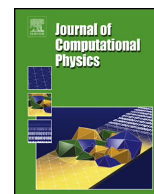
Journal of Computational Physics 458 (2022) 111086

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp



A class of finite element methods with averaging techniques for solving the three-dimensional drift-diffusion model in semiconductor device simulations

Qianru Zhang [a,b], Qin Wang [a,b], Linbo Zhang [a,b], Benzhuo Lu [a,b,*]

[a] LSEC, NCMIS, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[b] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 4 March 2021
Received in revised form 12 February 2022
Accepted 15 February 2022
Available online 22 February 2022

Keywords:
Three-dimensional drift-diffusion model
Averaging technique
Finite element method
Semiconductor device

ABSTRACT

Obtaining a satisfactory numerical solution of the classical three-dimensional drift-diffusion (DD) model, widely used in semiconductor device simulations, is still challenging nowadays, especially when the convection dominates the diffusion. In this work, we propose a series of finite element schemes with different types of averaging techniques to discretize the three-dimensional continuity equations. Our methods are based on the classical finite element framework, quite different from those mixed finite element/volume methods that also employ inverse averaging techniques. At first, the Slotboom variables are employed to transform the continuity equations into self-adjoint second-order elliptic equations with exponentially behaved coefficients. Then four averaging techniques, denoted with A1-A4, are introduced to approximate the exponential coefficient with its average on every tetrahedral element of the grid. The first scheme calculates the harmonic average of the exponential coefficient on a whole tetrahedron, and the other three schemes calculate the average of the exponential coefficient on each edge of a tetrahedral element. Our methods can avoid the spurious non-physical numerical oscillations and guarantee the conservation of the computed terminal currents with a terminal current evaluation approach. In fact, these methods not only maintain numerical stability but also overcome the disadvantages of some stabilization methods that cannot guarantee the conservation of the terminal currents, such as the streamline-upwind Petrov-Galerkin (SUPG) method. Moreover, the derivation of these discretization methods does not need the dual Voronoi grid as that of the finite volume Scharfetter-Gummel (FVSG) method or other mixed finite element/volume methods with inverse averaging techniques, which greatly reduces the complexity of parallel implementations of our methods. Simulations on two realistic three-dimensional semiconductor devices are carried out to test the accuracy and stability of our methods. According to numerical results, we conclude that scheme A4 can produce more accurate numerical solutions than the other three schemes, especially when the bias applied on the electrode is high. Numerical results also show that the scheme A4 is more robust than the Zlámal finite element method [1] in high-bias cases, and it also performs better than the FVSG method and a tetrahedral mixed finite element method [2] on poor-

* Corresponding author at: LSEC, NCMIS, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

E-mail address: bzlu@lsec.cc.ac.cn (B. Lu).

https://doi.org/10.1016/j.jcp.2022.111086
0021-9991/© 2022 Elsevier Inc. All rights reserved.

quality grids. Scheme A4 is also employed to study rich physical properties of the n-channel MOSFET.

© 2022 Elsevier Inc. All rights reserved.

1. Introduction

The drift-diffusion (DD) model introduced by Van Roosbroeck [3], which is composed of two convection-diffusion-reaction (continuity) equations and a Poisson equation, is frequently used in modeling the electromagnetic and thermal behavior of semiconductor devices. So far, it is still a challenging task to obtain a satisfactory numerical solution to this mathematical model, especially when the convection dominates the diffusion. The convection domination arises from the electric field solved from the Poisson equation after appropriate scaling is employed. Large biases applied on the electrodes and essentially discontinuous doping profiles, which may jump from large positive values to large negative values across an extremely thin layer, can lead to a strong convection-dominated effect and make the numerical solution suffer from spurious non-physical oscillations if the mesh size is not small enough. Due to these problems, classical numerical methods, including standard finite element/difference/volume methods, usually have difficulty in dealing with the DD model.

