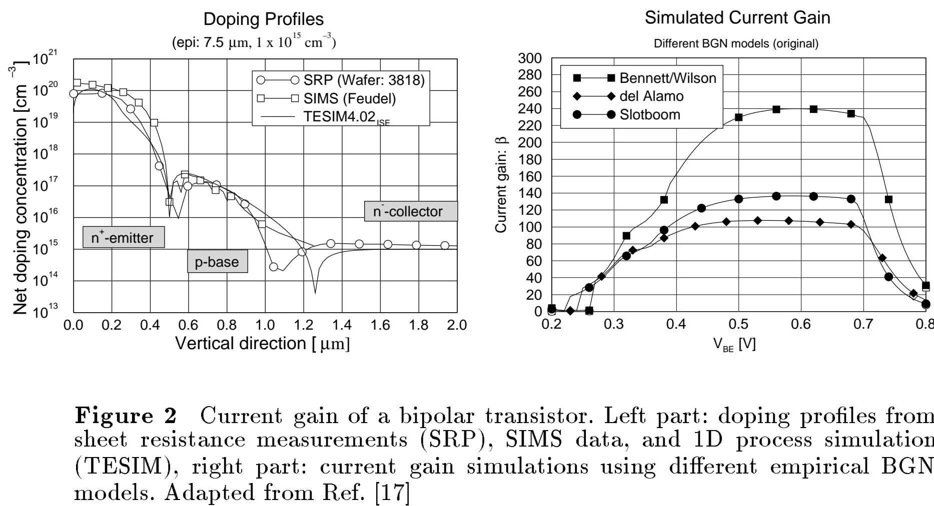

Figure 2 Current gain of a bipolar transistor. Left part: doping profiles from sheet resistance measurements (SRP), SIMS data, and 1D process simulation (TESIM), right part: current gain simulations using different empirical BGN models. Adapted from Ref. [17]

Application of different empirical BGN models has a tremendous effect on the simulated current gain, which is for an npn-transistor proportional to $n_{base}/p_{em} \sim \exp(-\Delta E_g/k_B T)$, where ΔE_g denotes the BGN in the emitter [18].

From a theoretical point of view BGN results from many-body effects and from potential fluctuations caused by the disordered impurities. In addition, electron-hole plasmas can modify the electron-impurity interaction.

3.1 Doping-Induced Rigid Shift of the Band Edges

The main contribution to BGN is caused by many-body effects. Low-concentration measurements of the band gap correspond to calculated effective one-particle band structures of an *ideal*, intrinsic semiconductor. At high doping concentrations or/and under high excitation the electron-impurity interaction and the strong correlation between the carriers result in qualitative changes of the energy spectrum. The difference between ideal and interacting dispersion, the so-called quasi-particle shift, is equivalent to a rigid shift of the band edges. Here and in the following paragraph, we disregard a possible electron-hole plasma.

Mahan [19] used a variational method to calculate the self-energy from electron-impurity scattering assuming that the donors are distributed on a *fcc*-lattice. Berggren and Sernelius [20] derived these contributions from second-order perturbation theory for a *random* system of impurities. They found the same order of magnitude for the impurity-induced shift of the valence band in n-type silicon. Logan and Egley [21] calculated the screening parameter for highly doped p-type silicon using band dispersions based on a 6×6 Hamiltonian and applying the full random phase approximation (RPA) expression of the dielectric function (dispersive screening). The screened potential of electron-electron and electron-impurity interaction then was used to determine the self-energies in a finite-temperature dielectric-response formalism: a statically screened Hartree-Fock exchange potential and the second-order perturbation term of electron-impurity interaction. The resulting BGN in p-type silicon agreed quantitatively with the photoluminescence data by Wagner [22] both at 20 K and 300 K. Following the lines of Mahan, i.e. using the many-body technique (ground state, no band tails, high density regime) but aiming at simple results usable in device simulation, Jain and Roulston [23] derived the formula

$$\frac{\Delta E_g}{R_{maj}} = 1.83\,\frac{\Lambda}{n_b^{1/3}}\frac{1}{r_s} + \frac{0.95}{r_s^{3/4}} + \left(1 + \frac{R_{min}}{R_{maj}}\right)\frac{1.57}{n_b^{1/3}\,r_s^{3/2}}\;. \qquad (3.1)$$

Here $R_{maj,min}$ denote the effective Rydberg energy of the majority or minority band, respectively, r_s is the density parameter defined from $4\pi(r_s\,a_B)^3/3 = 1/N_{dop}$, n_b is the number of valleys for n-type silicon or the band multiplicity for p-type silicon, and $\Lambda = 1$ (n-type) or $\Lambda = 0.75$ (p-type), respectively. Eq. (3.1) contains the four major contributions to BGN; (1) the shift of the majority band edge due to exchange interaction; (2) the shift of the minority band edge due to electron-hole interaction; (3) the shift of the majority band edge due to carrier-impurity interactions; and (4) the shift of the minority band edge due to carrier-impurity interactions. If potential fluctuations due to local changes in the density of impurities are neglected in many-body calculations, the band gap becomes a local quantity. Selloni and Pantelides [24, 25] have calculated the real density of states (DOS) of heavily doped n-type silicon applying the jellium model for the electron-electron interactions (Inkson [26]), linear-response theory for electron-impurity interaction with the impurities assumed to be on an ordered lattice, and in a third step, introduced disorder assuming fluctuations in the impurity concentration about its average value. They found that if multivalley scattering is neglected, as in Ref. [20], electron-impurity interactions cause only a negligible shift in the conduction band edge and do not change the free-electron character of the DOS near the band edge. In contrast, including intervalley scattering caused a shift comparable to that produced by electron-electron interactions and even modified the DOS near the band edge. With multivalley scattering the calculated zero-phonon photoluminescence spectra and their high-energy edges

