Journal of Engineering Mathematics **38:** 233–263, 2000.

Modelling nonlocal processes in semiconductor devices with exponential difference schemes

R. V. N. MELNIK[1] and HAO HE[2]
[1]*Mathematical Modelling of Industrial Processes, CSIRO Mathematical and Information Sciences – Sydney, Locked Bag 17, North Ryde, NSW 1670, Australia; Corresponding author (E-mail: Roderick.Melnik@cmis.csiro.au)*
[2]*Department of Theoretical Physics, School of Physics, University of Sydney, NSW 2006, Australia*

Received 7 December 1998; accepted in revised form 24 September 1999

Abstract. In this paper nonlocal quasi-hydrodynamic mathematical models describing non-equilibrium physical processes in semiconductor devices are considered. These processes cannot be adequately described with conventional drift-diffusion models. The primary numerical difficulty arises in the energy balance equation. Details of the discretisation for the continuity equations will be described along with a transformation of the energy balance equations to give computationally convenient forms. Effective exponential difference schemes are constructed and applied to modelling transport phenomena in semiconductors. Stability conditions, computational convergence and algorithmic realisations of the proposed schemes are discussed and numerical examples are given.

Key words: time relaxation, semiconductors, quasi-hydrodynamic models, exponential difference schemes

1. Introduction

During recent years microelectronics has provided a wide range of challenging mathematical problems. Amongst them are problems in describing the electron-hole plasma in semiconductor devices, plasmo-chemical etching, ion lithography, fluid and gas epitaxy processes and crystal growth. From the mathematical-physics viewpoint, a number of problems in computational microelectronics can be reduced to mathematical models involving stiff systems of ordinary differential equations and nonlinear partial differential equations including systems of the Navier-Stokes type and the kinetic Boltzmann equations with its variants [1, 2].

Technological advances in the field of microelectronics foster interdisciplinary research between mathematicians, physicists and engineers. The application of many classical algorithms to problems of computational electronics encounters serious mathematical difficulties and technological trends require continuous development of new and efficient numerical techniques. The problems of computational microelectronics become a challenge for applied mathematicians, and as a consequence, a great impetus to the further development of effective numerical methods.

The degree of integration in microelectronics and high configuration density with increasing power density of scattering lead to a situation where the problem of accounting for thermal regimes is critical in the design of microelectronic devices. This includes: (i) the analysis of thermoelectrical conditions of a device and the definition of functional characteristics accounting for local thermal regimes of each device on a substrate [3]; (ii) accounting for the possibility of 'self-heating' of devices.

Our main focus in this paper is the latter problem. The use of Extended Drift-Diffusion Models (EDDM) in the solution of this problem does not account for thermoflux of charge carriers. Typically, such models are obtained under the assumption of thermal equilibrium of charge carriers with the lattice. As a result, EDDM, similar to the classical drift-diffusion model, cannot describe today's semiconductor devices with sufficient accuracy.

In this work we consider and analyse non-local mathematical models which allow us to account for non-equilibrium effects and nonlocal processes in the electron-hole semiconductor plasma. However, an interplay between the oscillatory and diffusive character of transport processes causes major mathematical difficulties in studying transport phenomena which generally includes both parabolic and hyperbolic modes of dynamics. This requires nonlocal models that can describe a combined effect of long and short range forces.

We organise this paper as follows. In Section 2 we briefly describe a hierarchy of mathematical models constructed on the basis of relaxation time approximations. The main emphasis is given to the quasi-hydrodynamic model as an important alternative to the conventional drift-diffusion and kinetic models. In Section 3 we consider problems of flux approximations for the continuity equation and discuss extensions of such approximations to the energy balance equation. The main focus is given to monotone exponential schemes constructed for the discretisation of continuity and energy balance equations in the quasi-hydrodynamic model. Stability issues for these schemes are also discussed in this section. In Section 4 we propose two algorithms for computational implementation of the schemes discussed in Section 3. In Section 5 we specify the choice of the initial approximation and stopping criteria used in our algorithms. In Section 6 we present results of computational experiments. Conclusions and future directions are discussed in Section 7.

2. Modelling transport phenomena in semiconductors via relaxation-time approximations

In the most general setting, mathematical modelling of transport phenomena, including transport phenomena in semiconductors, originated from the Liouville equation for the evolution of the position-velocity probability density. Unfortunately, all Liouville-type models for semiconductor device modelling require the resolution of the following difficulties:

- 6M-dimensional μ-space, used in such models, is unrealistic for modelling many of today's devices;
- adequate models for the driving force as a combination of short-range and long-range interactions are not readily available [4].

In order to overcome these difficulties, it is a common practice to use a hierarchy of mathematical models for the description of transport phenomena [5]. In the semiconductor device context the basis for such a hirarchy can be provided by the concept of relaxation time. The applicability range of mathematical models in semiconductor device theory and its classification is eventually determined by certain functional relationships between the relaxation time and other characteristics of semiconductor plasma. Indeed, physical properties of semiconductor plasma are characterised by a number of fundamental lengths, such as:

- De-Broglie wave length, $\lambda = h/(m^*\tilde{v})$, where $\tilde{v}$ is the characteristic velocity of charge carrier motion, m^* is the effective carrier mass, and $h = 2\pi\hbar$;
- the length of momentum (impulse) relaxation, *i.e.* the length of the free mean path with respect to the momentum, $\lambda_p = \tilde{v}\tau_p$, where τ_p is the momentum relaxation time (the

time that describes the exchange of (quasi-)momentum between carriers and the crystal lattice);

- the length of energy relaxation or the length of 'cooling', $\lambda_\omega = \tilde{v}\sqrt{\tau_p \tau_\omega}$, where τ_ω is the energy relaxation time (the time that describes the exchange of energy between carriers and the crystal lattice).

We consider devices with characteristic dimension l for which at least one of the following inequalities holds

$$l \gg \lambda, \quad l \gg \lambda_p, \quad l \gg \lambda_\omega. \tag{2.1}$$

Strictly speaking, if any of the inequalities (2.1) is violated and l is commensurate with the fundamental lengths defined above, quantum effects may essentially influence the electric characteristics and parameters of devices such as hetero-structures with selective doping, devices with quantum holes and heterojunctions, and thin-layer MOS devices [5].

As follows from the definitions of λ, λ_p, λ_ω, in a specific practical situation the choice of model strongly depends on values of $\tilde{v}$, *i.e.* on the mechanism of scattering. Within a large range of temperatures in many applications $\lambda_p \ll \lambda_\omega$, for example, under scattering on acoustic phonons, we expect $\tau_p \ll \tau_\omega$. Surprisingly, the range of applicability of kinetic models may lie outside this inequality. Therefore, modelling semiconductor devices with kinetic models (the process that typically require the application of costly computational procedures) may not always be justified. In addition, the solution of kinetic models often contains a great deal of redundant information. Computation with the complete kinetic model is relatively efficient only when pair collisions of charge carriers weakly influence the charge transfer. However, if the frequency of pair collisions is fairly high (that is the case for large concentrations, $n \geq 10^{14}$ cm^{-3} and higher), then modelling of devices using kinetic models involves considerable difficulties.

The relaxation of mathematical models in semiconductor device theory may be provided by comparing the role of collisions with other scattering mechanisms. In this case we have to define the range of model applicability with respect to the mean time between the collision that characterises the momentum-and-energy exchange speed [5]. Initially, this consideration leads to two limiting cases that are discussed below.

– Kinetic models (KM) may be efficient in the case when

$$\tau_p \leq \tau_\omega \ll \tau. \tag{2.2}$$

In this case scattering of carriers on each other is not essential. Charge carriers cannot be considered as an independent thermodynamical system, because the scattering of carriers on imperfections of the lattice plays the dominant role. This may include momentum scattering on charged impurity ions, as well as on acoustic/piezoelectric and optical phonons.

– Hydrodynamic models (HDM) are confined to the case

$$\tau \ll \tau_p \ll \tau_\omega, \tag{2.3}$$

when carriers have enough time to exchange by energy and by momentum before the scattering on phonons (and other lattice impurities) becomes essential. In this case the electron-hole plasma (EHP) can be considered as an almost independent thermodynamical system that only weakly interacts with the lattice. We do not require that the temperature of lattice, T_l, should be equal to the carrier temperature (electron temperature, T_n, or hole temperature, T_p), but we think of the motion of the carrier system as a whole with respect to the lattice. Using analogy with fluid dynamics, we refer to the models based

on this reasoning as hydrodynamic. In these macroscopic models physical quantities are averaged over the whole carrier (electron/hole) population, and the sought-for information is substantially reduced compared to kinetic models. However, it is well known that locally such models may not correctly describe many important physical process such as impact ionisation caused by the influence of hot carrier subpopulations [7].

The electro-hydrodynamic model for an electron system is often written in the following way (see [8–10] and references therein):

$$\frac{\partial \mathbf{z}}{\partial t} = \boldsymbol{\zeta} + \left(\frac{\partial \mathbf{z}}{\partial t}\right)_{\text{col}}, \tag{2.4}$$

where

$$\mathbf{z} = (n, \mathbf{v}, W)^T, \quad \boldsymbol{\zeta} = (\mathcal{F}_1, \boldsymbol{\mathcal{F}}_2, \mathcal{F}_3)^T, \quad \mathcal{F}_1 = -\nabla \cdot (n\mathbf{v}), \tag{2.5}$$

$$\boldsymbol{\mathcal{F}}_2 = -\mathbf{v} \cdot \nabla \mathbf{v} - q\mathbf{E}_{\text{eff}}/m_n - \nabla(nT_n)/(m_n n), \tag{2.6}$$

$$\mathcal{F}_3 = -\nabla \cdot (\mathbf{v}W) - qn\mathbf{v} \cdot \mathbf{E}_{\text{eff}} - \nabla \cdot (\mathbf{v}nT_n) - \nabla \cdot \mathbf{q}, \tag{2.7}$$

T_n is the electron temperature given in energetic units, n is the electron concentration, $\mathbf{v}$ is their averaged velocity, W is the energy density (typically modelled by $W = 3nT_n/2 + m_n n\|\mathbf{v}\|^2$), $\mathbf{q}$ is the heat flow (typically modelled by the Fourier law $\mathbf{q} = -k\nabla T_n$), E_{eff} is the effective electric strain, and m_n is the effective electron mass (see Appendix for the list of notation). During recent years attempts have been made to improve hydrodynamic models using the method of moments and taking into account moments of higher orders [11].

A useful simplification of model (2.4)–(2.7), well investigated mathematically, is provided by

– Drift-diffusion models (DDM). However, the derivation of the DDM is usually based on a version of the Hilbert expansion and can be rigorously justified only for low carrier densities and small electric fields (see details in [5]). Due to the technological advances connected with the miniaturization and the use of materials other than silicon, this may not be sufficient for the adequate modelling of many new devices.

As a result, the development of the next generation of mathematical models and numerical methods for their solutions has become an important challenging problem in applied mathematics.

2.1. Quasi-hydrodynamic models: the range of applicability and physical parametrisation

A certain compromise between the model types described above gives quasi-hydrodynamic models (QHDM). In fact, there is a number of reasons in favor of the development of models other than hydrodynamic, kinetic and drift-diffusion types. Firstly, conditions for application of hydrodynamic models are quite restrictive from the physical point of view. Such conditions can be justified under conditions of strong injection or in application to low-bandgap materials when the intrinsic concentration of charge carriers is very large. Secondly, the application of kinetic models is connected with essential computational difficulties. Thirdly, although mathematical investigation of drift-diffusion type models has achieved some maturity, these models cannot predict important physical phenomena such as carrier heating or velocity overshoot. By

now it is clear that drift-diffusion type models are not compatible with technological advances. Moving to the next generation of mathematical models in this field means accounting for *non-equilibrium* and *non-local* behaviour of semiconductor plasma.

A wide area of applications is confined to the situation, which may not overlap with (2.2) or (2.3), when

$$\tau_p \le \tau \ll \tau_\omega. \tag{2.8}$$

From a physical point of view this means that charge carriers have enough time to repeatedly exchange by energy (but not by momentum!) before the scattering on phonons becomes essential. Plasma of charge carriers achieves its equilibrium after time τ, *i.e.* long before the time when the exchange between carriers and the lattice becomes noticeable. Hence, in this case our perception on carrier temperature is quite definite. Indeed, with respect to the energy, plasma of charge carriers can be considered as an *almost independent* thermodynamical system. Of course, it is not true any more with respect to the momentum because the scattering takes place mainly on impurities of the lattice. This approximation leads to mathematical models of quasi-hydrodynamic type. The area of applicability of such models is wider than that of hydrodynamic models, although physical simplifications connected with the application of quasi-hydrodynamic type models are similar to those for hydrodynamic models (see details in [5]).

Following [12–15], in the space-time region $\bar{G}^R = \{(x,t):\ 0 \le x \le L,\ 0 \le t \le T\}$ we consider the following quasi-hydrodynamic model for semiconductor device modelling

$$\begin{aligned} &\partial_{xx}\varphi = q(n-p-N)/\epsilon\epsilon_0, \quad \partial_t n - \partial_x J_n/q = F, \quad \partial_t p + \partial_x J_p/q = F, \\ &\partial_t \bar{\mathcal{E}}_n + \partial_x Q_n = -J_n\partial_x\varphi + P_n, \quad \partial_t \bar{\mathcal{E}}_p + \partial_x Q_p = -J_p\partial_x\varphi + P_p, \end{aligned} \tag{2.9}$$

where expressions for densities of carrier currents, J_n, J_p and flux energies, Q_n and Q_p have the following form

$$J_n = -qn\mu_n\partial_x\varphi + q\partial_x(D_n n), \quad J_p = -qp\mu_p\partial_x\varphi - \partial_x(D_p p), \tag{2.10}$$

$$Q_n = \beta_n T_n n\mu_n\partial_x\varphi - \beta_n\partial_x[T_n D_n n]/q, \quad Q_p = -\beta_p T_p p\mu_p\partial_x\varphi - \beta_p\partial_x[T_p D_p p]/q, \tag{2.11}$$

$\bar{\mathcal{E}}_n = 3nT_n/2$, $\bar{\mathcal{E}}_p = 3pT_p/2$ are the average densities of the electron and hole systems respectively, and F is an approximation to the contribution of the generation-recombination (and, possibly, ionisation) processes.