Since the Scharfetter-Gummel (SG) method [4] was firstly proposed for the one-dimensional DD model, the finite volume Scharfetter-Gummel (FVSG) method [5], also known as the box method, for higher-dimensional problems has been successfully employed in semiconductor device simulations up to now. This method uses boxes from the dual Voronoi grid as control volumes. The dual Voronoi grid only exists for Delaunay meshes. And it is challenging to generate Delaunay meshes, especially in three spatial dimensions [6]. If the semiconductor device includes two or more different material regions, the control volume covering different regions must be carefully treated. Therefore, the practical implementation of the FVSG scheme in three-dimensional semiconductor device simulations is complicated. Nowadays, many alternative discretization schemes have been proposed. A large family of methods derived from the classical finite element method (FEM) plays a crucial role in the numerical solution of the DD model. To avoid non-physical spurious oscillations of the numerical solutions to the convection-dominated continuity equations, people have developed many special schemes to improve the standard FEM: stabilization techniques [7–15], inverse averaging techniques [2,16–20], exponential fitting techniques [21–27], specifically designed basis functions [1,28–31], mixed finite element methods [32–35], and so on. Some stabilization techniques, such as the streamline upwind/Petrov-Galerkin (SUPG) method [8], can cut off negative oscillations (negative concentration) of carriers, but they may lead to a piling up of remaining positive oscillations [36,37]. Moreover, the stabilization approaches that enhance the stability of the numerical algorithm by adding diffusion terms or interior penalties (IP) cannot guarantee the conservation of the computed terminal currents [15,38]. Various inverse averaging techniques have been employed to deal with convection-diffusion problems, such as the simplex-averaged FEM [20]. This method is monotone, suitable for analysis, and does not need a dual Voronoi grid. In the derivation of the simplex-averaged FEM, the construction of the local simplex-averaged operators and the introduction of several special interpolation operators make this scheme suitable for the analysis. Our schemes directly use the Slotboom variables to transform the continuity equations into self-adjoint second order elliptic equations with exponentially behaved coefficients. Then we employ four different averaging techniques to deal with the exponential coefficients and compare them. The derivations in our work are relatively more convenient and straightforward for constructing the finial stiffness matrices. Exponential fitting techniques are also commonly used for solving convection-diffusion problems. In reference [25], the exponentially fitted edge fluxes are used to stabilize the FEM, and the employment of edge elements expands the edge fluxes into an H(curl)-conforming flux field inside each element. Different from [25], our schemes employ averaging techniques to improve the numerical stability and only need the linear finite element space. Apart from exponential fitting techniques, reference [39] utilizes quadrature rules and Newton's method to solve the nonlinear integral equation satisfied by the flux density. This scheme is used in a finite volume method which needs a dual Voronoi grid. We think that this scheme is an alternative approach and especially suitable for degenerate semiconductor devices satisfying non-Boltzmann statistics. Because in degenerate semiconductor device simulations, the continuity equation and the corresponding integral equation satisfied by the flux density are nonlinear. For non-degenerate semiconductor devices, the flux density can be analytically solved from the integral equation, and the classical FVSG scheme is recovered. Most of the above improved FEMs are mainly designed to solve the two-dimensional DD model, and some of them construct their finite element discretizations on the dual Voronoi grid as well. For some unordinary FEMs with specifically designed basis functions, their derivation is pretty complicated. In this work, we propose improved finite element schemes that can eliminate spurious numerical oscillations due to dominant convection terms and guarantee the conservation of the computed terminal currents for the three-dimensional model. Our schemes don't need the dual Voronoi grid or specifically designed finite element basis functions. Therefore, their derivation is simpler, and their parallel implementation is easier.

In semiconductor device simulations, the carrier concentration may vary extremely rapidly in some subregions of the device, and the variational form of the continuity equation is not symmetric. This implies that discretizing the carrier concentration directly with piecewise polynomials is not appropriate. However, the flux density varies moderately in the whole domain. Thus, many people apply the mixed finite element/volume methods to the DD model, in which the flux density is treated as a whole. In this work, the Slotboom variables [40] are introduced to eliminate the cross term in

the flux density and transform continuity equations into self-adjoint second-order elliptic equations with exponentially behaved coefficients. Then, our numerical difficulty lies in dealing with the exponential coefficients. The flux densities can be approximated with constant vectors on each element of the tetrahedral mesh due to their moderate variation. Correspondingly, the exponential coefficients are also approximated with constants in each tetrahedral element. Hence, finding appropriate approximation techniques is the key to our method. In some mixed finite element/volume methods, people have employed several inverse averaging techniques to deal with their singular coefficients on a control volume of the dual Voronoi grid. Inspired by these methods, we introduce four different forms of averaging techniques into the general FEM aiming at dealing with the exponential coefficients on general tetrahedral elements. In this work, we propose a series of schemes, denoted by A1-A4, based on the averaging techniques to discretize the three-dimensional continuity equations in the DD model. In scheme A1, the average of the exponential coefficient is calculated on a whole tetrahedral element. In schemes A2-A4, the average of the exponential coefficient is separately calculated on each edge of a tetrahedral element. Our methods can handle the layer oscillation problems and guarantee the conservation of the computed terminal currents with a terminal current evaluation approach. On the one hand, our methods possess the upwinding properties. On the other hand, they overcome the shortcoming that the terminal current conservation cannot be maintained for some stabilization schemes [15,38]. Later numerical results also show that our scheme A4 is more robust than the Zlámal finite element method [1] in high-bias cases. The Zlámal finite element method is a specifically designed FEM for solving the two- or three-dimensional DD model directly on triangle or tetrahedral elements without using the control volumes. Similar to it, our methods also don't need a dual Voronoi grid. The numerical results also show that scheme A4 performs better than the FVSG method and a tetrahedral mixed FEM [2] on poor-quality grids. The derivation of our methods is simpler than many FEMs specifically designed for solving the DD model, which greatly reduces the difficulty of their parallel implementation. The stiffness matrix can be readily built by looping over the tetrahedral elements of the mesh.

In the numerical experiments, we first use a simple cube test to check the accuracy of our methods. Then, two classical three-dimensional semiconductor devices, a $p-n$ junction and an n-channel MOSFET, are employed to evaluate the effectiveness of our methods. The $p-n$ junction is regarded as a benchmark to test the methods. And we conclude that scheme A4 can produce numerical solutions that are more in line with the physical properties of the $p-n$ junction, compared with the other schemes A1-A3, especially in high-bias cases. In addition, scheme A4 can simulate the rich physical properties of the n-channel MOSFET well.