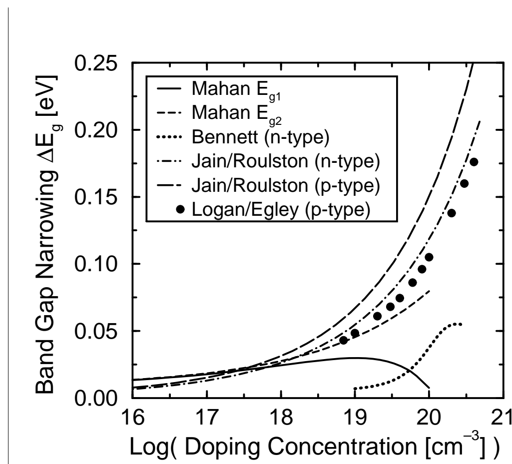

Figure 3 Comparison of different theoretical BGN models based on calculations of the rigid shifts of the band edges. References are given in the text. The points from Logan and Egley refer to 20 K.

in n-type silicon showed a remarkable agreement with experimental results. The disorder-model of Selloni and Pantelides only smoothed out the sharp features of their DOS, whereas Berggren and Sernelius [27] found that the model of complete disorder produces a large shift due to electron-impurity interactions even *without* the inclusion of multivalley scattering.

3.2 Band Tails

Besides many-body calculations under the assumption of certain dopant configurations, the band tailing has been studied by many authors ignoring exchange-correlation effects. The pioneering semi-classical treatment by Kane [28] overestimated the number of states deep in the tails, but becomes accurate with increasing energy. Halperin and Lax [29, 30] corrected the failure of the semi-classical method in the deep band tail region by adding the kinetic energy of localization [31]. Their tail model fails for energies near the unperturbed band edge. Sa-yakanit [32, 33, 34] used the Feynman path integral method, which yields an exact expression of the average over the random potential. He obtained the semi-classical result and the result of Halperin and Lax as limiting cases for small times (high energy) and large times (low energies), respectively. The calculated DOS curves in Kane's and Sa-yakanit's models intersect at a certain energy below the unperturbed band edge. Sa-yakanit proposed to use his model up to this intersection point, but Kane's model for higher energies. All approximations leading to Gaussian statistics for the potential fluctuations require

high doping levels (high-density limit) [31]. Van Mieghem *et al.* [35] modeled the effect of band tailing for a non-interacting system as an equivalent downward shift of the Fermi level. After various approximations (high-density limit, screening calculated with unperturbed DOS) an analytical expression was presented which actually can be simplified to the following form:

$$\Delta E_F = -\frac{0.41\, R_{\mathit{eff}}}{r_s^{7/8}} \;, \tag{3.2}$$

where R_{eff} is the effective Rydberg energy. ΔE_F was implemented into a device simulator as additional BGN effect. For silicon the shift amounts about 15 meV–35 meV in the density range $3 \times 10^{19}\,\mathrm{cm}^{-3} - 1 \times 10^{21}\,\mathrm{cm}^{-3}$. Hence, for emitter dopings larger than $10^{20}\,\mathrm{cm}^{-3}$ the current gain of bipolar transistors becomes strongly reduced.

3.3 Plasma-Induced Shift of the Band Edges

High injection levels occur frequently in silicon devices even in lightly doped regions as in the base and collector of bipolar transistors, in photoconductive switches or in concentrator solar cells under conditions of strong optical excitation. In heavily doped regions the presence of an electron-hole (e-h) plasma reduces the dopant-carrier interaction with a corresponding reduction of the BGN because of the screening by the excess carriers. Abram *et al.* [36] used the plasmon pole approximation (Hedin 1969 [37]) including a q^4 term in the plasmon dispersion relation which gives an accuracy close to the Lindhard dielectric function [37]. Lowney [38] generalized the zero temperature theory of Abram *et al.* to room temperature. A fit to his results was proposed by Shaheed *et al.* [39]: $\Delta E_{g,e-h} = 3.81 \times 10^{-6}\, n^{0.38}$ meV with the plasma density n in cm^{-3}. This model was implemented into a device simulator within an iteration loop assuming a certain value for the band offset X ($\Delta E_c = X\, \Delta E_{g,e-h}$, $\Delta E_v = (1 - X)\, \Delta E_{g,e-h}$). Since in a bipolar transistor the plasma-induced BGN is very nonuniform both across the emitter-base junction and throughout the base, it affects the barrier for minority carriers in the junction and the effective drift field in the base. Both effects depend on the value of the band offset X. It was found that $X = 45\%$ gave the best fit to measured DC-current gain characteristics in the temperature range 77 K – 310 K.

Zimmermann [40] derived an RPA expression for the quasi-particle shift in a symmetric e-h plasma (important for laser modeling and the on-state of power diodes), which is valid at all temperatures:

$$\Delta\mu_c^{eh}(r_s, \mathcal{T}) = -\frac{3.24\, R_{\mathit{eff}}}{r_s^{3/4}\,(1 + 0.0478\, r_s^3\, \mathcal{T}^2)^{1/4}} \;. \tag{3.3}$$

Here, the density parameter r_s is calculated with the plasma density and $\mathcal{T} = k_B T / R_{eff}$. For R_{eff} the reduced effective mass has to be used. Eq. (3.3) reproduces both the correct limit for $T = 0$ and the limit of the Debye shift at high temperatures [40]. The case of an asymmetrical plasma (extrinsic semiconductor with plasma excitation) has been worked out by Roesler *et al.* [41]. They considered the ions as dynamical quantities with infinite mass assuming complete disorder. Their final RPA result is restricted to $T = 0\,\mathrm{K}$, and a fit formula was only presented for GaP.