The Peltier coefficients, β_n and β_p in (2.11) for typical Si and GaAs semiconductors take values between 2 and 3 and can be well approximated by the following formulae

$$\beta_n = 2{\cdot}5 + \xi_n, \quad \beta_p = 2{\cdot}5 + \xi_p, \tag{2.12}$$

where $\xi_n = \mathrm{d}\log\mu_n(T_n)/\mathrm{d}\log T_n$, $\xi_p = \mathrm{d}\log\mu_p(T_p)/\mathrm{d}\log T_p$ [13]. The first terms in the RHS of the energy balance equations represent the velocity of Joule heating/cooling. The second terms represent the velocity of energy losses induced by scattering on the lattice and are modelled by the formulae

$$P_n = n(T_l - T_n)/\tau_\omega^n, \quad P_p = p(T_l - T_p)/\tau_\omega^p, \tag{2.13}$$

where temperature is considered in energy units, and τ_ω^n, τ_ω^p are the average energy relaxation times for electrons and holes respectively. We can easily include the velocity of energy exchange between electrons and holes (and *vice versa*) as well as the energy of electron and hole subsystems due to non-elastic collisions (recombination and ionization). Without loss of mathematical generality, we do not include these processes in our numerical procedures. As physical support for this simplification, we note that in many semiconductor structures like high-speed diodes and transistors, transition periods are small compared to characteristic times of energy exchange between carriers and the time of recombination/generation of carriers.

We also assume that carriers in different valleys have the same effective temperature. Then if n_i is the concentration of electrons in the ith valley we have $n = \sum_i n_i$. In this case properties of the carrier systems are characterized by the fact that average mobilities ($\mu_n = \mu_n(T_n)$, $\mu_p = \mu_p(T_p)$), diffusion coefficients ($D_n = D_n(T_n)$, $D_p = D_p(T_p)$), and times of energy relaxation ($\tau_\omega^n = \tau_\omega^n(T_n)$, $\tau_\omega^p = \tau_\omega^p(T_p)$) are dependent on carrier temperatures (see [13]). We can approximate these dependencies by using a number of models known in the literature [10, 13, 16]. Typically, it is assumed that impulse scattering takes place mainly on acoustic phonons, so that the energy relaxation time can be approximated as a sum of two terms. The first one takes into account deformational acoustic phonons, and the second is due to between-valley acoustic phonons. For example, for the electron system we have [13]

$$\frac{1}{\tau_\omega^n(T_n)} = \frac{1}{\tau_a(T_n/T_l)^{-1/2}} + \frac{\exp(-\hbar\omega_0/T_n)}{\tau_o(T_n/T_l)^{1/2}}, \tag{2.14}$$

where in the second term we use a 'one'-phonon approximation ($\hbar\omega_0$ is the mean energy of an optical phonon), and τ_a, τ_o are temperature-dependent time-constants that characterise deformational and between-valley acoustic phonons. When the lattice temperature is close to 300° K, the contribution of the first term for Si devices becomes smaller. Another approximation often used in the literature (see [10] and references therein) has the following form

$$\tau_\omega^n = \frac{m_n\mu_n^0}{2q}\frac{T_0}{T_n} + \frac{3\mu_n^0}{2q(v_s^n)^2}\frac{T_nT_0}{T_n+T_0}, \tag{2.15}$$

where the velocity saturation, v_s^n, depends on the lattice temperature [9, 6] (typically it is of the order $10^6 - 10^7$ cm s^{-1}). In order to avoid unnecessary technicalities we follow [16] by setting

$$\mu_n = \mu_n^0(T_n/T_l)^q, \quad \mu_p = \mu_p^0(T_p/T_l)^q, \tag{2.16}$$

$$\tau_\omega^n = \tau_{\omega,0}^n(T_l/T_n)^s, \quad \tau_\omega^p = \tau_{\omega,0}^p(T_l/T_p)^s, \tag{2.17}$$

where q and s are determined by the dominant relaxation mechanisms of the momentum and energy. Computational results, reported in Section 6, were obtained for $q = s = 0$ with the low-field mobilities taken as $\mu_n^0 = 1300$ cm^2/Vs, $\mu_p^0 = 400$ cm^2/Vs and the energy relaxation times set as $\tau_\omega^n = \tau_\omega^p = 0{\cdot}4 \times 10^{-12}$ s [12].

As for the dependencies of the diffusion coefficients on carrier temperatures, we admit that the numerical procedures developed in the next sections can be easily generalized to the general type of dependence

$$D_n(T_n)/\mu_n(T_n) = \tilde{f}_1(T_n), \quad D_p(T_p)/\mu_p(T_p) = \tilde{f}_2(T_p). \tag{2.18}$$

To be specific, we developed the computational procedures under the assumption of the Einstein-type relationship, that is

$$D_n(T_n) \sim T_n \mu_n(T_n) \quad D_p(T_p) \sim T_p \mu_p(T_p) \tag{2.19}$$

with the constant of proportionality equal to $k_b/q = 8{\cdot}61738 \times 10^{-5}$ eV/K.

Initial conditions for the model are

$$n(x,0) = n_0(x), \quad p(x,0) = p_0(x), \quad T_n(x,0) = T_p(x,0) = T_l, \quad 0 \le x \le L. \tag{2.20}$$

We assume that the functions $n_0(x)$ and $p_0(x)$ in the initial conditions are defined as equilibrium values of densities for electrons and holes, that is

$$p_0(x) n_0(x) = n_{ie}^2, \quad n_0(x) - p_0(x) - N = 0, \tag{2.21}$$

where n_{ie} is the effective intrinsic concentration of carriers.

In the general case, boundary conditions depend on the type of modelling structure. In this paper we require:
(a) equality of carrier temperature and lattice temperature

$$T_n(0,t) = T_p(L,t) = T_l; \tag{2.22}$$

(b) conditions of quasi-neutrality and infinite velocity of recombination (thermodynamic equilibrium):

$$p - n + N = 0, \ pn = n_{ie}^2, \ x \in \partial G^R = \{0, L\}, \tag{2.23}$$

from where it is easy to get

$$n = \frac{N}{2} + \sqrt{\left(\frac{N}{2}\right)^2 + n_{ie}^2}, \quad p = -\frac{N}{2} + \sqrt{\left(\frac{N}{2}\right)^2 + n_{ie}^2}, \ x \in \partial G^R = \{0, L\}. \tag{2.24}$$

For the potential, boundary conditions are standard [4]

$$\varphi(0,t) = 0, \quad \varphi(L,t) = U + \varphi_{\text{cont}}, \tag{2.25}$$

where U is the applied voltage and φ_{cont} is the contact potential difference determined by the formula $\varphi_{\text{cont}} = \varphi_T \log(n(t,L)/n_{ie}$ (obtained as a consequence of $n = n_{ie} \exp((\varphi - \varphi_n)/\varphi_T)$ by setting $\varphi_n = U$). In other words, we assume that the bias is applied at the right contact, while the left contact is grounded. In this case we require the conjugating conditions

$$\varphi(0,0) = 0, \ \varphi(L,0) = U + \varphi_{\text{cont}} \tag{2.26}$$

to be satisfied. We note that if $\varphi_{\text{cont}}>0$ then the case $U<0$ corresponds to forward bias, and the case $U>0$ corresponds to reverse bias.

For the effective intrinsic concentration, n_{ie}, in (2.21), (2.23), (2.24), (2.26) one has to use an empirical model, for example (see [17, 6, 18] and references therein):

$$n_{ie} = n_{\text{int}}(T) \exp(q \Delta E_g/(2T)), \tag{2.27}$$

where T is the absolute temperature taken in energy units (multiply by the factor k_b/q), and ΔE_g is an *experimentally* measured parameter known as the effective bandgap narrowing. The energy gap itself, $E_g = E_c - E_v$, defined as the difference between the bottom of the conduction band, E_c, and the ceiling of the valence band, E_v, may change with different doping profiles and temperature. It is known, for example, that for highly doped material and high temperatures the bandgaps become smaller. Formula (2.27) is meant to take into account such changes. The intrinsic concentration, n_{int}, in formula (2.27) depends on the effective number of states in the conduction and valent zones (N_c and N_v, respectively). It is common practice to use the following formula for its approximation

$$n_{\text{int}} = \sqrt{N_c N_v}\exp(-E_g/2k_bT), \tag{2.28}$$

where

$$N_c = 2\left(\frac{2\pi m^*_{dn} k_b T}{h^2}\right)^{1.5} M_c, \quad N_v = 2\left(\frac{2\pi m^*_{dp} k_b T}{h^2}\right)^{1.5}, \tag{2.29}$$

M_c is the number of equivalent minima in the conduction zone, and m^*_{dn}, m^*_{dp} are the density-of-state of effective masses of electrons and holes respectively (see [18], p. 17). It is easy to see that when $\Delta E_g \to 0$, the effective intrinsic concentration can be well approximated by the intrinsic concentration. This assumption is used in our code where we set $n_{ie} \approx n_{\text{int}} = 1{\cdot}45 \times 10^{10}$ cm^{-3} (see [18], p.850).

The recombination model was chosen to account for recombination on defects induced by dopants (the Shockley-Read-Hall recombination) and between-zone Auger recombination:

$$F(n, p) = \frac{pn - n_{ie}^2}{\tau_n(p + n_{ie}) + \tau_p(n + n_{ie})} + (pn - n_{ie}^2)(c_n n + c_p p), \tag{2.30}$$

where the carrier lifetimes and coefficients of Auger recombination are set as follows $\tau_n = 1 \cdot 7 \times 10^{-5}$ s, $\tau_p = 3 \cdot 95 \times 10^{-4}$ s, $c_n = 2{\cdot}9 \times 10^{-31}$ cm^6/s, $c_p = 1{\cdot}2 \times 10^{-31}$ cm^6/s. For the problem where impact ionisation plays a significant role we have to add the velocity of ionization term

$$G_p - G_n = \alpha_p J_p - \alpha_n J_n, \tag{2.31}$$

where

$$J_n = -qnv_n, \; J_p = qpv_p \tag{2.32}$$

and α_n, α_p are field-dependent carrier ionisation rates defined as the number of electron-hole pairs generated by an electron/hole per unit distance travelled [18]. Although in our numerical examples, presented in Section 6, only the Shockley–Read–Hall recombination was considered, our code is easily adaptable to account for other processes such as ionisation.

2.2. Normalization procedure and challenges in the computational treatment of nonlocal models

The magnitudes of dependent variables in the quasi-hydrodynamic model critically vary amongst each other, leading to a substantial computational cost of associated numerical procedures. In order to reduce the cost, effective normalisation procedures have to be implemented for the quasi-hydrodynamic model [17, 4].

We reduce model (2.9)–(2.11) to the following normalised system considered in the space-time region $\bar{G} = \{(x,t) :\ 0 \le x \le 1,\ 0 \le t \le \tilde{T}\}$ (see notation in Appendix and details in [5])

$$\begin{aligned} &\partial_{xx}\varphi = n - p - N, \quad \partial_t n - \partial_x J_n = F, \quad 3/2\partial_t(nT_n) + \partial_x Q_n = -J_n\partial_x\varphi + P_n, \\ &\partial_t p + \partial_x J_p = F, \quad 3/2\partial_t(pT_p) + \partial_x Q_p = -J_p\partial_x\varphi + P_p, \end{aligned} \tag{2.33}$$

where

$$J_n = -n\mu_n\partial_x\varphi + \partial_x(T_n\mu_n n), \quad J_p = -p\mu_p\partial_x\varphi - \partial_x(T_p\mu_p p), \tag{2.34}$$

$$Q_n = \beta_n T_n n\mu_n\partial_x\varphi - \beta_n\partial_x[T_n D_n n], \quad Q_p = -\beta_p T_p p\mu_p\partial_x\varphi - \beta_p\partial_x[T_p D_p p]. \tag{2.35}$$

The system (2.33)–(2.35) is supplemented by the normalised initial

$$n(x,0) = \tilde{n}_0(x),\ p(x,0) = \tilde{p}_0(x),\ T_n(x,0) = T_p(x,0) = 1, \tag{2.36}$$

and boundary conditions

$$p - n + N = 0, \quad pn = n_{ie}^2, \quad T_n = T_p = 1, \quad x \in \partial\bar{G} = \{0,1\}, \tag{2.37}$$

$$\varphi(0,t) = 0, \quad \varphi(1,t) = \tilde{U} + \tilde{\varphi}_{\text{cont}}. \tag{2.38}$$

It is also assumed that the condition $J_n(x,0) = J_p(x,0) = 0$ and the normalised conjugating conditions $\varphi(0,0) = 0$, $\varphi(1,0) = \tilde{U} + \tilde{\varphi}_{\text{cont}}$ are satisfied.

Model (2.33)–(2.38) allows us to adequately describe a number of non-stationary physical phenomena in semiconductor devices, including carrier heating and velocity overshoot. One of the main features of this model is accounting for a non-equilibrium and non-local character of electron-hole semiconductor plasma, a feature absent in the classical drift-diffusion model. Since technological advances lead to further reduction of device sizes, a higher density of configuration and power density of scattering, non-local and non-equilibrium phenomena are becoming increasingly important in device simulation.