The rest of this paper is organized as follows. In Section 2, we introduce the mathematical model (DD model) employed in semiconductor device simulations and various types of boundary conditions. In section 3, we present the central part of our methods, in which four different forms of averaging techniques are derived for dealing with the exponentially behaved coefficients in symmetrized continuity equations with the help of the Slotboom variables. In Section 4, we discuss the evaluation of terminal currents using our schemes. Numerical experiments on two realistic three-dimensional semiconductor devices are conducted in Section 5 to evaluate the accuracy and stability of our methods.

2. The mathematical model

2.1. The drift-diffusion (DD) model

The geometrical model of a semiconductor device is a bounded domain $\Omega \subset \mathbb{R}^3$, which is comprised of a semiconductor part Ω_S, and, with regard to metal-oxide-semiconductor field-effect transistors (MOSFETs), one or more subdomains of thin oxide adjacent to Ω_S, denoted with Ω_O. In this work, we consider the classical steady-state DD equations commonly used in semiconductor device simulations. These equations are

$$\begin{cases} -\nabla \cdot \epsilon \nabla \psi = q(p-n+C), & \text{in } \Omega, \\ \dfrac{1}{q}\nabla \cdot \mathbf{J}_n - R_n = 0, & \text{in } \Omega_S, \\ -\dfrac{1}{q}\nabla \cdot \mathbf{J}_p - R_p = 0, & \text{in } \Omega_S, \end{cases} \tag{2.1}$$

with

$$\begin{cases} \mathbf{J}_n = -qn\mu_n \nabla \psi + qD_n \nabla n, \\ \mathbf{J}_p = -qp\mu_p \nabla \psi - qD_p \nabla p, \end{cases} \tag{2.2}$$

where ψ[V] — content in the square bracket is the physical unit of the corresponding physical quantity — is the electrostatic potential, $\epsilon[CV^{-1}\ \text{cm}^{-1}]$ is the dielectric constant, $q[C]$ is the fundamental electron charge, n and $p[\text{cm}^{-3}]$ are the electron and hole concentrations inside the semiconductor ($n|_{\Omega_O} \equiv p|_{\Omega_O} \equiv 0$). $C = N_D - N_A[\text{cm}^{-3}]$ is the doping profile, which is assumed to be a given datum of the problem in terms of the donor and acceptor concentrations N_D and N_A. Source terms R_n, $R_p[\text{cm}^{-3}\,\text{s}^{-1}]$ can be interpreted as the net recombination/generation rate of carriers in unit time and volume. For simplicity, we set $R_p = R_n = 0$, that is, we do not consider the carrier recombination and generation effects. Moreover,

the oxide region is assumed to be a perfect insulator, which implies that $\mathbf{J}_n \cdot \mathbf{n}\big|_{\Omega_O} \equiv \mathbf{J}_p \cdot \mathbf{n}\big|_{\Omega_O} \equiv 0$. We assume that the temperature $T[K]$ of the crystal is constant, and also suppose the following Einstein's relations

$$D_n = \mu_n \frac{k_B T}{q}, \qquad D_p = \mu_p \frac{k_B T}{q} \tag{2.3}$$

for carrier mobilities μ_n, $\mu_p[\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}]$ and carrier diffusion coefficients D_n, $D_p[\mathrm{cm}^2\,\mathrm{s}^{-1}]$, where $k_B[VCK^{-1}]$ is the Boltzmann constant. The carrier mobilities are treated as constants and calculated using the constant low field mobility model [41] in this paper.

2.2. The nonlinear Poisson equation

In our work, the Poisson equation, i.e., the first equation in the DD model (2.1), is solved with the help of the quasi-Fermi levels ϕ_n, ϕ_p. Referring to the Maxwell-Boltzmann statistics, we get

$$n = n_{\mathrm{ie}} \exp\left(\frac{q}{k_B T}(\psi - \phi_n)\right), \qquad p = n_{\mathrm{ie}} \exp\left(\frac{q}{k_B T}(\phi_p - \psi)\right), \tag{2.4}$$

where n_{ie} is the intrinsic concentration of the semiconductor. Substituting (2.4) into the Poisson equation, we get the following nonlinear Poisson equation

$$-\nabla \cdot \epsilon \nabla \psi = q\left\{ n_{\mathrm{ie}}\left[\exp\left(\frac{q}{k_B T}(\phi_p - \psi)\right) - \exp\left(\frac{q}{k_B T}(\psi - \phi_n)\right)\right] + C\right\}, \quad \text{in } \Omega. \tag{2.5}$$

2.3. Boundary conditions

The boundary conditions of DD equations may be different for various kinds of semiconductor devices. In this paper, we mainly consider three categories of boundary conditions: the nonhomogeneous Dirichlet conditions (for ideal ohmic contacts and gate contacts), the nonhomogeneous Neumann conditions (for oxide-semiconductor interfaces), and the homogeneous Neumann conditions (for boundaries without contacts).