3.4 BGN Models in Device Simulation

From the above review of theoretical work on BGN we may conclude that up to now there is no satisfactory unified theory of many-body effects and disorder for silicon at finite temperatures. Moreover, for device simulation purposes some analytical expression $\Delta E_g(n, p, N_A, N_D, T)$ similar to Eq. (3.3) is desirable, extended to the general case. However, even provided the existence of such a model, its application would lead to a complicated iteration $n = n[\Delta E_g(n, ...)]$. Hence, in todays device simulators BGN is mostly accounted for by oversimplified models of the form (del Alamo *et al.* [42])

$$\Delta E_g^{app} = 18.7 \times 10^{-3} \ln \left(\frac{N_D}{7 \times 10^{17}} \right) \text{ eV} \tag{3.4}$$

for $N_D \geq 7 \times 10^{17}\,\mathrm{cm}^{-3}$, and zero at lower doping levels. Eq. (3.4) was derived for n-type silicon by studying simultaneously pn-product, diffusion length and lifetime of minority carriers in bipolar transistors. Because this electrical method yields a gap narrowing, which contains the effect of everything not considered elsewhere in an analytical or numerical transistor model, del Alamo *et al.* claimed only to measure an "*apparent gap narrowing*" ΔE_g^{app}. It is not surprising that the model (3.4) from experience [17] is the most successful BGN model for the simulation of current gains in bipolar transistors.

Let us come back to the question how far a model like Eq. (3.4) reflects the complicated physics of BGN in a device simulation. First, it is implicitly assumed that both band edges shift by $\Delta E_g(N_{imp})/2$, i.e. it is not distinguished between conduction and valence band edge, and the BGN is considered to be a function of the total doping concentration N_{imp} only. Furthermore, the different physical effects of electron-electron interaction, carrier-impurity interaction, electron-hole interaction, and random potential fluctuations, as well as their respective contribution to the total BGN are not separated from each other. Therefore, such models yield e.g. equal gap narrowing both in neutral and depleted regions of a device. The last problem becomes particularly severe for the influence of BGN on the electron-hole pair generation in heavily doped and depleted regions, as in the case of band-to-band tunneling or impact ionization.

4 Band-to-Band Tunneling

As a result of device scaling very shallow junctions with high doping levels and steep gradients came into use in recent years. Leakage currents due to defect-assisted tunneling (DAT) and band-to-band tunneling (BBT) were observed in the emitter-base junction of bipolar transistors [43, 44], in critical interface regions of trench transistor DRAM cells [45, 46], and at the drain edge of MOSFETs [47, 48]. On the other hand, tunnel generation is intentionally used for (band-to-band ?) tunneling-induced substrate hot-electron injection (BBISHE) in non-volatile memories [49].

4.1 Microscopic Theory

BBT in silicon is phonon-assisted, which was experimentally shown already in the early sixties. By measuring the derivative of the conductance in silicon Esaki diodes at 4.2 K, Chynoweth *et al.* [50] could reveal twelve phonon and phonon-combination energies, which agreed well with results of neutron scattering studies. Later Logan and Chynoweth [51] succeeded to decompose the tunneling current into a phonon-unassisted current (excess current), a TA phonon-assisted current and a TO phonon-assisted component. The first calculations of phonon-assisted BBT were presented by Keldysh [52, 53] and, independently, by Price and Radcliffe [54] in 1958. Keldysh used second order perturbation theory, Houston-type wave functions [55], and the saddle-point method, whereas Price and Radcliffe applied the Wentzel-Kramers-Brillioun (WKB) approximation. Keldysh's result was improved by Kane [56]. In all these papers the problem was solved by determining the transmission coefficient of an electron striking the junction barrier (illustrated in Fig. 4) and then calculating the current by the number of generated carriers. The connection between transmission probability and current density is unnecessary, if a macroscopic quantity is calculated which directly determines the BBT current. This was done for the first time by Enderlein and Peuker [57] for a direct semiconductor. They used a Kubo formula [58] for the differential conductivity of the crystal in a strong electric field. The band-to-band part of the conductivity is determined by the off-diagonal elements of the one-particle density matrix and arises, because the electrons change their place when penetrating the barrier. The application of this method to the phonon-assisted BBT in silicon was presented by Schenk [59]. There, electron-phonon collisions were taken into account as a momentum source for the tunneling electrons, and the crystal Hamiltonian was treated in effective mass approximation (EMA), fully accounting for the anisotropy of the six conduction band valleys. Since the size of the direct gap of silicon does not change dramatically within the first Brillioun zone, transitions via both intermediate states opposite to the band extrema of the indirect gap are equally important. In the course of the

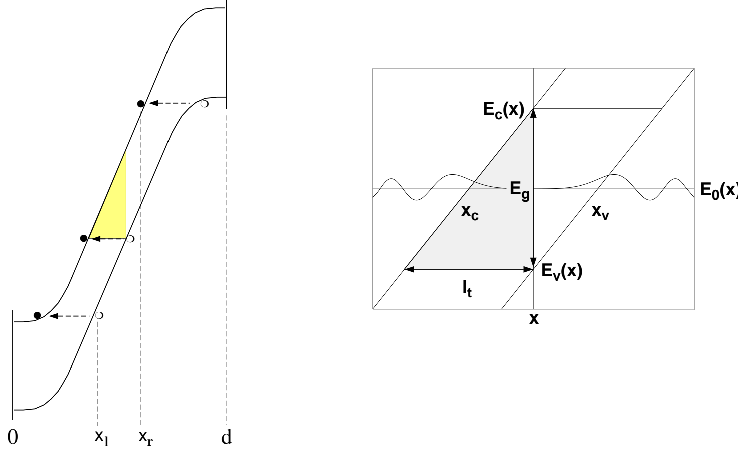

Figure 4 Electron-hole pair generation by band-to-band tunneling. Left part: Boundaries x_r for electron generation and x_l for hole generation, respectively. The marked area is the tunnel barrier. Right part: Locality at x is suggested by the maximum overlap of the Airy functions within the gap. E_0 denotes the transition energy and l_t the tunneling length.