In order to adequately describe these phenomena of semiconductor plasma it is often unnecessary to invoke the Boltzmann model, solution of which is known to be costly and 'noisy' with a great deal of redundant information [19]. An important direction in engineering applications of semiconductor device theory is the analysis of 'intermediate' (between the Boltzmann and drift-diffusion) models, such as (2.33)–(2.38). These models require efficient computational procedures, the development and justification of which is a challenging problem in applied mathematics.

In contrast to DDM, for which numerical methods have undergone extensive development, starting with Gummel's work (see, for example, references in [20, 4] and others), efficient numerical methods constructed for nonlocal-type models are at the beginning of their development. Modelling with nonlocal models incurs considerable mathematical difficulties, the overcoming of which is a challenging problem in applied mathematics [2]. For example, considering models of quasi-hydrodynamic type, we are dealing with fairly complex, strongly nonlinear problems of coupled field theory. Therefore, the development of effective numerical algorithms is required for the investigation of physical processes within the framework of such models. A large amount of publications is devoted to results of computations for specific devices [12, 10, 21, 22, 9, 7, 20, 23, 24, 25, 11]. However, the analysis of cost-effective algorithms is still lacking in the literature.

Mathematical modelling of non-local phenomena such as ballistic transfer and the velocity overshoot in semiconductor plasma has progressed since the early 1980's (although physical effects were described long before that time). The common feature of models for such phenomena is the accounting for macroscopic parameters for which balance laws are written with respect to the average energy. As a result, in contrast to DDM (where average energy is a local function of the field) such models are classified as nonlocal models.

A typical example of non-local mathematical models in semiconductor device theory is provided by the quasi-hydrodynamic model. A straightforward mathematical procedure for the solution of (2.33)–(2.38) is the Newton–Raphson method, applied to the discretised system of nonlinear equations [12, 25, 10]. This procedure is quite costly when applied to realistic semiconductor devices. Usually we have to apply special techniques in order to obtain convergence for the whole discretized system. For example, we may apply a sequentially-simultaneous algorithm which increases the convergence rate by conducting internal (adiabatic) iterations under fixed carrier temperature for three equations of DDM (prior to solving the coupled system of all five equations).

Alternative approaches to the solution of problem (2.33)–(2.38) are often based on different types of splitting algorithms [22]. The application of such approaches to strongly coupled problems encountered considerable difficulties in the context of semiconductor devices, especially for large electric fields. Another group of approaches uses different versions of the macro-particle method, where carrier collisions are modelled by Monte-Carlo type procedures [26]. The methods in this group are known to be typically costly and 'noisy' in the computational sense. The principal problem with the macro-particle approach lies in the adequate modelling of pair collisions, a problem remains open to a large extent [27]. References to other recently developed computational procedures for semiconductor device models can be found in [28, 1].

In the next sections, using the quasi-hydrodynamic model as a typical example of non-local models, we demonstrate the main ideas of the construction of effective numerical schemes which can also be applied to hydrodynamic and classical drift-diffusion models.

3. Numerical approximations for nonlocal models of quasi-hydrodynamic type

One of the most important properties required by difference schemes in semiconductor device theory is monotonicity. Indeed, we have to guarantee that the solutions of the continuity and energy balance equations are nonnegative ($n, p, T_n, T_p \geq 0$) for any function of the potential φ. Let us consider these issues in some details.

First, we introduce a non-uniform grid in $\bar{G}$

$$\hat{\omega}_{h\tau} = \hat{\omega}_h \times \hat{\omega}_\tau, \tag{3.1}$$

where

$$\hat{\omega}_h = \left\{ x_{i+1} = x_i + h_i,\ i = 0, ..., N, x_0 = 0,\ x_{N+1} = 1,\ \sum_{i=0}^{N} h_i = 1 \right\},$$

$$\hat{\omega}_\tau = \left\{ t_j = t_{j-1} + \tau_j,\ j = 1, ..., K-1,\ t_0 = 0,\ t_K = T_f, \sum_{j=1}^{K-1} \tau_j = T_f \right\}.$$

We will compute the values of φ, n, p, T_n, and T_p in the 'whole' nodes (*i.e.* x_i, $i = 0, 1, \ldots, N+1$), whereas the values of J_n, J_p, Q_n, Q_p, and $E = -\nabla\varphi$ will be computed in the data-driven (flux) nodes (*i.e.* $x_{i+1/2}$, i=0, ... ,N).

3.1. Flux approximations and transformation of the energy balance equations to forms amenable to computational efficiency

The approximation of fluxes is a long-standing problem in many applied problems for which solutions have steep gradients. In the context of semiconductor device modelling, we recall that even in the DDM, for which current density is defined as $J_n = -\mu_n n\nabla\varphi + \mu_n \nabla n$, the application of standard approximations is impeded because of the very restrictive condition on the space step discretisation which follows from the *maximum principle*. In order to obtain this condition, one can use, for example, the theorem on monotonicity of three-point difference operators (so called the Karetkina lemma, see [14, 15, 5] and references therein). Typically, the conditions of this theorem will be satisfied for

$$h < 2/E^*, \text{ where } E^* = \max_{i=1,\ldots N} |E_{i+1/2}|. \tag{3.2}$$

In the case of the QHDM (2.33)–(2.38), the monotonicity condition for the standard current-density approximation in the form

$$J_{n,i+1/2} = \frac{D(T_{i+1})n_{i+1} - D(T_i)n_i}{h_{i+1}} - \frac{n_{i+1}\mu(T_{i+1}) + n_i\mu(T_i)}{2}\frac{\varphi_{i+1} - \varphi_i}{h_{i+1}} \tag{3.3}$$

coincides with (3.2).

The first monotone difference scheme in a semiconductor-device-modelling context was first reported by D. L. Scharfetter and H. K. Gummel (see references, for example, in [20, 29, 14] and others). These types of schemes, known to the mathematical community as *exponential*, constitute an important tool in the integration of stiff ordinary differential equations [30, 31]. For ODEs they are typically unconditionally stable and, what is very important, positivity of the solution is guaranteed if the solution of the differential problem is expected to be positive. They can also be constructed without major difficulties in the case of partial differential equations when the spatial differential operator can be reduced to the self-conjugate form (see [32] and references therein). For the continuity equations of the classical DDM, the idea of such a reduction has been intensively investigated. If the Boltzmann statistics is assumed, then the exponential change of variables

$$n = n_{ie}\exp(\varphi)\Phi_n, \quad p = n_{ie}\exp(-\varphi)\Phi_p, \tag{3.4}$$

leads to an essential simplification of the current densities which become linearly dependent on quasi-potentials Φ_n, Φ_p. From the mathematical point of view, this is a very attractive feature of the model that, in turn, leads to a number of effective algorithms [20]. Such algorithms were also constructed in the case of Fermi statistics [29]. However, it should be noted that the practical value of all such schemes is essentially dependent on the quality of approximation of the strongly nonlinear RHS of the continuity equation.

In the 70's and early 80's, works in the application of non-local models to semiconductor device simulation were conducted predominantly with Monte-Carlo type procedures, and in those papers where difference methods were used, questions of scheme quality have not been adequately explored. The turning point in the development of difference methods for the

QHDM was the work of Tang [33], where the Scharfetter–Gummel approximation was generalized to the case of a particular type of non-local model. The current density approximation was considered in the following form (we omit the indexes n and p for the simplicity):

$$J_{i+1/2} = \frac{1}{h_{i+1}} \left[D(T_{i+1}) n_{i+1} f_1 \left(\frac{\varphi_{i+1} - \varphi_i}{T_{i+1/2}} \right) - D(T_i) n_i f \left(\frac{\varphi_{i+1} - \varphi_i}{T_{i+1/2}} \right) \right], \tag{3.5}$$

where

$$f(x) = x \exp(x)/(\exp(x) - 1), \quad f_1(x) = x/(\exp(x) - 1) \tag{3.6}$$

are Bernoulli functions (bearing a computer code in mind, we recall that $f_1(x) = f(-x)$), and the quantities $T_{i+1/2}$ may be approximated by any value of temperature on the integration interval $[x_i, x_{i+1}]$, for example, T_i, $(T_i + T_{i+1})/2$, T_{i+1}. As we expect, such approximation turns into the Scharfetter–Gummel approximation when $T_{i+1} \to T_i$.

Unfortunately in the general case, even the Scharfetter–Gummel-type approximation cannot guarantee absolute stability neither for the DDM nor for nonlocal models. We can only claim the conservation of solution positiveness for certain 'model' problems for the continuity equation, such as

$$\left(\mu_n \left(\frac{\partial n}{\partial x} - n \frac{\partial \varphi}{\partial x} \right) \right)_x = 0. \tag{3.7}$$

We can also claim the conservation of positiveness property for a specific computational experiment. However, if we consider the problem with even the 'simplest' recombination model (say, the Shockley–Read–Hall recombination), then the stability of the method depends on the method of linearization of the recombination term. Of course, there exist linearisation procedures for the recombination term (such as the Seidman–Choo procedure) that satisfy all conditions of the monotonicity theorem. And yet, if other processes such as ionization are dominant and therefore have to be taken into account, then the RHS linearization is typically a heuristic procedure, aimed at the achievement of numerical stability and algorithm convergence.

Challenging mathematical and computational problems also arise in the approximation of energy balance equations for nonlocal models. Since the approximation for the energy flux, analogous to the Scharfetter–Gummel approximation of the current density, is known [33, 34], the main challenge is to transform the energy balance equations to forms that are most suitable for an efficient computational implementation.

In contrast to the continuity equation, the energy balance equation *cannot be readily reduced* to a 'divergent' or 'conservation' form (see [13, 32] and references therein). In the semiconductor-device modelling context, the main problem with the energy balance equation lies with the presence of the product between the current density and the electric field strength ($J_n \times E$ or $J_p \times E$), that has a 'non-divergent' structure. Hence, one cannot immediately apply the general theory developed for the construction of monotone difference schemes (see [32, 31] and references therein). However, since the product between the current density and the electric field strength provides the key to the nonlocal coupling between the effective carrier temperature and the electric field, the problem of its efficient approximation has to be dealt with.

Following [33, 13, 14], in [5] it was shown that the energy fluxes can be trasformed to forms where all derivatives of $\mathcal{E}_n$ and $\mathcal{E}_p$ are 'covered' by the symbol of divergence. In particular, the energy balance equation for the system of electrons can be written in the form

$$3\partial_t \mathcal{E}_n/2 = \partial_x Q_n^* + S_n(T_n, \varphi)\mathcal{E}_n, \tag{3.8}$$

where

$$\mathcal{E}_n = nT_n, \;\; S_n = \mu_n(T_n)\partial_{xx}\varphi + \mu_n(T_n)(\partial_x\varphi)^2 + (1 - T_n)/(\tau_\omega^n(T_n)T_n), \tag{3.9}$$

$$Q_n^* = \beta_n \partial_x(D_n(T_n)\mathcal{E}_n) - (1 + \beta_n)\mu_n(T_n)\mathcal{E}_n\partial_x\varphi. \tag{3.10}$$

Similarly, the balance energy equation for the system of holes can be transformed to the following form

$$3\partial_t \mathcal{E}_p/2 = \partial_x Q_p^* + S_p(T_p, \varphi)\mathcal{E}_n, \tag{3.11}$$

where

$$\mathcal{E}_p = pT_p, \;\; S_p = -\mu_p(T_p)\partial_{xx}\varphi + \mu_p(T_p)(\partial_x\varphi)^2 + (1 - T_p)/(\tau_\omega^p(T_p)T_p), \tag{3.12}$$

$$Q_p^* = \beta_p \partial_x(D_p(T_p)\mathcal{E}_p) + (1 + \beta_p)\mu_p(T_p)\mathcal{E}_p\partial_x\varphi. \tag{3.13}$$

The representations (3.8) and (3.11) allow us to construct *monotone exponential* difference schemes (see [32] and references therein) for nonlocal models applied to semiconductor device simulation.

3.2. Monotone exponential schemes for the continuity and energy balance equations

Major mathematical and computational challenges in the numerical solution of system (2.33)–(2.38) are connected with efficient approximation of the energy balance equation [19, 17, 13, 15, 33].

The stationary case provides us with a clear picture of computational difficulties which can be 'hidden' in the non-stationary case by an appropriate reduction of the time step. Indeed, in this case standard approximations lead us to the restriction on the space discretisation step similar to (3.2) (see details in [33, 13, 5])

$$h < 2\beta/E^*, \quad \text{where} \quad E^* = \max_{i=1,\dots,N_0} |E_{i+1/2}|, \tag{3.14}$$

which may be burdensome for high electric fields. To overcome this condition is difficult. Indeed, if a splitting technique is used for the numerical solution of (2.33)–(2.38) and condition (3.14) is violated, then in the general case the positiveness of the solution cannot be guaranteed. In order to relax the requirement (3.14), we apply exponential difference schemes.