Ohmic contacts: On ohmic contacts, denoted with Γ_C, where external voltages $V_{\mathrm{ext}}\big|_{\Gamma_C}$ are applied to electrically drive the device, boundary conditions for the electrostatic potential ψ and concentrations of carriers n, p are all Dirichlet conditions:

$$\begin{aligned} n\big|_{\Gamma_C} &= \frac{C + \sqrt{C^2 + 4n_{\mathrm{ie}}^2}}{2}, \\ p\big|_{\Gamma_C} &= \frac{-C + \sqrt{C^2 + 4n_{\mathrm{ie}}^2}}{2}, \\ \psi\big|_{\Gamma_C} &= V_{\mathrm{ext}}\big|_{\Gamma_C} + \frac{k_B T}{q} \ln\left(\frac{n}{n_{\mathrm{ie}}}\right) = V_{\mathrm{ext}}\big|_{\Gamma_C} - \frac{k_B T}{q} \ln\left(\frac{p}{n_{\mathrm{ie}}}\right). \end{aligned}$$

Gate contacts: Gate contacts, denoted with Γ_G, are located over the oxide region Ω_O where external voltages $V_{\mathrm{ext}}\big|_{\Gamma_G}$ are applied to control the current flow between the input-output contacts of the device. Moreover, $n\big|_{\Omega_O} \equiv p\big|_{\Omega_O} \equiv 0$, so we just need to give the boundary condition of the electrostatic potential ψ, which is also a Dirichlet condition:

$$\psi\big|_{\Gamma_G} = V_{\mathrm{ext}}\big|_{\Gamma_G} - \phi_{\mathrm{ms}},$$

where ϕ_{ms} is the work function difference between the metal and an intrinsic reference semiconductor.

Oxide-semiconductor interfaces: On the interfaces between the oxide and the semiconductor, denoted with Γ_I, flux densities vanish, that is,

$$\mathbf{J}_n \cdot \mathbf{n}\big|_{\Gamma_I} = \mathbf{J}_p \cdot \mathbf{n}\big|_{\Gamma_I} = 0,$$

where $\mathbf{n}$ is the out normal vector of Γ_I. The electrostatic potential ψ satisfies the following conditions:

$$[\psi]_{\Gamma_I} = 0, \qquad \left[\epsilon \frac{\partial \psi}{\partial \mathbf{n}}\right]_{\Gamma_I} = \sigma,$$

where σ is the surface charge density on Γ_I, and $[\cdot]$ denotes the jump function.

Boundaries without contacts: Outer boundaries of the device that have no contacts, denoted with Γ_N, are treated with ideal Neumann boundary conditions:

$$\left.\frac{\partial \psi}{\partial \mathbf{n}}\right|_{\Gamma_N} = 0,$$
$$\mathbf{J}_n \cdot \mathbf{n}\big|_{\Gamma_N} = \mathbf{J}_p \cdot \mathbf{n}\big|_{\Gamma_N} = 0.$$

2.4. Scaling of the DD model

For computational convenience, we make the electrostatic potential and the quasi-Fermi levels dimensionless with the following scaling:

$$\psi \leftarrow \frac{q\psi}{k_B T}, \quad \phi_n \leftarrow \frac{q\phi_n}{k_B T}, \quad \phi_p \leftarrow \frac{q\phi_p}{k_B T}. \tag{2.6}$$

By taking (2.3) and (2.6) into DD equations (2.1)-(2.2), we obtain the scaled DD model including the nonlinear Poisson equation (2.5):

$$\begin{cases} -\nabla \cdot \epsilon \nabla \psi = \dfrac{q^2}{k_B T}\left[n_{\text{ie}}\left(\exp(\phi_p - \psi) - \exp(\psi - \phi_n)\right) + C\right], & \text{in } \Omega, \\ -\dfrac{1}{q}\nabla \cdot \mathbf{J}_n = -\nabla \cdot (D_n(\nabla n - n\nabla\psi)) = 0, & \text{in } \Omega_S, \\ \dfrac{1}{q}\nabla \cdot \mathbf{J}_p = -\nabla \cdot (D_p(\nabla p + p\nabla\psi)) = 0, & \text{in } \Omega_S. \end{cases} \tag{2.7}$$

3. Finite element discretization of continuity equations with averaging techniques

In semiconductor device simulations, the electron and hole concentrations may vary extremely rapidly near the interface of differently doped subregions, such as the depletion regions. Also, the variational forms of continuity equations are not symmetric. Therefore, it may be inappropriate to discretize carrier concentrations directly with piecewise polynomials. Moreover, the flux densities J_p and J_n vary moderately in the whole domain Ω. Thus, many people use the mixed finite element/volume methods to solve the DD model, in which the flux density is always treated as a whole. In this work, we introduce the scaled Slotboom variables [40]

$$\Phi_n = n\exp(-\psi), \qquad \Phi_p = p\exp(\psi),$$

aiming at eliminating the cross terms in the flux densities and transforming the continuity equations in (2.7) into a set of self-adjoint second-order elliptic partial differential equations with exponential coefficients $\exp(\psi)$ and $\exp(-\psi)$:

$$\begin{cases} -\dfrac{1}{q}\nabla \cdot \mathbf{J}_n = -\nabla \cdot (D_n \exp(\psi)\nabla\Phi_n) = 0, & \text{in } \Omega_S, \\ \dfrac{1}{q}\nabla \cdot \mathbf{J}_p = -\nabla \cdot (D_p \exp(-\psi)\nabla\Phi_p) = 0, & \text{in } \Omega_S. \end{cases} \tag{3.1}$$

Then our numerical difficulty lies in dealing with the exponential coefficients $\exp(\psi)$ and $\exp(-\psi)$.