derivation, a combination of three-particle Green's functions occurs, which must be decoupled by RPA. After utilizing the resulting delta functions, a sevenfold nontrivial integral remains to be evaluated analytically. Explicit expressions for the matrix elements (two momentum and two electron-phonon) have to be found. The final form of the phonon-assisted BBT rate suitable for device simulation reads

$$R_t = \frac{2.47\times 10^{21}}{\mathrm{cm^3 s}} F^2 \sum_{\alpha=x,y,z} \frac{\sqrt{m_\perp^\alpha m_\parallel^\alpha}}{m_0} \left[f_B \,\mathrm{H}\left(x_\alpha^\mp\right) + (f_B+1)\,\mathrm{H}\left(x_\alpha^\pm\right)\right](f_v - f_c) \ , \tag{4.5}$$

with

$$\mathrm{H}(x) = \frac{\mathrm{Ai(x)}}{\mathrm{x^2}} + \frac{\mathrm{Ai'(x)}}{\mathrm{x}} + \mathrm{Ai_1(x)} \ , \qquad \mathrm{x^\pm} = (\mathrm{E_g} \pm \hbar\omega_0)\left(\frac{8\mu_\parallel^\alpha}{\mathrm{q^2\hbar^2F^2}}\right)^{1/3} \ , \tag{4.6}$$

$$\frac{1}{\mu_\parallel^\alpha} = \frac{m_l - (m_l - m_t)F_\alpha^2/F^2}{m_t m_l} + \frac{1}{m_v} \ , \quad m_\perp^\alpha = \frac{m_t m_l}{m_t - (m_t - m_l)F_\alpha^2/F^2} \ . \tag{4.7}$$

F is measured in $\mathrm{Vcm^{-1}}$, $m_{t,l,0}$ denote transverse, longitudinal, and rest mass, respectively, f_B is the phonon occupation number, $\hbar\omega_0$ a representative phonon energy, Ai' the derivative of the Airy function, and $\mathrm{Ai_1}$ its integral [60]. The

upper sign has to be applied in the case of reverse biased junctions (generation, $f_v > f_c$), whereas the lower sign holds for forward biased junctions (recombination, $f_v < f_c$) in order to account for energy conservation in the electron-phonon system [56]. The carrier distribution functions have to be evaluated at the transition energy $E_0(x)$ given by

$$E_0(x) = \frac{m_v\, E_v(x) + m_{\|}^{\alpha}\ [E_c(x) \pm \hbar\omega_0]}{m_{\|}^{\alpha} + m_v}\ . \tag{4.8}$$

The model (4.5) is not based on the WKB method and includes finite temperatures as well as the anisotropy of the conduction band valleys. Apart from some approximations to enable the complicated analytical integration, which are not very severe (e.g. neglection of phonon dispersion and taking $f_{c,v}$ at the local band extrema), two major uncertainties arise. The first originates from the momentum matrix element containing Bloch factors. The second concerns the average hole mass m_v. One has to expect that the strong electric field responsible for BBT leakage will remove the band degeneracy at $\boldsymbol{k} = 0$ and, therefore, will also change the effective hole mass to be used in the tunneling model. Because m_v enters the BBT rate in the exponent, a change within the range defined by average light and average heavy hole masses is followed by a large change of the rate. This is shown in Fig. 5. Using the heavy hole mass instead of the light hole

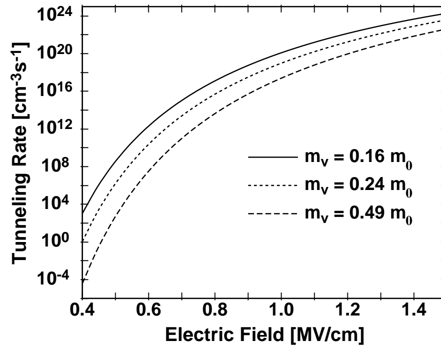

Figure 5 BBT rate in silicon for three values of the effective hole mass: $m_v = 0.16\, m_0$ (light hole mass) and $m_v = 0.49\, m_0$ (heavy hole mass) and $m_v = 0.24\, m_0$. The direction of the applied field is [111].

mass decreases the rate by seven orders of magnitude for $F = 4 \times 10^5\ \mathrm{Vcm}^{-1}$ and still by two orders of magnitude at $F = 1.5 \times 10^6\ \mathrm{Vcm}^{-1}$. Applying two-level perturbation theory yields just twice the value of a reduced hole mass built from the heavy and light hole masses, i.e. $0.24\, m_0$. Using this value gives the dotted curve in Fig. 5.

4.2 Evaluation of BBT Model Parameters

The above mentioned uncertain parameters of the model can be fixed by simulating a tunnel diode with precisely known doping profile. The 1D, vertical junction with given area and SIMS profile allows to fit both the matrix element in the prefactor and the value of the tunneling mass in the exponential by means of an Arrhenius plot. For this, the tunnel peak of the forward IV-characteristic and the reverse branch can be used, where BBT is the dominant mechanism. Fig. 6 represents a simulation of the tunnel diode described by Esaki and Miyahara in Ref. [61] with the device simulator DESSIS$_{-\mathrm{ISE}}$ [62]. The unknown doping profile was adjusted by "reverse modeling" assuming a Gaussian distribution and fitting the forward BBT current to the experimental tunnel peak. In the simulation, BBT is "switched off" sharply, when the diffusion potential of the diode becomes smaller than the band gap. In reality, a further increase and a subsequent smooth "switch-off" occurs as a consequence of the band tails. The question mark in Fig. 6 points to the voltage range between 0.1 V and 0.3 V with negative differential resistance, where band tails or DAT can hardly explain the strong current. It is obvious that the real DOS as well as the right BGN model have a large effect on the simulated tunnel current. In the simulation BGN was turned off, which was considered a better approximation in the depletion zone as compared to any empirical BGN model from quasi-neutral regions. However, the misfit of the peak maximum also indicates that BGN has a non-negligible influence on the position of the tunnel "hump".