3.2.1. *Continuity equations*

We recall the procedure for the construction of exponential difference schemes on the example of the 1D stationary continuity equation in the absence of recombination/generation/ionisation processes

$$\frac{\partial J_n}{\partial x} = 0. \tag{3.15}$$

The standard change of variables $n(x) \to n^{\text{new}}(x)$ in this case (see [13, 14, 33] and references therein) is

$$n(x) = n^{\text{new}}(x) \exp\left[\int_{x_0}^{x} \left\{\frac{1}{T_n}\frac{\partial\varphi(\xi)}{\partial\xi} - \frac{1}{D_n(T_n)}\frac{\partial D_n(\xi)}{\partial\xi}\right\} \mathrm{d}\xi\right], \tag{3.16}$$

where x_0 is an arbitrary number such that $x_0 < x$. Substitution (3.16) in (2.34) leads to the following expression for the current density:

$$J_n(x) = D_n(T_n) \exp\left[\int_{x_0}^{x} \left\{\frac{1}{T_n}\frac{\partial\varphi(\xi)}{\partial_\xi} - \frac{\log D_n(\xi)}{\partial_\xi}\right\} \mathrm{d}\xi\right] \frac{\partial n^{\text{new}}}{\partial x}. \tag{3.17}$$

If we now integrate this expression on the interval $[x_i, x_{i+1}]$ (assuming that the quantities $(J_n)_{i+1/2}$, $(D_n)_{i+1/2}$, $(\mu_n)_{i+1/2}$ are constants) and return to the old variable $n(x)$ we obtain

$$(J_n)_{i+1/2} = \frac{1}{\int_{x_i}^{x_{i+1}} J_n^*(x)dx}(D_n)_{i+1/2}[n_{i+1}J_n^*(x_{i+1}) - n_i], \tag{3.18}$$

where

$$\begin{aligned} J_n^*(x) = \ & \exp\left[-\int_{x_i}^{x} \left\{\frac{1}{T_n(\xi)}\frac{\partial\varphi(\xi)}{\partial\xi} - \frac{\partial \log D_n(\xi)}{\partial\xi}\right\} \mathrm{d}\xi\right] = \\ & \exp\left[-\frac{\varphi(x) - \varphi(x_i)}{T_n(x^*)} + \log\frac{D_n(x)}{D_n(x_i)}\right]. \end{aligned} \tag{3.19}$$

Assuming, for example, that for $x^* \in [x_i, x_{i+1}]$

$$T_n(x^*) = \text{const} = \tfrac{1}{2}((T_n)_i + (T_n)_{i+1}) = (T_n)_{i+1/2}, \tag{3.20}$$

we can transform expression (3.18) into the following form:

$$\begin{aligned} (J_n)_{i+1/2} = \ & \frac{D_n((T_n)_i)}{h}\frac{\varphi_{i+1} - \varphi_i}{(T_n)_{i+1/2}}\left[\exp\frac{\varphi_{i+1} - \varphi_i}{(T_n)_{i+1/2}} - 1\right]^{-1} \times \\ & \times \left[n_{i+1}\frac{(D_n)_{i+1}}{(D_n)_i} - n_i \exp\left(\frac{\varphi_{i+1} - \varphi_i}{(T_n)_{i+1/2}}\right)\right], \end{aligned} \tag{3.21}$$

which coincides with the approximation (3.5). Now, if we integrate Equation (3.15) on the interval $[x_{i-1/2}, x_{i+1/2}]$, we obtain that

$$[(J_n)_{i+1/2} - (J_n)_{i-1/2}]h = 0. \tag{3.22}$$

Substitution of the corresponding expressions for $(J_n)_{i\pm1/2}$ (see (3.21)) in (3.22) leads to the following difference scheme:

$$[\Lambda_n(\varphi, T_n)n]_i = \frac{A_i}{h}n_{i-1} + \frac{B_i}{h}n_{i+1} - \frac{C_i}{h}n_i = 0, \tag{3.23}$$

where

$$A_i = \frac{(D_n)_{i-1}}{h} f\left(\frac{\varphi_i - \varphi_{i-1}}{(T_n)_{i-1/2}}\right), \quad B_i = \frac{(D_n)_{i+1}}{h} f_1\left(\frac{\varphi_{i+1} - \varphi_i}{(T_n)_{i+1/2}}\right), \quad C_i = A_{i+1} + B_{i-1}. \tag{3.24}$$

Applying the above procedure to the non-stationary continuity equation on the non-uniform grid (3.1) we obtain

$$\frac{n_i^{l+1}-n_i^l}{\tau_{l+1}}=\frac{1}{h_i^*}[A_i^n n_{i-1}^{l+1}+B_i^n n_{i+1}^{l+1}-C_i^n n_i^{l+1}]+F_i, \tag{3.25}$$

where index l indicates the corresponding time-layer and the coefficients A_i^n, B_i^n and C_i^n are determined by the following formulae

$$A_i^n=\frac{D_n[(T_n)_{i-1}^{l+1}]}{h_i}f\left(\frac{\varphi_i^{l+1}-\varphi_{i-1}^{l+1}}{(T_n)_{i-1/2}^{l+1}}\right),\quad h_i=x_i-x_{i-1},\quad h_i^*=\frac{h_i+h_{i+1}}{2}, \tag{3.26}$$

$$B_i^n=\frac{D_n[(T_n)_{i+1}^{l+1}]}{h_{i+1}}f_1\left(\frac{\varphi_{i+1}^{l+1}-\varphi_i^{l+1}}{(T_n)_{i+1/2}^{l+1}}\right),\quad C_i^n=A_{i+1}^n+B_{i-1}^n. \tag{3.27}$$

Similarly, we derive the exponential difference scheme for the continuity equation for holes:

$$\frac{p_i^{l+1}-p_i^l}{\tau_{l+1}}=\frac{1}{h_i^*}[A_i^p p_{i-1}^{l+1}+B_i^p p_{i+1}^{l+1}-C_i^p p_i^{l+1}]+F_i, \tag{3.28}$$

where

$$A_i^p=\frac{D_p[(T_p)_{i-1}^{l+1}]}{h_i}f_1\left(\frac{\varphi_i^{l+1}-\varphi_{i-1}^{l+1}}{(T_p)_{i-1/2}^{l+1}}\right), \tag{3.29}$$

$$B_i^p=\frac{D_p[(T_p)_{i+1}^{l+1}]}{h_{i+1}}f\left(\frac{\varphi_{i+1}^{l+1}-\varphi_i^{l+1}}{(T_p)_{i+1/2}^{l+1}}\right),\ C_i^p=A_{i+1}^p+B_{i-1}^p \tag{3.30}$$

Remark 3.1. From the computational point of view, splitting algorithms are quite appealing in application to (2.33)–(2.38). However, in the application of such algorithms, a special care should be taken in approximating F in the right-hand side of (3.25) and (3.28). For example, for the approximation of F in the RHS of (3.25), the values of n^{l+1} can be found from the approximation of the Poisson equation. In this case the values of p have to be taken from the time layer l, which may significantly slow down the convergence. If convergence is satisfactory, then the computed value of n^{l+1} can be used for the approximation of F in the RHS of (3.28).

3.2.2. *Energy-balance equations*

Now we are in a position to consider approximation procedures for the most difficult equations in system (2.33)–(2.38), for energy balance equations. Our approach is different from that proposed in [33]. We recall that balance energy equations can be reduced to the forms amenable to computationally efficient schemes. For example, for the electron system we have

$$\begin{aligned}\frac{3}{2}\frac{\partial\mathcal{E}_n}{\partial t}=&\ \beta_n\frac{\partial^2[D_n(T_n)\mathcal{E}_n]}{\partial x^2}-(1+\beta_n)\frac{\partial}{\partial x}\left[\mu_n(T_n)\mathcal{E}_n\frac{\partial\varphi}{\partial x}\right]+\\&+\mu_n(T_n)\mathcal{E}_n\frac{\partial^2\varphi}{\partial x^2}+\mathcal{E}_n\frac{\mu_n(T_n)}{T_n}\left(\frac{\partial\varphi}{\partial x}\right)^2-\frac{\mathcal{E}_n(1-1/T_n)}{\tau_\omega^n(T_n)},\end{aligned} \tag{3.31}$$

where $\mathcal{E}_n=nT_n$. As was noted in [13], we can also obtain Equation (3.31) from the third equation of system (2.33) by using the following identity

$$\frac{\partial}{\partial x}[\mu_n(T_n)nT_n]\frac{\partial \varphi}{\partial x} = \frac{\partial}{\partial x}\left[(\mu_n(T_n)nT_n)\frac{\partial \varphi}{\partial x}\right] - \mu_n(T_n)nT_n\frac{\partial^2 \varphi}{\partial x^2}. \tag{3.32}$$

In order to construct an exponential difference scheme for Equation (3.31), we follow a procedure similar to that for the continuity equation. We introduce the change of variables $\mathcal{E}_n \to \mathcal{E}_n^{\text{new}}$ which is analogous to (3.16)

$$\mathcal{E}_n(x) = \mathcal{E}_n^{\text{new}}(x)\exp\left[\int_{x_0}^{x}\left\{\frac{1+\beta_n}{\beta_n}\frac{1}{T_n(\xi)}\frac{\partial \varphi(\xi)}{\partial \xi} - \frac{\partial \log D_n(\xi)}{\partial \xi}\right\}\mathrm{d}\xi\right], \tag{3.33}$$

where, as above, x_0 is an arbitrary number such that $x_0 < x$. As a result of transformations analogous to (3.17)–(3.22) we get the following difference scheme

$$\frac{3}{2}\frac{(\mathcal{E}_n)_i^{l+1} - (\mathcal{E}_n)_i^{l}}{\tau_{l+1}} = (\Lambda_{T_n}(\varphi^{l+1}, T_n^{l+1})\mathcal{E}_n^{l+1})_i, \tag{3.34}$$

where

$$(\Lambda_{T_n}(\varphi, T_n)\mathcal{E}_n)_i = \frac{1}{h_i^*}[\tilde{A}_i^n(\mathcal{E}_n)_{i-1} + \tilde{B}_i^n(\mathcal{E}_n)_{i+1} - \tilde{C}_i^n(\mathcal{E}_n)_i] -$$

$$-\left\{\mu_n[(T_n)_i]\varphi_{\bar{x}\hat{x},i} - \frac{\mu_n[(T_n)_i]}{(T_n)_i}(\varphi_{\check{x},i})^2 + \frac{1}{\tau_\omega^n[(T_n)_i]} - \frac{1}{\tau_\omega^n[(T_n)_i](T_n)_i}\right\}(\mathcal{E}_n)_i, \tag{3.35}$$

and the coefficients of this difference scheme are defined as follows

$$(\tilde{A})_i^n = \frac{\beta_n D_n[(T_n)_{i-1}^{l+1}]}{h_i} f\left(\frac{1+\beta_n}{\beta_n}\frac{\varphi_i^{l+1} - \varphi_{i-1}^{l+1}}{(T_n)_{i-1/2}^{l+1}}\right), \tag{3.36}$$

$$(\tilde{B})_i^n = \frac{\beta_n D_n[(T_n)_{i+1}^{l+1}]}{h_{i+1}} f_1\left(\frac{1+\beta_n}{\beta_n}\frac{\varphi_{i+1}^{l+1} - \varphi_i^{l+1}}{(T_n)_{i+1/2}^{l+1}}\right), \tag{3.37}$$

$$\tilde{C}_i^n = \tilde{A}_{i+1}^n + \tilde{B}_{i-1}^n. \tag{3.38}$$

We use standard difference-scheme notation (see [32] and references therein) and denote the second and the first central difference derivatives on the non-uniform grid (3.1) by

$$\varphi_{\bar{x}\hat{x},i} = \frac{1}{h_i^*}\left[\frac{\varphi_{i+1} - \varphi_i}{h_{i+1}} - \frac{\varphi_i - \varphi_{i-1}}{h_i}\right] \quad \text{and} \quad \varphi_{\check{x},i} = \frac{\varphi_{i+1} - \varphi_{i-1}}{2h_i^*},$$

respectively.

We construct the scheme analogous to (3.34)–(3.38) for the solution of the energy balance equation for the hole system:

$$\frac{3}{2}\frac{(\mathcal{E}_p)_i^{l+1} - (\mathcal{E}_p)_i^{l}}{\tau_{l+1}} = (\Lambda_{T_p}(\varphi^{l+1}, T_p^{l+1})\mathcal{E}_p^{l+1})_i, \tag{3.39}$$

where

$$(\Lambda_{T_p}(\varphi, T_p)\mathcal{E}_p)_i = \frac{1}{h_i^*}[\tilde{A}_i^p(\mathcal{E}_p)_{i-1} + \tilde{B}_i^p(\mathcal{E}_p)_{i+1} - \tilde{C}_i^p(\mathcal{E}_p)_i] -$$

$$\left\{\mu_p[(T_p)_i]\varphi_{\bar{x}\hat{x},i} - \frac{\mu_p[(T_p)_i]}{(T_p)_i}(\varphi_{\check{x},i})^2 + \frac{1}{\tau_\omega^p[(T_p)_i]} - \frac{1}{\tau_\omega^p[(T_p)_i](T_p)_i}\right\}(\mathcal{E}_p)_i, \tag{3.40}$$

and the coefficients of this difference scheme are defined as follows

$$(\tilde{A})_i^p = \frac{\beta_p D_p[(T_p)_{i-1}^{l+1}]}{h_i} f_1\left(\frac{1+\beta_p}{\beta_p}\frac{\varphi_i^{l+1} - \varphi_{i-1}^{l+1}}{(T_p)_{i-1/2}^{l+1}}\right), \tag{3.41}$$

$$(\tilde{B})_i^p = \frac{\beta_p D_p[(T_p)_{i+1}^{l+1}]}{h_{i+1}} f\left(\frac{1+\beta_p}{\beta_p}\frac{\varphi_{i+1}^{l+1} - \varphi_i^{l+1}}{(T_p)_{i+1/2}^{l+1}}\right), \tag{3.42}$$

$$\tilde{C}_i^p = \tilde{A}_{i+1}^p + \tilde{B}_{i-1}^p, \quad \text{with} \quad f_1(x) = f(-x). \tag{3.43}$$

3.2.3. *Monotonicity and stability*

All coefficients of difference schemes (3.34)–(3.38), (3.39)–(3.43) preserve the positiveness property, namely it is easy to see that

$$\tilde{A}_i^n, \tilde{B}_i^n, \tilde{C}_i^n > 0, \quad \tilde{A}_i^p, \tilde{B}_i^p, \tilde{C}_i^p > 0. \tag{3.44}$$