In this paper, we propose four schemes with different kinds of averaging techniques to deal with the exponential coefficients, denoted respectively by A1-A4. To present our idea better, we first establish our method with the following boundary conditions:

$$\Phi_n|_{\Gamma_D} = \Phi_p|_{\Gamma_D} = 0,$$
$$D_n\exp(\psi)\nabla\Phi_n \cdot \mathbf{n}\big|_{\Gamma_N} = D_p\exp(-\psi)\nabla\Phi_p \cdot \mathbf{n}\big|_{\Gamma_N} = 0,$$

where $\partial\Omega_S = \Gamma_D \cup \Gamma_N$. When dealing with other kinds of boundary conditions, our methods can be easily derived by making corresponding changes.

Let $H^1(\Omega_S)$ be the Sobolev space of weakly differentiable functions and $H_D^1(\Omega_S) = \{v \in H^1(\Omega_S) \mid v = 0 \text{ on } \Gamma_D\}$. The variational form of the first continuity equation in (3.1) is to find $\Phi_n \in H_D^1(\Omega_S)$ such that

$$\int_{\Omega_S} D_n \exp(\psi)\nabla\Phi_n \cdot \nabla v \, d\Omega_S = 0, \quad \forall v \in H_D^1(\Omega_S). \tag{3.2}$$

Similarly, the variational form of the second continuity equation in (3.1) is to find $\Phi_p \in H_D^1(\Omega_S)$ satisfying

$$\int_{\Omega_S} D_p \exp(-\psi)\nabla\Phi_p \cdot \nabla v \, d\Omega_S = 0, \quad \forall v \in H_D^1(\Omega_S). \tag{3.3}$$

Let $\mathcal{T}_h$ be a tetrahedral mesh over the semiconductor part of the device Ω_S, $X_h = \{q_i\}_{i=1}^{N_v}$ be the set of all vertices of $\mathcal{T}_h$, and $T \in \mathcal{T}_h$ denote a tetrahedron in the tetrahedral mesh. We choose the test function v in the piecewise linear finite element space $V_h \subset H_D^1(\Omega_S)$, and denote it with v_h. The Slotboom variables Φ_n and Φ_p are respectively discretized by $\Phi_{nh} = \sum_i \Phi_{nh}(q_i)\varphi_i$, $\Phi_{ph} = \sum_i \Phi_{ph}(q_i)\varphi_i$, where φ_i denotes the linear Lagrangian basis function at q_i. Due to the moderate variation of flux density in the whole computational domain Ω_S, J_n (respectively J_p) can be approximated with a constant vector on each tetrahedral element of the mesh. So we also approximate the exponential coefficient $\exp(\psi)$ (respectively $\exp(-\psi)$) with the piecewise constant $E(\psi)_T$ (respectively $E(-\psi)_T$) on each tetrahedron T. Then the discrete forms of (3.2)-(3.3) become

$$
\begin{aligned}
0 &= \sum_{T\in\mathcal{T}_h} \int_T D_n \exp(\psi)\nabla\Phi_{nh}\cdot\nabla v_h\, dT \approx \sum_{T\in\mathcal{T}_h} D_n E(\psi)_T \int_T \nabla\Phi_{nh}\cdot\nabla v_h\, dT \\
&= \sum_{T\in\mathcal{T}_h} D_n E(\psi)_T \sum_{q_i\in T} \Phi_{nh}(q_i) \int_T \nabla\varphi_i\cdot\nabla v_h\, dT,
\end{aligned}
\tag{3.4}
$$

and

$$
\begin{aligned}
0 &= \sum_{T\in\mathcal{T}_h} \int_T D_p \exp(-\psi)\nabla\Phi_{ph}\cdot\nabla v_h\, dT \approx \sum_{T\in\mathcal{T}_h} D_p E(-\psi)_T \int_T \nabla\Phi_{ph}\cdot\nabla v_h\, dT \\
&= \sum_{T\in\mathcal{T}_h} D_p E(-\psi)_T \sum_{q_i\in T} \Phi_{ph}(q_i) \int_T \nabla\varphi_i\cdot\nabla v_h\, dT.
\end{aligned}
\tag{3.5}
$$

We now describe in detail the computation of the element-wise stiffness matrices of the continuity equations in (2.7), i.e., $A_n = (a_{ij}^T)_{T\in\mathcal{T}_h}$ and $A_p = (b_{ij}^T)_{T\in\mathcal{T}_h}$. On a tetrahedral element T, let v_h take the associated piecewise linear finite element basis functions, then we have

$$
\begin{aligned}
D_n E(\psi)_T \int_T \nabla\Phi_{nh}\cdot\nabla\varphi_i\, dT &= D_n E(\psi)_T \sum_{q_j\in T} \Phi_{nh}(q_j) \int_T \nabla\varphi_j\cdot\nabla\varphi_i\, dT \\
&\triangleq D_n E(\psi)_T \sum_{q_j\in T} \Phi_{nh}(q_j) e_{ij}^T,
\end{aligned}
\tag{3.6}
$$

and

$$
\begin{aligned}
D_p E(-\psi)_T \int_T \nabla\Phi_{ph}\cdot\nabla\varphi_i\, dT &= D_p E(-\psi)_T \sum_{q_j\in T} \Phi_{ph}(q_j) \int_T \nabla\varphi_j\cdot\nabla\varphi_i\, dT \\
&\triangleq D_p E(-\psi)_T \sum_{q_j\in T} \Phi_{ph}(q_j) e_{ij}^T.
\end{aligned}
\tag{3.7}
$$