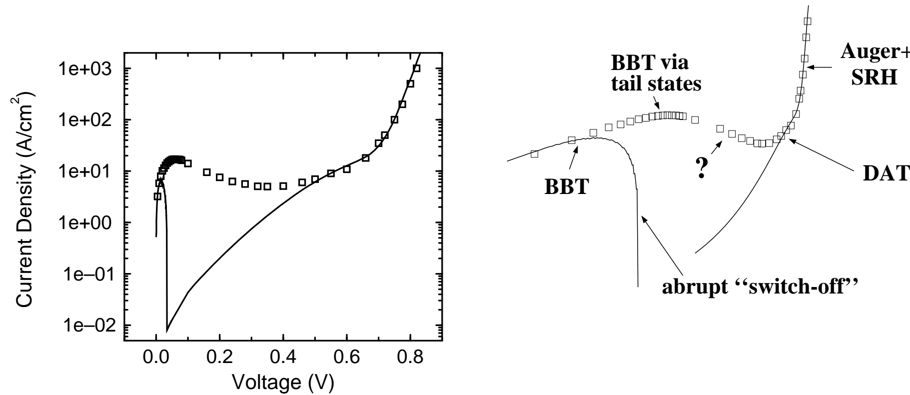

Figure 6 Simulation of the 298 K IV-characteristic of the silicon tunnel diode from Ref. [61]. Left part: Symbols – experimental data, solid curve – device simulation. Right part: Corresponding log-log plot with indication of the dominant recombination mechanisms.

4.3 Pseudo-local Formulation of the BBT Rate

In view of simulation of MOSFETs, artificial electron-hole pair generation directly at the oxide-silicon interface, where no final conduction band states are available but the normal electric field peaks, must be excluded. Therefore, a thin sheet with a thickness of half the BBT length does not contribute to the BBT rate in DESSIS$_{-\mathrm{ISE}}$. In this way a time-consuming searching procedure is avoided. Similarly, a corresponding thin sheet is excluded from the DAT domain. Fig. 7 illustrates the effect in the case of BBT for a gated diode with 10 V contact bias and 10 V gate voltage. The difference in the corresponding IV-characteristics

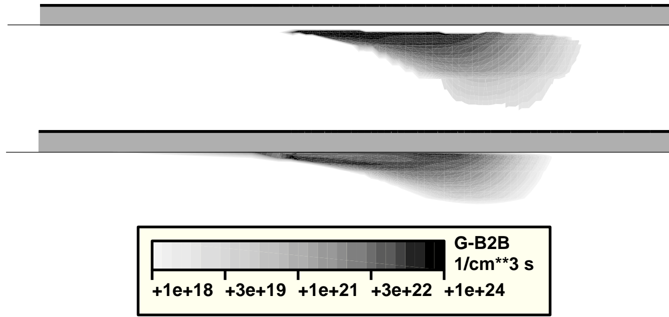

Figure 7 Distribution of the BBT rate beneath the gate oxide of a MOS-gated diode. Upper part: with exclusion of a thin sheet where no final conduction band states exist, lower part: without exclusion.

is shown in Fig. 8. If the rate is set to zero in the region of vanishing tunnel probability, the onset of the breakdown current is shifted to higher voltages, as demonstrated in the right part of Fig. 8. The same exclusion routine is applied to pn-junctions. As they become steeper, both the local generation models and the drift-diffusion approach itself start to fail. Simulating very narrow junctions with local DAT and BBT models, one can observe the following effect: At a certain reverse bias, DAT or BBT can generate such a large number of electron-hole pairs at the point of maximum field strength, that the np-product reaches the value of $n_{i,eff}^2$, and consequently, the rate will go into a pronounced minimum at this point. In the self-consistent solution of the transport equations this leads to a self-saturation of the generation current. In local versions of generation rates the spatial separation of the generated "carriers" (i.e. classical point charges moving along a trajectory) is not taken into account, which results in the above mentioned effect.

The question arises how a tunneling rate can be efficiently formulated in the picture, where the carriers appear or disappear as classical particles, i.e. at the

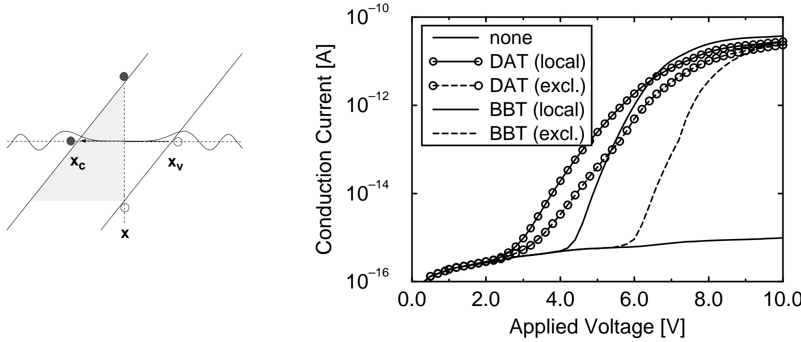

Figure 8 Left part: Nonlocal generation of electrons and holes by BBT at the classical turning points. Right part: BBT and DAT IV- characteristics of the gated diode in Fig. 7.