Unfortunately, this fact cannot guarantee monotonicity of the constructed schemes. The sign of the functions near $(\mathcal{E}_n)_i$ and $(\mathcal{E}_p)_i$ in the expressions (3.35) and (3.40) cannot be defined *a priori*, and in the general case it may change. For example, we can not claim that on each time step and in each space point x_i the conditions $n_i \geq 0$, $T_n - 1 \geq 0$ (and, respectively, $p_i \geq 0$, $T_p - 1 \geq 0$) will always be satisfied. We can only claim that sufficient conditions for monotonicity given in the Karetkina lemma will be satisfied if in addition to (3.44) the conditions $(G_n)_i \geq 0$, $(G_p)_i \geq 0$ are also satisfied. In the context of the approximation of energy balance equations these conditions lead to the following inequalities (see also [13, 14])

$$(G_n)_i \& = \& - \mu_n[(T_n)_i]\varphi_{\bar{x}\hat{x},i} - \frac{\mu_n[(T_n)_i]}{(T_n)_i}(\varphi_{\check{x},i})^2 + \frac{(T_n)_i - 1}{\tau_\omega^n[(T_n)_i](T_n)_i} + \frac{1{\cdot}5}{\tau_{l+1}} \geq 0, \tag{3.45}$$

$$(G_p)_i \& = \& \mu_p[(T_p)_i]\varphi_{\bar{x}\hat{x},i} - \frac{\mu_p[(T_p)_i]}{(T_p)_i}(\varphi_{\check{x},i})^2 + \frac{(T_p)_i - 1}{\tau_\omega^p[(T_p)_i](T_p)_i} + \frac{1{\cdot}5}{\tau_{l+1}} \geq 0. \tag{3.46}$$

It is easy to verify [13] that inequalities (3.45) and (3.46) will be satisfied if

$$\tau < 1{\cdot}5/(E^*)^2. \tag{3.47}$$

In the stationary case conditions (3.45) and (3.46) can be simplified to

$$\begin{aligned}(G_n)_i = {} & -\mu_n[(T_n)_i]\varphi_{\bar{x}\hat{x},i} - \mu_n[(T_n)_i](\varphi_{\check{x},i})^2/(T_n)_i + \\ & + ((T_n)_i - 1)/(\tau_\omega^n[(T_n)_i](T_n)_i) \geq 0,\end{aligned} \tag{3.48}$$

$$\begin{aligned}(G_p)_i = {} & \mu_p[(T_p)_i]\varphi_{\bar{x}\hat{x},i} - \mu_p[(T_p)_i](\varphi_{\check{x},i})^2/(T_p)_i + \\ & + ((T_p)_i - 1)/(\tau_\omega^p[(T_p)_i](T_p)_i) \geq 0.\end{aligned} \tag{3.49}$$

and will be satisfied when (3.14) holds.

It is clear that in the case of large gradients of the potential (*i.e.* in high electric fields) both conditions, (3.14) and (3.47) are very restrictive computationally. Hence, when modelling non-stationary processes in non-highly doped semiconductors, purely explicit schemes may become a competitive alternative to the proposed schemes due to their minimal computational cost per time-step. However, such explicit schemes typically require the time-step to be of order $1/\max_{0\le x\le 1} N$ (i.e. $\tau = \mathcal{O}(1/\max_{0\le x\le 1} N)$) (see, for example, [35]). This causes problems when in the RHS of the Poisson equation we have a large dopant concentration N [36]. The use of purely implicit schemes cannot resolve all difficulties; firstly, because in the general case such schemes cannot guarantee absolute stability of the numerical algorithm (subject to the approximation of F), and, secondly, the computational cost for their numerical realization on each time-step integration substantially increases, especially for devices with two types of carriers. Therefore, one of the most promising directions in the development of efficient numerical schemes in semiconductor device theory lies with semi-implicit schemes.

4. Algorithmic realizations of semi-implicit schemes for quasi-hydrodynamic models

Semi-implicit schemes have been extensively applied in modelling semiconductor devices with drift-diffusion types of models (see [37, 35, 13] and references therein). Basic ideas of their algorithmic implementation are typically connected either with Mock's scheme, Polsky–Rimshans's scheme, or a self-consistent scheme. We describe these ideas below.

1. In the Mock scheme [37] the potential is determined from the continuity equation for the *total current* rather than from the Poisson equation as in standard procedures. This scheme is not conservative and contains the disbalance term of the numerical nature of the order $\mathcal{O}(\tau)$ ($\tau = \max_{j=1,\dots,K-1} \tau_j$). The presence of this term deteriorates the scheme accuracy in practice when the time-step increases in spite of the absolute stability of this scheme for the linear case.
2. In the Polsky-Rimshans scheme [35] the potential is sought in two stages similar to prediction-correction procedures. First, from the continuity equation for the *total current*, we find a prediction $\varphi^{l+1/2}$. Then we correct it using the Poisson equation written for the time layer $l+1$. The Poisson equation on each step is solved with the accuracy $\mathcal{O}(\tau^3)$.
3. In the *self-consistent scheme* proposed in [13] we determine the potential on the $(l+1)$-time layer through n^l and p^l. Then we compute n^{l+1} and p^{l+1} using φ^{l+1}. On the next step we determine the potential using the purely implicit scheme for the Poisson equation. In doing so, we find the values n^{l+1} and p^{l+1} (for example, $n^{l+1} = n^l + \tau n_t$) from the semi-implicit scheme for the continuity equation which in the homogeneous case has the following form

$$(n^{l+1} - n^l)/\tau_{l+1} = [D_n((T_n)^l)n^l]_{\bar{x}\hat{x}} - (a^l \varphi_{\bar{x}}^{l+1})_{\hat{x}}, \tag{4.1}$$

where

$$a_i = (n_{i-1}\mu_n((T_n)_{i-1}) + n_i \mu_n((T_n)_i))/2. \tag{4.2}$$

From a computational point of view, the last scheme is very attractive if we use the central-difference approximation for the current density. Indeed, in this case we get a linear equation with respect to φ^{l+1}. Otherwise, if we use the exponential scheme, we have to solve a nonlinear

equation in order to determine φ^{l+1}. Similar to Mock's scheme, in this self-consistent scheme we have a disbalance term $\mathcal{O}(\tau)$, which in the stationary-regime limit tends to zero.

4.1. SIMPLEST SEMI-IMPLICIT SCHEMES AND DECELERATION OF CONVERGENCE

In recent years the interest to the application of semi-implicit schemes to non-local models has been increasing. One of the simplest algorithms of this type is what is known as the 'relaxation-to-the-stationary-regime' method, which is widely used for the solution of drift-diffusion models (see references in [14]). In order to determine unknowns $(\varphi, n, T_n, p, T_p)$ of system (2.33)–(2.38) with this method, all equations are solved alternately, but for the solution of each equation an implicit scheme (with respect to the leading variable of that equation) is applied. Conceptually, this algorithm is a nonlinear Gummel-type algorithm. If devices are modelled with the DDM, this algorithm typically provides the user with good convergence when applied to devices with low and middle levels of doping. Unfortunately, for high-doping-level devices convergence of this algorithm may seriously deteriorate. For example, modelling bipolar transistors with this method, we observe that for high forward voltages the concentration of majority carriers approaches the concentration of minority carriers in the vicinity of junctions. As a result we have a strong coupling between concentrations of both types of carriers through the potential function (partly induced by the requirement of the quasi-neutrality at the boundaries). This causes difficulties in numerical simulation of such devices one of which is slow convergence. It is possible that physical reasons for the deceleration of the convergence of simplest semi-implicit algorithms may be different from the described above. For example, it is well-known that the coupling between electrostatic potential and carrier concentrations in MOS-transistors that work in strong inversion regimes also increases. Other sources of coupling in modelling transient processes may be caused by the bias current. In all such situations we may expect a decrease in the rate of convergence of simplest semi-implicit algorithms.

Ultimately, the roots of the described computational difficulties lie with the quality of approximation of recombination/generation/ionisation terms. We recall that in the standard Gummel algorithm, all values of concentrations are taken from the previous time-layer, that may be unsatisfactory for many problems. However, from the computational viewpoint it is very attractive to apply a Gummel-type algorithm to the solution of the QHDM (2.33)–(2.38), where we have to solve a system of five, rather that three (as in the DDM), strongly coupled nonlinear equations. Although for the last few decades attempts have been made to modify the Gummel algorithm in order to include a special treatment of the recombination/generation/ionisation terms, efficiency of the coupling of discretised equations in the Gummel-type algorithms critically depend on the type of modelling device and the strength of applied electric field. In this paper such a coupling is performed through the Boltzmann statistics and a Newton-type solver. In order to clarify the idea of such an algorithm we recall the connection between the *mixed basis* (n, p, φ), and the *hybrid basis*, $(\Phi_n, \Phi_p, \varphi)$ in the case of the classical drift-diffusion model:

$$n = n_{\text{int}} \exp\left[\frac{\gamma \Delta E_G + \varphi - \varphi_n}{\varphi_T}\right], \quad p = n_{\text{int}} \exp\left[\frac{(1-\gamma)\Delta E_G + \varphi_p - \varphi}{\varphi_T}\right], \tag{4.3}$$

where ΔE_G is the effective bandgap narrowing [5], n_{int} is the intrinsic concentration, φ_T is the thermal potential, and γ is the experimentally measured parameter that takes into account the asymmetry factor. If we set $\gamma = 0{\cdot}5$, then formulae (4.3) will simplify to (see formula (2.27))

$$n = n_{ie} \exp\left[\frac{\varphi - \varphi_n}{\varphi_T}\right], \quad p = n_{ie} \exp\left[\frac{\varphi_p - \varphi}{\varphi_T}\right] \tag{4.4}$$

or, finally, to

$$n = n_{ie} \exp\left(\frac{\varphi}{T}\right) \Phi_n, \quad p = n_{ie} \exp\left(-\frac{\varphi}{T}\right) \Phi_p, \tag{4.5}$$

where $\Phi_n = \exp(-\varphi_n/T)$ and $\Phi_p = \exp(\varphi_p/T)$ are the Fermi quasi-levels, and the temperature, T, is taken in energy units (multiplied by the factor k_b/q [5]).

4.2. Conditionally coupled algorithm of the first order

The Fermi-quasi-level representation (4.5) can be effectively implemented into a conditionally coupled semi-implicit algorithm (in a sense that quantities φ, n and p can be made coupled via the Boltzmann statistics). Here we propose a generalization of this algorithm to the quasi-hydrodynamic model and describe it on the example of the one-type carrier system (further details can be found in [38]).

Algorithm 4.1.

1. We choose initial approximations for φ, n, T_n and calculate F;
2. We sequentially solve the continuity and energy balance equations for computed values of φ^m and F^{m+1} (as before, m is the index of external iterations with $m+1$ being current); for the solution of the energy balance equation we organise the following 'internal' coupling procedure:
   (a) assuming the steadiness of Fermi quasi-levels we solve the discretized balance energy equation with respect to the corrections $\delta(T_n)_{k+1}^{m+1}$;
   (b) we compute the values of temperature on the *current internal iteration*, $(T_n)_{k+1}^{m+1} = (T_n)_k^{m+1} + \delta(T_n)_{k+1}^{m+1}$ and predict the values of concentration for just computed new values of temperature using the formula (assuming steadiness of Fermi quasi-levels):
   $$n_{k+1}^{m+1} = n_{ie}\left(\frac{n_k^{m+1}}{n_{ie}}\right)^{\frac{(T_n)_k^{m+1}}{(T_n)_{k+1}^{m+1}}}; \tag{4.6}$$
   (c) we set $(T_n)_k^{m+1} = (T_n)_{k+1}^{m+1}$, $n_k^{m+1} = n_{k+1}^{m+1}$ and go to a new *internal iteration* by setting $k := k+1$ and returning to (a); such internal iterations are performed until the given accuracy is achieved or the given number of times;
3. We perform a new 'incomplete' external iteration for the computational block (n, T_n) (*i.e.* step 2); such 'incomplete external' iterations are performed either up to the complete convergence or given number of times;
4. Then we solve the discretized Poisson equation
$$\begin{aligned} &F_1(\varphi_{i-1}^{m+1}, \varphi_i^{m+1}, \varphi_{i+1}^{m+1}) = \frac{\varphi_{i+1}^{m+1} - \varphi_i^{m+1}}{h_{i+1}} - \frac{\varphi_i^{m+1} - \varphi_{i-1}^{m+1}}{h_i} - \\ &h_i^*\left[n_i^m \exp\left(\frac{\varphi_i^{m+1} - \varphi_i^m}{(T_n)_i^m}\right) - p_i^m \exp\left(\frac{-\varphi_i^{m+1} + \varphi_i^m}{(T_n)_i^m}\right) - N\right] = 0, \end{aligned} \tag{4.7}$$

using the Newton method (details are in [38]) and assuming the relationships $n^{l+1} = n^l \exp((\varphi^{l+1} - \varphi^l)/(T_n)^l)$, and $p^{l+1} = p^l \exp((\varphi^l - \varphi^{l+1})/(T_n)^l)$ (which are valid under constant temperatures); in other words, we organise a one more cycle of internal iterations:

$$\left(\frac{\partial F_1}{\partial \varphi_{i-1}}\right)\Bigg|_m \delta\varphi_{i-1}^{m+1} + \left(\frac{\partial F_1}{\partial \varphi_i}\right)\Bigg|_m \delta\varphi_i^{m+1} + \left(\frac{\partial F_1}{\partial \varphi_{i+1}}\right)\Bigg|_m \delta\varphi_{i+1}^{m+1} = -F_1\Big|_m, \tag{4.8}$$

where $\delta\varphi_k^{m+1} = \varphi_k^{m+1} - \varphi_k^m$;

5. Steps 2–4 complete one external iteration; external iterations are performed until the convergence is reached.

Remark 4.1. 'Incomplete' external iterations for the computational block (n, T_n) couples discretised versions of continuity and energy balance equations. The problem of the increase of the convergence rate for these iterations is addressed with the prognostic formula (4.6) (obtained under the steadiness of Fermi quasi-levels on the current iteration).