Note that $e_{ij}^T = \int_T \nabla\varphi_j\cdot\nabla\varphi_i\, dT$ includes some geometric information of the tetrahedron T, and it holds for linear Lagrangian finite element basis functions that $e_{ii}^T = -\sum_{j\neq i} e_{ij}^T$. Therefore, equations (3.6)-(3.7) can be written as

$$
\begin{aligned}
D_n E(\psi)_T \int_T \nabla\Phi_{nh}\cdot\nabla\varphi_i\, dT &= -D_n \sum_{q_j\in T, q_j\neq q_i} E(\psi)_T (\Phi_{nh}(q_i) - \Phi_{nh}(q_j)) e_{ij}^T \\
&\triangleq -D_n \sum_{q_j\in T, q_j\neq q_i} E(\psi)_{\mathcal{E}_{ij}} (\Phi_{nh}(q_i) - \Phi_{nh}(q_j)) e_{ij}^T,
\end{aligned}
$$

and

$$
\begin{aligned}
D_p E(-\psi)_T \int_T \nabla\Phi_{ph}\cdot\nabla\varphi_i\, dT &= -D_p \sum_{q_j\in T, q_j\neq q_i} E(-\psi)_T (\Phi_{ph}(q_i) - \Phi_{ph}(q_j)) e_{ij}^T \\
&\triangleq -D_p \sum_{q_j\in T, q_j\neq q_i} E(-\psi)_{\mathcal{E}_{ij}} (\Phi_{ph}(q_i) - \Phi_{ph}(q_j)) e_{ij}^T.
\end{aligned}
$$

Inspired by the above equalities, we intend to approximate $\exp(\psi)$ and $\exp(-\psi)$ by averaging either over the whole tetrahedron or its edges, that is $E(\psi)_T$ (respectively $E(-\psi)_T$) may be a constant on each tetrahedral element or take different values $E(\psi)_{\mathcal{E}_{ij}}$ (respectively $E(-\psi)_{\mathcal{E}_{ij}}$) on different edges of every tetrahedral element. This leads to the following different schemes.

3.1. Averaging over the whole tetrahedron T

We first introduce the derivation of our scheme A1. Using harmonic averages to treat singular coefficients is natural in mixed methods [42,43]. The harmonic average has been proven to be able to provide a better result than the general mean value in the one-dimensional case, especially when the singular coefficient exhibits sharp variations or is even discontinuous on mesh elements. So we approximate the exponential coefficients with their respective harmonic averages over the whole tetrahedral element T as follows:

$$E(\Psi)_T = \left(\frac{1}{|T|}\int_T e^{-\Psi} dT\right)^{-1},$$

where $\Psi = \pm\psi$. We assume that Ψ is linear on T and $\Psi_i = \Psi(q_i)$, $q_i \in T$, $i = 1, 2, 3, 4$. Then referring to Appendix A in [2], we have

$$\begin{aligned} I(T) = \frac{1}{|T|}\int_T e^{\Psi} dT = e^{\Psi_1}\Bigg[&\frac{6e^{\Psi_4-\Psi_1}}{\Psi_4-\Psi_1}\left(\frac{e^{\Psi_3-\Psi_4}}{\Psi_3-\Psi_4}B^{-1}(\Psi_2-\Psi_3) - \frac{1}{\Psi_3-\Psi_4}B^{-1}(\Psi_2-\Psi_4)\right) \\ &-\frac{6}{\Psi_4-\Psi_1}\left(\frac{e^{\Psi_3-\Psi_1}}{\Psi_3-\Psi_1}B^{-1}(\Psi_2-\Psi_3) - \frac{1}{\Psi_3-\Psi_1}B^{-1}(\Psi_2-\Psi_1)\right)\Bigg] \\ \triangleq\, & e^{\Psi_1}\tilde{I}^T(\Psi). \end{aligned} \tag{3.8}$$

Here, $B(t)$ is the Bernoulli function defined by

$$B(t) = \begin{cases} \dfrac{t}{e^t - 1}, & t \neq 0, \\ 1, & t = 0. \end{cases}$$

For numerical stability, if the difference between two nodal values of Ψ_i $(i = 1, 2, 3, 4)$ is very little, the corresponding terms should be calculated using Taylor expansions. For more details, please refer to Appendix A in [2].