classical turning points. In the rigid quantum-mechanical derivation of the BBT rate in Subsection 4.1 locality is suggested by the sharp overlap of the envelope wave functions within the band gap. Hence, the prefactor in front of the “driving force” $f_c - f_v$ is determined with the local field strength at this point. By calculating the carrier distributions at the transition energy E_0 but not at the band edges $E_c(\boldsymbol{x})$ and $E_v(\boldsymbol{x})$, respectively, defines a reasonable local version, which reads after re-formulation of the driving force in terms of densities at the turning points:

$$f_c - f_v = \frac{n(\boldsymbol{x}_c)p(\boldsymbol{x}_v)\, e^{\frac{\Delta E_{F,n}+\Delta E_{F,p}}{k_B T}} - N_c N_v}{\left[N_c + n(\boldsymbol{x}_c)e^{\frac{\Delta E_{F,n}}{k_B T}}\right]\left[N_v + p(\boldsymbol{x}_v)e^{\frac{\Delta E_{F,p}}{k_B T}}\right]}\ , \tag{4.9}$$

where $\Delta E_{F,n} = E_{F,n}(\boldsymbol{x}) - E_{F,n}(\boldsymbol{x}_c)$ and $\Delta E_{F,p} = E_{F,p}(\boldsymbol{x}_v) - E_{F,p}(\boldsymbol{x})$ are the drop of the quasi Fermi levels over the tunnel distances $l_{tn} = |\boldsymbol{x} - \boldsymbol{x}_c|$ and $l_{tp} = |\boldsymbol{x}_v - \boldsymbol{x}|$, respectively (compare Fig. 8). The fully nonlocal description leads to

$$f_c(\boldsymbol{x}_c) - f_v(\boldsymbol{x}_v) = \frac{n(\boldsymbol{x}_c)p(\boldsymbol{x}_v) - N_c N_v}{[N_c + n(\boldsymbol{x}_c)]\,[N_v + p(\boldsymbol{x}_v)]}\ . \tag{4.10}$$

Obviously, the drop of the quasi Fermi potentials over the tunneling length is disregarded in Eq. (4.9). This difference is negligible up to a certain reverse bias. However, for extremely strong generation such that $n(\boldsymbol{x})p(\boldsymbol{x}) \rightarrow n_{i,eff}^2$, a saturation of the BBT occurs in the local version (4.9). By experience, this effect (occurring only near breakdown) vanished using the nonlocal description

in DESSIS$_{-\mathrm{ISE}}$ even for the steepest junctions. Practically, the nonlocal version was realized by a pseudo-local treatment, where the quasi Fermi potentials are extrapolated to the turning points using constant gradients [63].

Fig. 9 compares simulated and measured jV-characteristics of a p^+n^+-diode fabricated as a test structure in the BiCMOS process of Microelectronic Marin. The doping profile was generated by process simulation with DIOS$_{-\mathrm{ISE}}$ [64]. The peak donor and acceptor concentrations were $2 \times 10^{20}\,\mathrm{cm}^{-3}$ and $5 \times 10^{19}\,\mathrm{cm}^{-3}$, respectively. It can be clearly seen how the local DAT rate (which is superior to

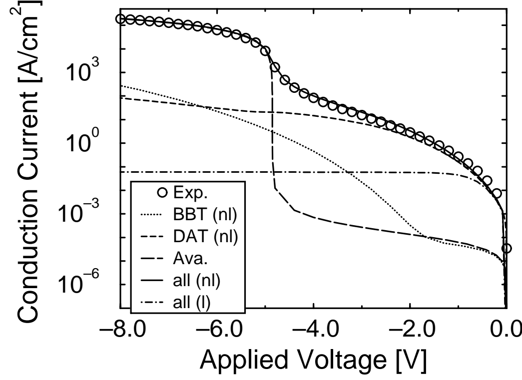

Figure 9 Reverse-bias jV-curves of a p^+n^+-diode simulated with nonlocal (nl) and local (l) models of BBT and DAT in comparison with measured data.

BBT up to -6 V) causes a saturation of the total current for biases larger than -1 V, which even prevents impact ionization from igniting at -5 V.

Finally, we briefly discuss the consequences if a bulk generation rate is treated like a barrier transmission problem. In the latter concept the net rate would have to be defined at the classical turning points, where the carriers appear in generation and disappear in recombination, respectively. For clarity we consider the 1D Esaki diode of Fig. 4. The rate (4.5) becomes different now for electrons and holes, respectively:

$$R^n_{net}(x) = A[F(x + l_{tn})]\ \{f_c(x) - f_v[x + l_t(x)]\}\ \Theta(x_r - x)\ , \quad (4.11)$$

$$R^p_{net}(x) = A[F(x - l_{tp})]\ \{f_c[x - l_t(x)] - f_v(x)\}\ \Theta(x - x_l)\ , \quad (4.12)$$

where l_{tn} and l_{tp} denote the distances to the point of maximum overlap of the valence and conduction band states. Inserting these expressions into the continuity equations for electrons and holes and skipping all the other contributions to the current, one obtains

$$
\begin{aligned}
\nabla\cdot\boldsymbol{J} &= \nabla\cdot\left(\boldsymbol{j}_n+\boldsymbol{j}_p\right)\\
&= qA[F(x)]\left(\{f_c(x)-f_v[x+l_t(x)]\}\,\Theta(x_r-x)-\right.\\
&\qquad\left.-\ \{f_c[x-l_t(x)]-f_v(x)\}\,\Theta(x-x_l)\right)\ , \qquad (4.13)
\end{aligned}
$$

where we have set for the prefactor $A[F(x+l_{tn})] = A[F(x-l_{tp})] = A[F(x)]$ because of the assumption of a constant field *over the tunneling distance* in (4.5). Obviously, $\nabla\cdot\boldsymbol{J}\neq 0$. Integrating in the limits 0 and d it follows that

$$
\begin{aligned}
\boldsymbol{J}(d)-\boldsymbol{J}(0) &= q\int_{x_l}^{x_r+x_l} dx\, A[F(x-x_l)]\,\{f_c(x-x_l)-f_v\left[x-x_l+l_t(x)\right]\}-\\
&\qquad -\ q\int_{x_l}^{d} dx\, A[F(x)]\,\{f_c\left[x-l_t(x)\right]-f_v(x)\}\ . \qquad (4.14)
\end{aligned}
$$