4.3. Coupling procedures using the Boltzmann statistics

'Incomplete' external iterations for the computational block (n, T_n), employed in Algorithm 4.1, couple continuity and energy balance equations by the prognostic formula (4.6). This increases the rate of algorithm convergence. The prognostic formula (4.6) is formally obtainable from (4.5) under the assumption of steadiness of Fermi quasi-levels. This formula is applied only to the current iteration and is modified on the next iteration when new values of the temperature become available. In order to derive this formula we use the idea of (4.5) assuming that

$$n = n_{ie} \exp\left(\frac{\varphi - \varphi_n}{T_n}\right), \tag{4.9}$$

where T_n is the electron temperature in energy units. Hence, for the k^{th} iteration we have that

$$\log\left(\frac{n_k}{n_{ie}}\right) = \frac{\varphi - \varphi_n}{(T_n)_k} \quad \text{or} \quad \varphi - \varphi_n = (T_n)_k \log\left(\frac{n_k}{n_{ie}}\right). \tag{4.10}$$

Using (4.9) and (4.10) and assuming the constant value of the potential over two subsequent internal iterations, we obtain that

$$n_{k+1} = n_{ie} \exp\left[\frac{\varphi - \varphi_n}{(T_n)_{k+1}}\right] = n_{ie} \exp\left[\frac{(T_n)_k}{(T_n)_{k+1}} \ln\left(\frac{n_k}{n_{ie}}\right)\right] = n_{ie}\left(\frac{n_k}{n_{ie}}\right)^{\frac{(T_n)_k}{(T_n)_{k+1}}}. \tag{4.11}$$

Formula (4.11) allows us to couple the values of concentrations over two subsequent internal iterations through the values of temperature. Using (4.11) as a predictor, we may rewrite the discretized energy balance equation in the form

$$F_2\left((T_n)_{i-1}^{m+1},(T_n)_i^{m+1},(T_n)_{i+1}^{m+1}\right)=\left\{\tilde{A}_i(n_{ie})_{i-1}\left(\frac{n_{i-1}^m}{(n_{ie})_{i-1}}\right)^{\frac{(T_n)_{i-1}^m}{(T_n)_{i-1}^{m+1}}}(T_n)_{i-1}^{m+1}+\right.$$

$$\left.+\tilde{B}_i(n_{ie})_{i+1}\left(\frac{n_{i+1}^m}{(n_{ie})_{i+1}}\right)^{\frac{(T_n)_{i+1}^m}{(T_n)_{i+1}^{m+1}}}(T_n)_{i+1}^{m+1}-\tilde{C}_i(n_{ie})_i\left(\frac{n_i^m}{(n_{ie})_i}\right)^{\frac{(T_n)_i^m}{(T_n)_i^{m+1}}}(T_n)_i^{m+1}\right\}-$$

$$-\left[-\mu\varphi_{\check{x}\hat{x}}-\frac{\mu}{(T_n)_i^{m+1}}(\varphi_{\check{x}})^2+\frac{(T_n)_i^{m+1}-1}{\tau_\omega^n(T_n)_i^{m+1}}\right]\times(n_{ie})_i\left(\frac{n_i^m}{(n_{ie})_i}\right)^{\frac{(T_n)_i^m}{(T_n)_i^{m+1}}}(T_n)_i^{m+1}=0, \tag{4.12}$$

and the linearization procedure reduces this equation to the equation

$$\left(\frac{\partial F_2}{\partial(T_n)_{i-1}}\right)\Bigg|_m\delta(T_n)_{i-1}^{m+1}+\left(\frac{\partial F_2}{\partial(T_n)_i}\right)\Bigg|_m\delta(T_n)_i^{m+1}+$$

$$\left(\frac{\partial F_2}{\partial(T_n)_{i+1}}\right)\Bigg|_m\delta(T_n)_{i+1}^{m+1}=-F_2\Big|_m, \tag{4.13}$$

where $\delta(T_n)_j^{m+1}=(T_n)_j^{m+1}-(T_n)_j^m,\ \ j=i-1,i,i+1.$

4.4. Conditionally coupled algorithm of the second order

Algorithm 4.1 provides a computationally efficient tool for modelling a wide range of semiconductor devices. However, its main drawback lies with the assumption of constancy of the potential over two subsequent internal iterations. This assumption may not be fulfilled in the case when the coupling between the continuity and the energy balance equations is sufficiently strong. For example, difficulties may arise in the application of Algorithm 4.1 to the modelling of such devices as microwave PIN diodes that work in reverse-bias regimes [39, 40, 41, 42]. Clearly that in such cases Algorithm 4.1 has to be modified to account for an additional computational block for (p, T_p). Moreover, if blocks (n, T_n) and (p, T_p) are to be treated sequentially, then a coupling between them has to be implemented (it can be done, for example, through the computational block solving the Poisson equation). When the strong coupling between continuity and energy balance equations is an intrinsic feature of the problem, we propose the *conditionally coupled algorithm of the second order,* which we refer to as Algorithm 4.2. Its most noticeable difference from Algorithm 4.1 is the absence of the iterative cycle inside of the block (n, T_n) (*i.e.* 'incomplete' external iterations). Details of the solution strategy with the second order conditionally coupled algorithm can be found in [38]. Here we only notice that a new QHDM computational block in Algorithm 4.2 contains the following steps:

1. Computation of values of carrier temperatures, T_n and T_p using exponential difference schemes (3.34)–(3.38) and (3.39)–(3.43), respectively;
2. Computation of concentrations, n and p, taken into account computed values of T_n and T_p using exponential difference schemes (3.25)–(3.27) and (3.28)–(3.30), respectively;
3. Recalculation of the potential φ taken into account the computed values of n, p, T_n, T_p.

The prediction stage for Algorithm 4.2 is realised by solving the classical drift-diffusion model.

Remark 4.2. Although the first order conditionally coupled algorithm described in Section 4.2 may meet serious computational challenges when applied to strongly coupled problems, after sufficient number of iterations it will typically provides the user with a plausible qualitative picture of the main characteristics of devices. However, this simplified approach may not be adequate when the investigation is focused at the physical phenomena in semiconductor plasma, rather than at output characteristics of the device.

5. Initial approximations, stopping criteria and the solution of linearized problems

In order to guarantee the convergence of the algorithms described in Sections 3 and 4, we have to take special efforts in constructing an appropriate initial approximation. As the initial approximation for the QHD computational block in Algorithm 4.2 we use the output from the solution of the DDM. Since the later model also requires an initial approximation, we use the assumption of quasi-neutrality

$$\rho = n - p - N = 0 \tag{5.1}$$

and thermal equilibrium

$$pn = n_{ie}^2 \tag{5.2}$$

in order to construct such an approximation. Using (5.1), (5.2) and assuming that $n = n_{ie} \exp((\tilde{U} - \varphi)/T_n)$, $p = n_{ie} \exp((\varphi - \tilde{U})/T_n)$ we determine the initial approximation for the potential as follows

$$\varphi = \tilde{U} + T_n \text{sign}(N) \log \left[\frac{|N|}{2n_{ie}} + \sqrt{\left(\frac{N}{2n_{ie}}\right)^2 + 1} \right] \approx \tilde{U} + T_n \text{sign}(N) \log \left(\frac{|N|}{n_{ie}} \right). \tag{5.3}$$

Then, the initial approximations for carrier concentrations can be found from the formulae

$$n = n_{ie} \exp\left(\frac{\varphi}{T_n}\right), \quad p = n_{ie} \exp\left(-\frac{\varphi}{T_p}\right), \tag{5.4}$$

that couples concentrations and the potential in the equilibrium case. As the initial approximations for carrier temperatures we assume their equality to the lattice temperature.

Remark 5.1. Strictly speaking the initial approximations for the carrier temperatures have to be computed, because the assumption of their equality to the lattice temperature may be dubious in simulation of some semiconductor devices such as reverse-bias PIN microwave diodes [40, 42]. However, in our numerical experiments we did not observe a deviation of the computed initial-temperature from the given equilibrium values for more than 8% (this was observed only in the neighbourhoods of p-n junctions).

Modelling semiconductor devices in high electric fields with the proposed schemes may lead to computational overflow due to the exponential character of these schemes. In order to avoid it, a special treatment of the Bernoulli functions $f(x)$, $f_1(x)$ and their derivatives has to be implemented in the case when $x \to 0$ (details of this treatment in our experiments can be found in [38]).

The choice of stopping criteria for numerical algorithms is another important issue in modelling semiconductor devices. In our code we use the following criterion

$$\epsilon_{\varpi}^{*} = \begin{cases} \max\limits_{i} |\varpi_i^{k+1} - \varpi_i^{k}|, & |\varpi_i^{k+1}| \le 1, \\ \max\limits_{i} \dfrac{|\varpi_i^{k+1} - \varpi_i^{k}|}{|\varpi_i^{k+1}|}, & |\varpi_i^{k+1}| > 1, \end{cases} \tag{5.5}$$

where ϖ is the corresponding function, for example, φ, n, T_n, etc. Other criteria may also be chosen (see [13] and references therein). For example, in the one-dimensional case we may estimate the error of the conservative property of the total current that flows through the endpoints of the structure (*i.e.* endpoints of the interval $[0, 1]$). An inconvenience of this criterion becomes obvious for non-stationary problems where this quantity has to be checked at each moment of the transient process and the bias current has to be taken into account.

Finally, we consider technical issues of the implementation of computational blocks (n, T_n) and (p, T_p) connected with the solution of linearized systems of two coupled equations, continuity equation and the energy balance equation. For the sake of simplicity, we consider the stationary electron system without the recombination term

$$\Phi_{1i}\left(n_{i,i\pm1}^{m+1}, (T_n)_{i,i\pm1}^{m+1}\right) = 0, \quad \Phi_{2i}\left(n_{i,i\pm1}^{m+1}, (T_n)_{i,i\pm1}^{m+1}\right) = 0, \tag{5.6}$$

where

$$\Phi_{1i} = \frac{1}{h_i^{*}}\left(A_i n_{i-1}^{m+1} + B_i n_{i+1}^{m+1} - C_i n_i^{m+1}\right), \tag{5.7}$$

$$\begin{aligned} \Phi_{2i} = {} & \frac{1}{h_i^{*}}\left(\tilde{A}_i(\mathcal{E}_n)_{i-1}^{m+1} + \tilde{B}_i(\mathcal{E}_n)_{i+1}^{m+1} - \tilde{C}_i(\mathcal{E}_n)_i^{m+1}\right) - \\ & \left[-\mu_n \varphi_{\bar{x}\hat{x},i} - \frac{\mu_n}{(T_n)_i}(\varphi_{\breve{x}})^2 + \frac{(T_n)_i^{m+1} - 1}{\tau_{\omega}^{n}(T_n)_i^{m+1}}\right](\mathcal{E}_n)_i^{m+1}, \end{aligned} \tag{5.8}$$

$m + 1$ denotes the current iteration of Newton's iterative process on which we determine corrections to the solution, $\delta\varpi_k^{m+1} = \varpi_i^{m+1} - \varpi_i^{m}$, and i is the index of the space grid point. Coefficients A_i, B_i, C_i and $\tilde{A}_i$, $\tilde{B}_i$, $\tilde{C}_i$ in (5.7)–(5.8) are determined by formulae (3.26)–(3.27) (or (3.29)–(3.30)) and (3.36)–(3.38) (or (3.41)–(3.43)), respectively.