Our past numerical experience also shows that the Slotboom variables Φ_{nh} and Φ_{ph} are not suitable in practical computations. Therefore, the normal variables n_h, p_h are the unknowns we finally solve. For simplicity, we set $\Psi_1 = \Psi_i$ in (3.8) – other options are also possible, then the equations (3.6)-(3.7) become

$$\begin{aligned} D_n E(\psi)_T \int_T \nabla\Phi_{nh}\cdot\nabla\varphi_i\, dT &= D_n \sum_{q_j\in T}\left(\frac{e^{\psi_j}}{|T|}\int_T e^{-\psi} dT\right)^{-1} n_h(q_j) e_{ij}^T \\ &= \sum_{q_j\in T} D_n\left(e^{\psi_j-\psi_i}\tilde{I}^T(-\psi)\right)^{-1} e_{ij}^T n_h(q_j) \end{aligned}$$

and

$$\begin{aligned} D_p E(-\psi)_T \int_T \nabla\Phi_{ph}\cdot\nabla\varphi_i\, dT &= D_p \sum_{q_j\in T}\left(\frac{e^{-\psi_j}}{|T|}\int_T e^{\psi} dT\right)^{-1} p_h(q_j) e_{ij}^T \\ &= \sum_{q_j\in T} D_p\left(e^{\psi_i-\psi_j}\tilde{I}^T(\psi)\right)^{-1} e_{ij}^T p_h(q_j). \end{aligned}$$

Furthermore, the nonzero entries of the element-wise stiffness matrices $A_n = (a_{ij}^T)_{T\in\mathcal{T}_h}$ and $A_p = (b_{ij}^T)_{T\in\mathcal{T}_h}$ are respectively given by

$$a_{ij}^T = D_n\left(e^{\psi_j-\psi_i}\tilde{I}^T(-\psi)\right)^{-1} e_{ij}^T \quad \text{and} \quad b_{ij}^T = D_p\left(e^{\psi_i-\psi_j}\tilde{I}^T(\psi)\right)^{-1} e_{ij}^T.$$

3.2. Averaging over the edges of the tetrahedron T

In this subsection, we consider three averaging techniques to calculate the averages of the exponential coefficients on the edge $\mathcal{E}_{ij} = \overline{q_i q_j}$ of a tetrahedron T. The first one uses the trapezoidal rule, and the other two techniques refer to the harmonic (inverse) averaging strategies [42], which have been successfully applied in two-dimensional mixed methods. The average of the exponential coefficient on the edge $\mathcal{E}_{ij} = \overline{q_i q_j}$ is denoted with

$$E(\Psi)_{\mathcal{E}_{ij}} = \left(\frac{\int_{q_i}^{q_j} e^{\pm\Psi} \, ds}{|\mathcal{E}_{ij}|} \right)^{\pm 1} \triangleq I(\mathcal{E}_{ij}),$$

where $\Psi = \pm\psi$ as before. In this subsection, the unknowns we finally solve are still the normal variables n_h and p_h for the same reason as before.

At first, we employ the trapezoidal quadrature formula and the inverse averaging technique, then we have

$$I(\mathcal{E}_{ij}) = \frac{\int_{q_i}^{q_j} e^{\Psi} ds}{|\mathcal{E}_{ij}|} \approx e^{\Psi_i} \left(\frac{1 + e^{\Psi_j - \Psi_i}}{2} \right),$$

and

$$I(\mathcal{E}_{ij}) = \left(\frac{\int_{q_i}^{q_j} e^{-\Psi} ds}{|\mathcal{E}_{ij}|} \right)^{-1} \approx e^{\Psi_j} \left(\frac{2}{1 + e^{\Psi_j - \Psi_i}} \right).$$

Their corresponding schemes are respectively denoted as A2 and A3. Referring to the derivation process of scheme A1, we can get the nonzero entries of the element-wise stiffness matrices $A_n = (a_{ij}^T)_{T \in \mathcal{T}_h}$ and $A_p = (b_{ij}^T)_{T \in \mathcal{T}_h}$ of scheme A2:

$$a_{ij}^T = \begin{cases} \frac{D_n}{2}(1 + e^{\psi_i - \psi_j}) e_{ij}^T, & j \neq i, \\ -\sum_{k \neq i} \frac{D_n}{2}(1 + e^{\psi_k - \psi_i}) e_{ik}^T, & j = i, \end{cases}$$
$$b_{ij}^T = \begin{cases} \frac{D_p}{2}(1 + e^{\psi_j - \psi_i}) e_{ij}^T, & j \neq i, \\ -\sum_{k \neq i} \frac{D_p}{2}(1 + e^{\psi_i - \psi_k}) e_{ik}^T, & j = i. \end{cases}$$

Similarly, the nonzero entries of the element-wise stiffness matrices $A_n = (a_{ij}^T)_{T \in \mathcal{T}_h}$ and $A_p = (b_{ij}^T)_{T \in \mathcal{T}_h}$ of scheme A3 are as follows:

$$a_{ij}^T = \begin{cases} D_n \frac{2}{1 + e^{\psi_j - \psi_i}} e_{ij}^T, & j \neq i, \\ -\sum_{k \neq i} D_n \frac{2 e^{\psi_k - \psi_i}}{1 + e^{\psi_k - \psi_i}} e_{ik}^T, & j = i, \end{cases}$$
$$b_{ij}^T = \begin{cases} D_p \frac{2}{1 + e^{\psi_i - \psi_j}} e_{ij}^T, & j \neq i, \\ -\sum_{k \neq i} D_p \frac{2 e^{\psi_i - \psi_k}}{1 + e^{\psi_i - \psi_k}} e_{ik}^T, & j = i. \end{cases}$$