In general, $A[F(x-x_l)]\neq A[F(x)]$ (the prefactor changes with the position in the junction), $x_r\neq d-x_l$ (the junction may be asymmetrical), and $l_t(x)\neq x_l$ (the tunneling length changes with the position in the junction). Hence $\boldsymbol{J}(d)\neq\boldsymbol{J}(0)$, i.e. the contact currents become different as consequence of the violation of local charge conservation.

5 Discussion

It has been the main purpose of this paper to demonstrate the problems when deriving physical models from "first principles" and bringing them into a form suitable for the implementation in a device simulator. From the outlined examples we may set up the following criteria for the often used term "physics-based": A physics-based model should result from a microscopic theory and contain all relevant effects observable from the macroscopic quantity. Parameters of the model are correlated to microscopic parameters like coupling constants, scattering cross sections, etc. These parameters can be tuned by comparison with experimental data. Suitability for complex device simulations includes the following: The model has to be analytical with preferably simple functionality. The implementation must be numerically robust. By experience, the model has to be local (pseudo-local), otherwise convergency problems will be inevitable in more complicated simulation examples. Finally, the model has to be tested for a large variety of cases. This last point, which is often ignored, is may be the most important one for the acceptance and successful application in an industrial environment.

Physics-based models allow a better understanding of the transport processes in a device. However, the demands for short computation times and numerical robustness require a compromise between physical accuracy and analytical

simplicity. The possible accuracy is also limited by the complexity of the problems. The example of BBT has shown that despite the enormous effort it is in principle impossible to describe all parameters exactly. Limits are posed by the complicated band structure and the not well predictable change of the DOS by heavy doping and strong electric fields. Therefore, it is crucial to estimate the range of validity of theoretical parameters, and, as it will be often necessary, to release the problematic ones for a fit to experimental data. Here, "fit" has the positive meaning of fine-tuning a physically well-defined quantity by comparison with suitable measurements.

Unfortunately, many experiments on the physics of silicon and silicon devices go back to the sixties or seventies, where e.g. doping profiles were not precisely known. Besides the desirable update of fundamental experiments, there are completely new problems coming up for which no data exist at all. In modern electrostatic discharge (ESD) protection devices current filamentation leads to local hot spots, where the lattice temperature can approach the melting point. Transport parameters at extremely high temperatures have never been measured, hence in simulations one can only hope that the extrapolation of the temperature dependence into the high-temperature range will at least qualitatively reproduce the actual behavior. Another class of problems is related to the spatial inhomogeneities resulting from todays VLSI technology, like heterojunctions, ultra-thin gate oxides, and narrow field peaks with the extension of a few tens of nanometers. Common approximations, as the WKB and the EMA approximations, become questionable in these cases. At the same time, the validity of the transport equations might be questioned, when the number of charge carriers in the active region of a device becomes very small.

Bibliography

[1] R. Stratton, *Phys. Rev.*, 126(6):2002–14, 1962

[2] K. Bløtekjær, *Ericsson Technics*, 22(2):125–183, 1966

[3] A. Bringer and G. Schön, *J. Appl. Phys.*, 64(5):2447–55, 1988

[4] E. M. Azoff, *Solid-State Electronics*, 30(9):913–917, 1987

[5] E. M. Azoff, In *Proc. NUMOS I Workshop*, pages 25–30, Los Angeles, 1987 Boole Press, Dublin.

[6] M. Rudan and F. Odeh, *COMPEL*, 5(3):149–183, 1986

[7] A. Forghieri, R. Guerrieri, P. Ciampolini, A. Gnudi, R. Rudan, and G. Baccarani, *IEEE Trans. on CAD*, 7(2):231–242, 1988

[8] R. K. Cook and J. Frey, *Compel*, 1(2):65–87, 1982

[9] K. Bløtekjær, *IEEE Trans. Electron Devices*, ED-17(1):38–47, 1970

[10] M. Fukuma and R. H. Uebbing, In *IEDM Technical Digest 184*, pages 621–624, 1984

[11] B. Meinerzhagen, In *Proceedings of the fifth NASECODE conference*, pages 42–59, 1988

[12] B. Meinerzhagen, *IEDM Tech. Digest*, pages 504–07, Dec. 1988

[13] M. A. Stettler, M. A. Alam, and M. S. Lundstrom, *IEEE Trans. Electron Devices*, 40(4):733–740, 1993

[14] S. Selberherr, *Analysis and Simulation of Semiconductor Devices.* Springer-Verlag, Wien, New York, 1984

[15] S. Selberherr, In C. M. Snowden, editor, *Semiconductor Device Modelling*, pages 70–88. Springer-Verlag Berlin Heidelberg, 1989

[16] G. Baccarani, M. Rudan, R. Guerrieri, and P. Ciampolini, In *Proc. of the Comett-Euroform*, DEIS-University of Bologna, Bologna, Italy, March 1991

[17] Roland Ryter, PhD thesis, Swiss Federal Institute of Technology, 1996

[18] S. M. Sze, *Physics of Semiconductor Devices.* John Wiley and Sons, 2nd ed., New York, 1981