The Equations (5.6), linearised with respect to correction terms, have the following form:

$$\begin{cases} \displaystyle\sum_{k=i,i\pm1}\left[\left(\frac{\partial\Phi_{1i}}{\partial n_k}\right)\Big|_m \delta n_k^{m+1} + \left(\frac{\partial\Phi_{1i}}{\partial (T_n)_k}\right)\Big|_m \delta (T_n)_k^{m+1}\right] = -\Phi_{1i}\Big|_m, \\ \displaystyle\sum_{k=i,i\pm1}\left[\left(\frac{\partial\Phi_{2i}}{\partial n_k}\right)\Big|_m \delta n_k^{m+1} + \left(\frac{\partial\Phi_{2i}}{\partial (T_n)_k}\right)\Big|_m \delta (T_n)_k^{m+1}\right] = -\Phi_{2i}\Big|_m, \end{cases} \tag{5.9}$$

where derivatives in (5.9) are computed by the following formulae

$$\frac{\partial\Phi_{1i}}{\partial n_{i-1}} = \frac{1}{h_i^{*}}A_i, \quad \frac{\partial\Phi_{1i}}{\partial n_{i+1}} = \frac{1}{h_i^{*}}B_i, \quad \frac{\partial\Phi_{1i}}{\partial n_i} = -\frac{1}{h_i^{*}}C_i, \quad \frac{\partial\Phi_{1i}}{\partial (T_n)_{i-1}} = \frac{1}{h_i^{*}}n_{i-1}^{m+1}(A_i)',$$

$$\frac{\partial\Phi_{1i}}{\partial (T_n)_{i+1}} = \frac{1}{h_i^{*}}n_{i+1}^{m+1}(B_i)', \quad \frac{\partial\Phi_{1i}}{\partial (T_n)_i} = -\frac{1}{h_i^{*}}n_i^{m+1}(C_i)',$$

$$\frac{\partial \Phi_{2i}}{\partial n_{i-1}} = \frac{1}{h_i^*}\tilde{A}_i(T_n)_{i-1}^{m+1}, \quad \frac{\partial \Phi_{2i}}{\partial n_{i+1}} = \frac{1}{h_i^*}\tilde{B}_i(T_n)_{i+1}^{m+1},$$

$$\frac{\partial \Phi_{2i}}{\partial n_i} = -\frac{1}{h_i^*}\tilde{C}_i(T_n)_i^{m+1} - \left[-\mu_n\varphi_{\bar{x}\hat{x},i}^{m+1} - \frac{\mu_n}{(T_n)_i^{m+1}}(\varphi_{\hat{x}}^{m+1})^2 + \frac{(T_n)_i^{m+1}-1}{\tau_\omega^n(T_n)_i^{m+1}}\right](T_n)_i^{m+1},$$

$$\frac{\partial \Phi_{2i}}{\partial (T_n)_{i-1}} = \frac{1}{h_i^*}n_{i-1}^{m+1}\left[\tilde{A}_i + (T_n)_{i-1}^{m+1}(\tilde{A}_i)'\right], \quad \frac{\partial \Phi_{2i}}{\partial (T_n)_{i+1}} = \frac{1}{h_i^*}n_{i+1}^{m+1}\left[\tilde{B}_i + (T_n)_{i+1}^{m+1}(\tilde{B}_i)'\right],$$

$$\frac{\partial \Phi_{2i}}{\partial (T_n)_i} = -\frac{1}{h_i^*}n_i^{l+1}\left[\tilde{C}_i + (T_n)_i^{m+1}(\tilde{C}_i)'\right] + \mu_n\varphi_{\bar{x}\hat{x},i}^{m+1}n_i^{m+1} - \frac{n_i^{m+1}}{\tau_\omega^n}.$$

As a result, we have a large sparse system of linear equations with the matrix $2(N+1) \times 2(N+1)$ that has a block-tridiagonal structure. More precisely, it consists of 4 blocks each of which has $(N+1)+2N$ non-zero elements. Hence, in the most general case the total number of non-zero matrix entries cannot exceed $12N+4$.

Such systems of linear equations may be effectively solved using direct methods that use the technology of sparse matrices (see [43] and references therein). In our experiments we used two packages for the solution of arising sparse systems. In the first package the data was packed into coupled lists. The program for the solution contains the algorithm for ordering and minimization of the number of non-zero elements, algorithms of symbolic and numerical factorization which are based on the representation of sparse matrices given by Singhal and Vlach (see references in [43]). The second package was based on the *sparse solver* presented in [44].

One of the main features of mathematical problems in semiconductor device theory is a large scattering of unknown quantities, the difficulty that has to be dealt with even for the dimensionalised systems of PDEs. Since classical iterative methods require at least estimates of spectrum boundaries for the guaranteed convergence, they may not be good candidates in the context of semiconductor device modelling. It is more appropriate to apply methods that do not require explicit knowledge of parameters that estimate the matrix spectrum. In this sense variational-type methods such as Kreig's method or methods based on biconjugate gradients seem to be very promising. However, when these methods are applied, the procedure for preconditioning requires special attention [45].

6. Numerical experiments

The constructed numerical schemes have been applied to modelling physical effects in electron-hole plasma of semiconductors.

As an example we present results on the modelling of a $n^+ - n - n^+$ ballistic diode. This device is often used to model the $n^+ - n - n^+$ channel in MEtal-Semiconductor Field-Effect Transistors (MESFET) and the modelling of this device is considered by a number of authors as a benchmark example [10, 17, 9]. The simulated diode is a unipolar device with $n^+ - n - n^+$ structure that has a central n region of length 0·4 μm bounded by two n^+ regions of length 0·1 μm each. The n^+ regions are doped at density $N = 5 \times 10^{17}$ cm^{-3} while the n region is doped at $N = 2 \times 10^{15}$ cm^{-3} (see Figure 1 (left)).

On Figure 1 (right) we give the electron-concentration distribution calculated for the applied biases 0·1 V, 0·5 V and 1·0 V. As one expects, for these applied voltages the concentration profile (see Figure 1 (right)) is similar to the shape of the doping distribution in the

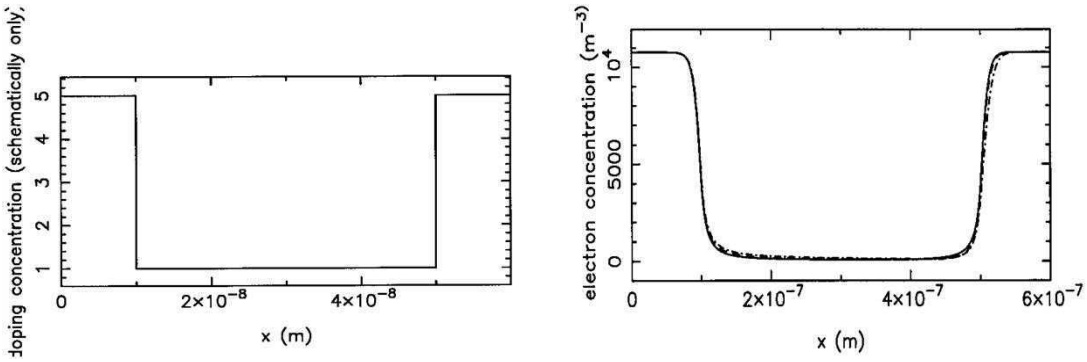

Figure 1. Dopant distribution in the ballistic diode (left) and concentration of electrons in normalised units (right).

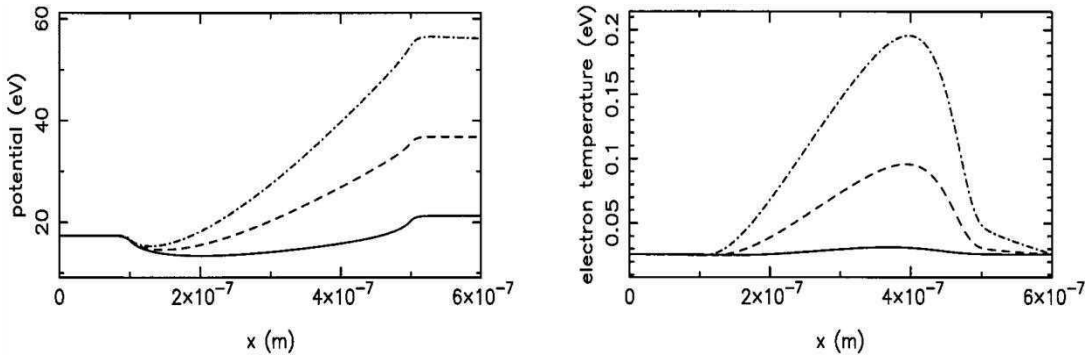

Figure 2. Electric potential (left) and electron-temperature distribution (right).

structure. When the bias increases we observe a drop in the concentration values in the right n^+ region. The electric potential as a function of the position at 0·1, 0·5 and 1·0 V bias is given on Figure 2 (left).

As follows from (2.38), we applied the bias at the right contact, while the left contact was grounded. This plot demonstrates electric field distribution over the semiconductor structure with the electron flow from left to right (note that this is opposite to the direction chosen in [9]). One can notice a slight drop in the electric field near the junction $n^+ - n$ (this drop leads to a slight 'cooling' of electrons reported, for example, in [9] and assigned to a strong diffusion effect opposite to the carrier motion) and its maximum value near the junction $n - n^+$.

Figure 2 (right) shows the electron-temperature distribution (presented in the energy units for $T_l = 0{\cdot}025$ eV) calculated for the three applied biases. A shift of the temperature peak to the right as the applied bias increases is clearly visible on this plot. This is in agreement with computational results obtained by other authors [17, 9].

The velocity profile, computed according to formula (2.32), is presented in Figure 3 (left).

We note that such a quantitative velocity overshoot cannot be identified with the classical drift-diffusion model. Finally we display the ratio v_n/c where c is the sound speed computed by the formula

$$c = \sqrt{\gamma T_n/m_n} \quad \text{with} \quad \gamma = \tfrac{5}{3}. \tag{6.1}$$

This ratio, presented on Figure 3 (right), is sometimes refer to as the Mach number [10].

In [42] we reported some computational results on the simulation of physical processes in the PIN diodes used extensively for microwave control applications such as microwave

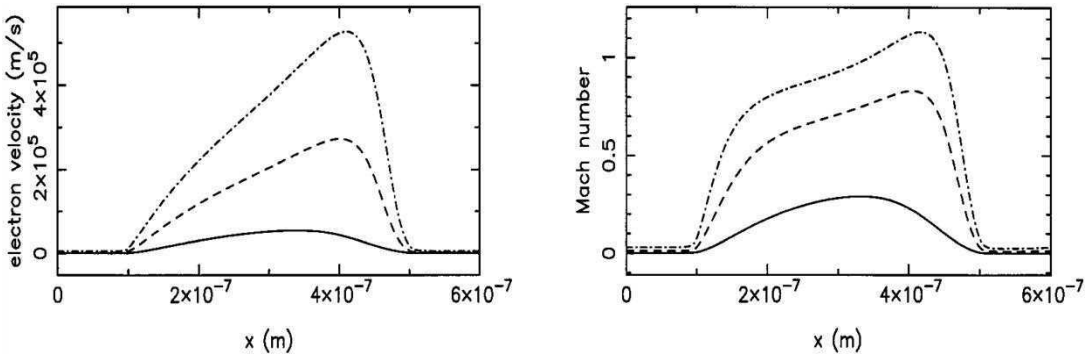

Figure 3. Electron velocity (left) and the Mach number in the ballistic diode (right).

switches and for electronically steered phased-array antennas [40]. Due to power-handling-capability requirements, the analysis of such devices has to include thermal effects. The results were obtained for a silicon $p^+ - i - n^+$ diode structure in the stationary case. Such diodes may be created by the molecular-beam epitaxy and are widely used as microwave switches. They may work in both direct and reverse biased regimes. Under forward-bias conditions, these devices exhibit a very low RF resistance, a higher conductivity and a larger breakdown compared to standard PN diodes; whereas under reverse-bias conditions they exhibit a very low constant capacitance. The latter case is also very interesting because the singular-perturbation-scaling technique can typically describe $p - n$ junctions under reverse biasing conditions only under small values of reverse biases. A singular perturbation analysis of reverse-biased semiconductor diodes for large applied biases is a difficult problem even in the case of the classical Van Roosbroek drift-diffusion model and is a topic of active research [39, 41]. These devices require further theoretical analysis and computational experiments using different models described in [5].

7. Conclusions and future directions

In this paper we considered a hierarchy of semiconductor device models using the relaxation time concept. In order to describe nonlocal, non-equilibrium processes in electron-hole semiconductor plasma, we focused on the class of quasi-hydrodynamic models which provides a reasonable compromise between kinetic, hydrodynamic and drift-diffusion models. These models belong to a wider class of nonlocal models which require the development of effective numerical procedures.

For the investigation of non-equilibrium and non-local processes in semiconductors we proposed exponential monotone schemes and developed their algorithmic realisations. The issues of the approximation of fluxes for these models, the problems of computational stability of the algorithmic realisations of the proposed schemes as well as their application to the modelling of transport phenomena in semiconductor devices have been discussed. The results of theoretical analysis were demonstrated with computational experiments. We note that numerical schemes constructed in this work may be effectively applied to the investigation of EHP in the region of collector junction of bipolar transistors (BJT) as well as in the drain region of the Metal-Oxide Semiconductors (MOS). They can be used as a ‘building’ block for modelling semiconductor supelattices and other layered structures in acousto- and opto-electronics. The technological progress in the design of optoelectronics devices, such as laser

diodes (semiconductor lasers), LCD (Liquid Crystal Displays), light-emitting diodes (LEDs), and thin-film devices [46], requires further development of nonlocal mathematical models and efficient numerical methods for the investigation of physical processes in such devices.

One of the major challenges in the analysis of mathematical models arising in micro/opto-electronics consists of the investigation of a coupled system of equations with source terms:

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \mathbf{F_1}(\mathbf{u}, \mathbf{v})}{\partial \mathbf{x}}\mathbf{v} = \mathbf{G_1}(\mathbf{u}, \mathbf{v}), \quad A\frac{\partial \mathbf{v}}{\partial \mathbf{x}} = \mathbf{G_2}(\mathbf{u}, \mathbf{v}), \tag{7.1}$$

where $\mathbf{u}(\mathbf{x}, t), \mathbf{v}(\mathbf{x}, t) \in \mathbb{R}^m$, A is a constant real matrix, $(\mathbf{x}, t) \in \mathbb{R}^n \times \mathbb{R}_+$, $\mathbf{F_1}$ is a given vector function and $\mathbf{G_1}, \mathbf{G_2}$ are source terms. We recall that the system (7.1) is a stiff system of PDEs if the time scales introduced by the source terms, $\mathbf{G_1}$ and $\mathbf{G_2}$, are small compared to characteristic speeds and some appropriate length scale [47]. The mathematical analysis and the constructive solution of (7.1) in the class of piecewise-constant functions can be based on the approximation by Riemann problems which are simpler to solve than the standard Cauchy problem [48]. However, the solution of such a reduced problem may not exist. Alternatively, using the perturbation technique the system (7.1) can be reduced to a perturbed equation obtained by the substitution of $\mathbf{v}$, determined from the second equation of system (7.1), into the first equation. The perturbed equation is typically written with respect to a new (perturbed) variable $\mathbf{u}_\epsilon$ (see, for example, [48]). However, as a result of such a reduction, the definition of the parameter of perturbation, ϵ, in the reduced equation becomes coupled to the definition of the source terms and the natural space for perturbations becomes $\mathbb{L}_1$ rather than $\mathbb{L}_2$. Immediate difficulties arising from this fact are that the flow map of the solution of the reduced equation might not be differentiable with respect to linear structure of $\mathbb{L}_1$ and the contractivity of the flow for the perturbed equation with respect to $\mathbb{L}_1$-distance in the dimension higher than one cannot be guaranteed in general. These difficulties present a challenge for future work.