[19] G. D. Mahan, *J. Appl. Phys.*, 51(5):2634–46, 1980

[20] K.-F. Berggren and B. E. Sernelius, *Phys. Rev.*, B24(4):1971–86, 1981

[21] L. R. Logan and J. L. Egley, *Phys. Rev.*, B47(19):12532–39, 1993

[22] J. Wagner and A. del Alamo, *J. Appl. Phys.*, 63(2):425–29, 1988

[23] S. C. Jain and D. J. Roulston, *Solid-State Electronics*, 34(5):453–465, 1991

[24] A. Selloni and S. T. Pantelides, *Phys. Rev. Lett.*, 49(8):586–89, 1982

[25] S. T. Pantelides, A. Selloni, and R. Car, *Solid-State Electronics*, 28(1):17–24, 1985

[26] J. C. Inkson, *J. Phys. C*, 9:1177–83, 1976

[27] K.-F. Berggren and B. E. Sernelius, *Phys. Rev.*, B29(10):5575–80, 1984

[28] E. O. Kane, *Phys. Rev.*, 131(1):79–88, 1963

[29] B. I. Halperin and M. Lax, *Phys. Rev.*, 148(2):722–39, 1966

[30] B. I. Halperin and M. Lax, *Phys. Rev.*, 153(3):802–14, 1967

[31] E. O. Kane, *Solid-State Electronics*, 28(1):3–10, 1985

[32] V. Sa-yakanit, *Phys. Rev.*, B19(4):2266–75, 1979

[33] V. Sa-yakanit and H. R. Glyde, *Phys. Rev.*, B22(12):6222–32, 1980

[34] V. Sa-yakanit, W. Sritrakool, and H. R. Glyde, *Phys. Rev.*, B25(4):2776–80, 1982

[35] P. Van Mieghem, S. Decoutere, G. Borghs, and R. Mertens, *Solid-State Electronics*, 35(5):699–704, 1992

[36] R. A. Abram, G. N. Childs, and P. A. Saunderson, *J. Phys. C*, 17:6105–25, 1984

[37] L. Hedin and S. Lundqvist, *Solid State Physics*, 23:1, 1969

[38] J. R. Lowney, *J. Appl. Phys.*, 66(9):4279–83, 1989

[39] M. Reaz Shaheed and C. M. Maziar, *Solid-State Electronics*, 37(9):1589–94, 1994

[40] R. Zimmermann, *Many Particle Theory of Highly Excited Semiconductors.* Texte zur Physik, Band 18. BSB Teubner Verlagsgesellschaft Leipzig, 1988

[41] M. Rösler, F. Thuselt, and R. Zimmermann, *phys. stat. sol. (b)*, 118:303–317, 1983

[42] J. del Alamo, S. Swirhun, and R. M. Swanson, *Solid-State Electronics*, 28(1):47–54, 1985

[43] A. Cuthbertson and P. Ashburn, *IEEE Trans. Electron Devices*, ED-32 (2):242–247, 1985

[44] J. del Alamo and R. M. Swanson, *IEEE Electron Device Letters*, EDL-7(11):629–31, 1986

[45] S. Banerjee, D. Coleman, JR., W. Richardson, and A. Shah, *IEEE Trans. Electron Devices*, ED-35 (1):108–115, 1988

[46] S. H. Voldman, J. B. Johnson, T. D. Linton, and S. L. Titcomb, *IEDM Tech. Digest*, Dec.:349–52, 1990

[47] T. Y. Chan, J. Chen, P. K. Ko, and C. Hu, *IEDM Tech. Digest*, Dec.:718–21, 1987

[48] H. Hazama, *Extended Abstracts of the 22nd Conference on Solid State Devices and Materials, Sendai*, pages 303–306, 1990

[49] I.-C. Chen, D. J. Coleman, and C. W. Teng, *IEEE Electron Device Letters*, EDL-10(7):297–300, 1989

[50] A. G. Chynoweth, R. A. Logan, and D. E. Thomas, *Phys. Rev.*, 125 (3):877–81, 1962

[51] R. A. Logan and A. G. Chynoweth, *Phys. Rev.*, 131 (1):89–95, 1963

[52] L. V. Keldysh, *Soviet Physics JETP*, 6(4):763–770, 1958

[53] L. V. Keldysh, *Soviet Physics JETP*, 7(4):665–669, 1958

[54] P. J. Price and J. M. Radcliffe, *IBM Journal*, Oct.:364–371, 1959

[55] W. V. Houston, *Phys. Rev.*, 57:184–86, 1940

[56] E. O. Kane, *J. Appl. Phys.*, 32 (1):83–91, 1961

[57] R. Enderlein and K. Peuker, *phys. stat. sol. (b)*, 48:231–241, 1971

[58] R. Kubo, *J. Phys. Soc. Japan*, 12(6):570–86, 1957

[59] A. Schenk, *Solid-State Electronics*, 36(1):19–34, 1993

[60] D. E. Aspnes, *Phys. Rev.*, 147:554–561, 1966

[61] L. Esaki and Y. Miyahara, *Solid-State Electronics*, 1:13–21, 1960

[62] S. Müller, K. Kells, A. Benvenuti, J. Litsios, U. Krumbein, A. Schenk, and W. Fichtner, DESSIS 1.3.6: Manual. Technical report, ISE Integrated Systems Engineering AG, 1994

[63] Ulrich Krumbein, PhD thesis, Swiss Federal Institute of Technology, 1996

[64] N. Strecker, T. Feudel, and W. Fichtner, DIOS : Manual. Technical report, ETH Zurich, Integrated Systems Laboratory, ETH Zentrum, 1992