Acknowledgements

Authors were supported by grant USQ-PTRP 17989 and by Australian Research Council Small Grant 17906. We thank Dr. David Smith for his assistance at the final stage of preparation of this paper.

Appendix

The following notation for variables, constants and normalisation factors were used in this paper (the reader may consult [49, 18, 6, 5] for further details):
n, p: concentrations of the majority (electrons) and minority (holes) carriers, respectively;
φ, E: electrostatic potential and electric field strenth respectively;
T_n, T_p: carrier temperatures;
F: the generation/recombination/ionisation term;
P_n, P_p: rates of energy losses by scattering on the lattice for electrons and holes, respectively;
J_n, J_p: current densities;
Q_n, Q_p: energy densities;
D_n (D_p), μ_n (μ_p): diffusion and mobility coefficients;
n_{ie}: effective intrinsic concentration of carriers;
$\tilde{U}$, $\tilde{\varphi}_{\text{cont}}$: applied voltage and the contact potential difference, respectively;

$\tilde{n}_0$, $\tilde{p}_0$: initial concentrations of carriers;
f, f_1: Bernoulli's functions;
τ_ω^n, τ_ω^n: average energy relaxation times for electron and holes, respectively;
β_n, β_p: Peltier coefficients (taken 2·5);
v_s^n, v_s^p: saturation velocities of carriers;
L: length of the semiconductor structure;
φ_{cont}: contact potential;
U: applied voltage;
$T_* = 0{\cdot}0259$: normalisation factor for temperature;
$\varphi_* = 0{\cdot}0244$: normalisation factor for the potential;
$\mu_* = 1$, $D_* = T_*$: normalisation factors for the mobility and diffusion coefficient, respectively;
$t_* = 5{\cdot}0256 \times 10^{-6}$: time normalisation factor;
$n_* = 1{\cdot}2877 \times 10^{12}$: concentration normalisation factor;
$J_* = 1{\cdot}4349 \times 10^{-5}$: current density normalisation factor;
$\alpha_* = 2{\cdot}8571 \times 10^3$, $c_* = 1{\cdot}2 \times 10^{-19}$: normalisation factors for carrier ionisation and Auger recombination coefficients, respectively;
c_n, c_p: coefficients of the Auger recombination (taken $2{\cdot}9 \times 10^{-31}$ and $1{\cdot}2 \times 10^{-31}$, respectively);
q: electron charge ($q = |q|$ taken $1{\cdot}6 \times 10^{-19}$ Coulomb);
ϵ: relative dielectric permittivity of the semiconductor material (for Si it is 11·7 F/ cm^{-1});
ϵ_0: relative dielectric permittivity of vacuum (taken $8{\cdot}85 \times 10^{-14}$ F/ cm^{-1});
N, doping density of a device (the summarised concentration of dopants);
$\bar{\mathcal{E}}_n = 3nT_n/2$, $\bar{\mathcal{E}}_p = 3pT_p/2$: approximations of energy densities of carriers;
τ_ω^n, τ_ω^p: characteristic times of energetic relaxation (taken $0{\cdot}43 \times 10^{-12}$ s);
T_l: lattice temperature (taken 300 K);
α_n, α_p: coefficients of collision ionization (taken 1×10^{-3} and 1×10^{-4}, respectively);
m_e: electron mass;
$m_n = 0{\cdot}26m_e$: effective electron mass;
$c_s = \sqrt{\gamma T_n/m_n}$: sound speed, where $\gamma = 5/3$ is the polytropic gas constant;
k_b: Boltzmann constant (taken $1{\cdot}3 \times 10^{-23}$);
τ_n, τ_p: carriers life times (taken $1{\cdot}7 \times 10^{-5}$ s and $3{\cdot}95 \times 10^{-4}$ s, respectively).

References

1. C. Ringhofer, Computational methods for semiclassical and quantum transport in semiconductor devices. *Acta Num.* 6 (1997) 485–521.
2. A.A. Samarskii and B. N. Chetverushkin, Microelectronics as a New Object of Investigation in Applied Mathematics. *Comp. Math. and Cybern.: Vestnik Moscow Univ.* 3 (1986) 9–20.
3. R.V.N. Melnik, Correction for nonstationarity and internal nonlinearity in the analysis of integrated-circuits thermal parameters. *Radioelect. Comm. Syst*. 34 (1991) 84–86.
4. P.A. Markowich, C.A. Ringhofer and C. Schmeiser, *Semiconductor Equations.* Wien: Springer-Verlag (1991) 248 pp.
5. R.V.N. Melnik and H. He, Modelling nonlocal processes in semiconductor devices with exponential difference schemes (Part 1: Relaxation time approximations). Department of Mathematics and Computing, University of Southern Queensland, Tech. Rep. SC-MC-9822 (1998) 29 pp. (available at `http://www.sci.usq.edu.au/cgi-bin/wp/research/workingpapers`)
6. M. Shur, *Physics of Semiconductor Devices.* Englewood Cliffs: Prentice Hall (1990) 680 pp.

7. P. G. Scrobohaci and T.-W. Tang, Modeling of the Hot Electron Subpopulation and its Application to Impact Ionization in Submicron Silicon Devices. *IEEE Trans. Electron Devices* 41 (1994) 1197–1212.
8. K. Blotekjaer, Transport equations for electrons in two-valley semiconductors. *IEEE Trans. Electron Devices* ED-17 (1970) 38–47.
9. M. Rudan, F. Odeh and J. White, Numerical solution of the hydrodynamic model for a one-dimensional semiconductor device. *COMPEL* 6 (1987) 151–170.
10. C.L. Gardner, J.W. Jerome and D.J. Rose, Numerical Methods for the Hydrodynamic Device Model: Subsonic Flow. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* 8 (1989) 501–507.
11. Y. Zhang and M.El. Nokali, A Hydrodynamic transport model and its applications in semiconductor device simulation. *Solid-State Electron.* 36 (1993) 1689–1696.
12. R. K. Cook, Numerical Simulation of Hot-Carrier Transport in Silicon Bipolar Transistors. *IEEE Trans. Electron Devices* ED-30 (1983) 1103–1110.
13. L. J. Birukova *et al.*, Simulation algorithms for computing processes in electron plasma of submicron semiconductor devices. *Math. Model.* 1 (1989) 11–22.
14. E.D. Lyumkis *et al.*, Transient Semiconductor Device Simulation including energy balance equation. *COMPEL* 11 (1992) 311–325.
15. R.V.N. Melnik, Semi-Implicit Finite-Difference Schemes with Flow Correction for Quasi-Hydrodynamic Models of Semiconductor Devices. *Eng. Simulation* 12 (1995) 856–865.
16. V.A. Nikolaeva, V.I. Ryzhii and B.N. Cheverushkin, A numerical method for the simulation of two-dimensional semiconductor structures in the quasi-hydrodynamic approximation. *Sov. Phys. Dokl.* 33 (1988) 110–112.
17. Y. Apanovich, E. Lyumkis, B. S. Polsky *et al.*, Steady-State and Transient Analysis of Submicron Devices Using Energy Balance and Simplified Hydrodynamic Models. *IEEE Trans. Comp.-Aided Des. Integr. Circuits Syst.* 13 (1994) 702–711.
18. S. M. Sze, *Physics of Semiconductor Devices.* New York: John Wiley & Sons (1981) 868 pp.
19. N. R. Aluru, K. H. Law, P. M. Pinsky *et al.*, Space-Time Galerkin/Least-Squares Finite Element Formulation for the Hydrodynamic Device Equations. *IEICE Trans. Electron.* E77-C (1994) 227–235.
20. J. W. Slotboom, Computer aided two-dimensional analysis of bipolar transistor. *IEEE Trans. Electron. Devices* ED-20 (1973) 669–679.
21. A.H. Marshak and K.M. van Vliet, Electrical current in solids with position-dependent band structure. *Solid-State Electron.* 21 (1978) 417–427.
22. S.A. Mayorov, A.M. Melnikov and A.A. Rudenko, Modelling semiconductor microstructures in strong electric fields taking into account collision ionisation. *Math. Model.* 1 (1989) 23–32.
23. C.M. Snowden and D. Loret, Two-dimensional hot-electron models for short-gate-length GaAs MESFET's. *IEEE Trans. Electron. Devices* ED-34 (1987) 212–223.
24. R. Stratton, Diffusion of hot and cold electrons in semiconductor barriers. *Phys. Rev. B* 126 (1962) 2002–2014.
25. T.-W. Tang, X. X. Ou and D. X. Navon, Prediction of the velocity overshoot by a nonlocal hot-carrier transport model. In: J. J. H. Miller (ed.), *Proc. of the Fourth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits (NEMACOM IV)*. Dublin: Boole Press (1985) pp. 519–524.
26. C. Moglestue, *Monte Carlo Simulation of Semiconductor Devices.* New York: Chapman & Hall (1993) 326 pp.
27. P. A. Markowich, Diffusion Approximation of Nonlinear Electron Phonon Collision Mechanisms. *Model. Math. Anal. Numer. (M^2AN)* 29 (1995) 857–869.
28. J. W. Jerome, Algorithmic aspects of the hydrodynamic and drift-diffusion models. In: R.E. Bank, R. Burlirsch and K. Merten (eds.), *Mathematical Modelling and Simulation of Electrical Circuits and Semiconductor Devices.* Basel: Birkhauser-Verlag (1990) pp. 217–236.
29. B. S. Polsky and J. S. Rimshans, Two-dimensional numerical simulation of bipolar semiconductor devices taking into account heavy doping effects and Fermi statistics. *Solid-State Electron.* 26 (1983) 275–279.
30. T.D. Bui, A.K. Oppenheim, and D.T. Pratt, Recent advances in methods for numerical solution of ODE initial value problems. *J. Comp. Math.* 11 (1984) 283–296.
31. E.S. Oran and J.P. Boris, *Numerical Simulation of Reactive Flow.* New York: Elsevier (1987) 601 pp.
32. A. A. Samarskii and E. S. Nikolaev, *Numerical Methods for Grid Equations.* Basel: Birkhauser Verlag (1989) 588 pp.

33. T.-W. Tang, Extension of the Scharfetter-Gummel algorithm to the energy balance equation. *IEEE Trans. Electron. Devices* ED-31 (1984) 1912–1914.
34. T.-W. Tang and M.-K. Ieong, Discretization of Flux Densities in Device Simulations Using Optimum Artificial Diffusivity. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* 14 (1995) 1309–1315.
35. B. S. Polsky and J. S. Rimshans, Half-Implicit Difference Scheme for Numerical Simulation of Transient Processes in Semiconductor Devices. *Solid-State Electron.* 29 (1986) 321–328.
36. D.J. Widiger, Two-dimensional transient simulation of an idealized high electron mobility transistor. *IEEE Trans. Electron. Devices* ED-32 (1985) 1092–1103.
37. M. S. Mock, A time-dependent numerical model of the insulated-gate FET. *Solid-State Electron.* 24 (1981) 959–966.
38. R. V. N. Melnik and H. He, Modelling nonlocal processes in semiconductor devices with exponential difference schemes (Part 2: Numerical methods and computational experiments). Department of Mathematics and Computing, University of Southern Queensland, Tech. Rep. SC-MC-9831 (1998) 26 pp. (available at `http://www.sci.usq.edu.au/cgi-bin/wp/research/workingpapers`).
39. F. Brezzi, A.C.S. Capelo and L. Gastaldi, A singular perturbation analysis of reverse-biased semiconductor diodes. *SIAM J. Math. Anal.* 20 (1989) 372–387.
40. D. Kakati, C. Ramanan and V. Ramamurthy, Numerical analysis of electrophysical characteristics of semiconductor devices accounting for the heat transfer. In: J.J.H. Miller (ed.), *Proc. of the Fourth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits.* Dublin: Bool Press (1985) pp. 326–331.
41. C. Schmeiser, On strongly reverse biased semiconductor diodes. *SIAM J. Appl. Math.* 49 (1989) 1734–1748.
42. R. V. N. Melnik and K. N. Melnik, Modelling of Nonlocal Physical Effects in Semiconductor Plasma Using Quasi-Hydrodynamic Models. In: J. Noye, M. Teubner, A. Gill (eds.), *Computational Techniques and Applications: CTAC97.* Singapore: World Scientific (1998) pp. 441–448.
43. K. Singhal and J. Vlach, *Computer Methods for Circuit Analysis and Design.* New York: Van Nostrand Reinhold (1994) 712 pp.
44. W. T. Vetterling *et al.*, *Numerical Recipes Example Book (C).* New York: Cambridge University Press (1994) 325 pp.
45. A. Greenbaum, *Iterative Methods for Solving Linear Systems.* Philadelphia: SIAM (1997) 220 pp.
46. W. B. Leigh, *Devices for Optoelectronics.* New York: Marcel Dekker (1996) 315 pp.
47. A. Tveito and R. Winther, The solution of nonstrictly hyperbolic conservation laws may be hard to compute. *SIAM J. Sci. Comput.* 16 (1995) 320–329.
48. P.-A. Raviart and L. Sainsaulieu, A nonconservative hyperbolic systems modelling spray dynamics (Part 1: Solution of the Riemann problem). *Math. Models Methods Appl. Sci.* 5 (1995) 297–333.
49. C. Jacobini *et al.*, A review of some charge transport properties of silicon. *Solid-State Electron.* 20 (1977) 77–89.