Then, we assume that Ψ is linear on the edge $\mathcal{E}_{ij}$, that is

$$\Psi(\mathbf{x}) = \left(\frac{\Psi_j - \Psi_i}{|\mathcal{E}_{ij}|} \right) (\mathbf{x} - \mathbf{x}_{q_i}) + \Psi_i, \quad \mathbf{x} \in [\mathbf{x}_{q_i}, \mathbf{x}_{q_j}].$$

Then we employ the inverse averaging technique and get

$$I(\mathcal{E}_{ij}) = \left(\frac{\int_{q_i}^{q_j} e^{-\Psi} ds}{|\mathcal{E}_{ij}|} \right)^{-1} = \left(\int_{q_i}^{q_j} \frac{e^{-\Psi_i}}{|\mathcal{E}_{ij}|} \left(\frac{e^{\Psi_i}}{e^{\Psi_j}} \right)^{\frac{\mathbf{x} - \mathbf{x}_{q_i}}{|\mathcal{E}_{ij}|}} d\mathbf{x} \right)^{-1} = e^{\Psi_i} B(\Psi_i - \Psi_j).$$

The corresponding scheme is denoted as A4. The equations (3.6)-(3.7) related to the normal variables n_h, p_h are as follows

$$D_n E(\psi)_T \int_T \nabla \Phi_{nh} \cdot \nabla \varphi_i \, dT = -D_n \sum_{q_j \in T, q_j \neq q_i} E(\psi)_{\mathcal{E}_{ij}} (e^{-\psi_i} n_h(q_i) - e^{-\psi_j} n_h(q_j)) e_{ij}^T$$
$$= \left(-\sum_{q_j \in T, q_j \neq q_i} D_n B(\psi_i - \psi_j) e_{ij}^T \right) n_h(q_i) + \sum_{q_j \in T, q_j \neq q_i} \left(D_n B(\psi_j - \psi_i) e_{ij}^T \right) n_h(q_j),$$

and

$$D_p E(-\psi)_T \int_T \nabla \Phi_{ph} \cdot \nabla \varphi_i \, dT = -D_p \sum_{q_j \in T, q_j \neq q_i} E(-\psi)_{\mathcal{E}_{ij}} (e^{\psi_i} p_h(q_i) - e^{\psi_j} p_h(q_j)) e_{ij}^T$$

$$= \left(- \sum_{q_j \in T, q_j \neq q_i} D_p B(\psi_j - \psi_i) e_{ij}^T \right) p_h(q_i) + \sum_{q_j \in T, q_j \neq q_i} \left(D_p B(\psi_i - \psi_j) e_{ij}^T \right) p_h(q_j).$$

The nonzero entries of the element-wise stiffness matrices $A_n = (a_{ij}^T)_{T \in \mathcal{T}_h}$ and $A_p = (b_{ij}^T)_{T \in \mathcal{T}_h}$ can be written as

$$a_{ij}^T = \begin{cases} D_n B(\psi_j - \psi_i) e_{ij}^T, & j \neq i, \\ -\sum_{k \neq i} D_n B(\psi_i - \psi_k) e_{ik}^T, & j = i, \end{cases}$$

$$b_{ij}^T = \begin{cases} D_p B(\psi_i - \psi_j) e_{ij}^T, & j \neq i, \\ -\sum_{k \neq i} D_p B(\psi_k - \psi_i) e_{ik}^T, & j = i. \end{cases}$$

Later $p-n$ junction numerical experiments show that schemes A1-A3 perform poorly comparing with scheme A4. Therefore, we briefly analyze the upwinding property of scheme A4. We take the electron continuity equation (3.6) as an example. The analysis for the hole continuity equation (3.7) is similar. The electron current on the edge $\mathcal{E}_{ij}$ is

$$j_n|_{\mathcal{E}_{ij}} = -D_n B(\psi_i - \psi_j) e_{ij}^T n_h(q_i) + D_n B(\psi_j - \psi_i) e_{ij}^T n_h(q_j).$$

We notice that

$$B(t) = \begin{cases} 1, & t \to 0, \\ 0, & t \to +\infty, \\ -t, & t \to -\infty. \end{cases}$$

Then we get

$$j_n|_{\mathcal{E}_{ij}} = \begin{cases} D_n e_{ij}^T \left(n_h(q_j) - n_h(q_i) \right), & \text{when } \psi_i = \psi_j, \\ D_n (\psi_i - \psi_j) e_{ij}^T n_h(q_i), & \text{when } \psi_i \ll \psi_j, \\ D_n (\psi_i - \psi_j) e_{ij}^T n_h(q_j), & \text{when } \psi_i \gg \psi_j. \end{cases} \tag{3.9}$$

From (3.9), we find that the electron current expression of scheme A4 is similar to that derived from the central difference scheme when the electric field is zero, namely $\psi_i = \psi_j$, as well. Moreover, when $\psi_i \ll \psi_j$ or $\psi_i \gg \psi_j$, the electron current given by scheme A4 is similar to the first-order upwinding form of the finite difference method, which implies that scheme A4 can produce physical current in the case of a large electric field.

Our later numerical results confirm the above analyses.

4. Evaluation of approximate terminal currents

Getting the ohmic contact currents is a goal of semiconductor device simulations. In this section, we present a method for the evaluation of approximate terminal currents using our finite element schemes introduced above. We also prove that our schemes can maintain the conservation of the terminal currents well. This property is verified with later numerical experiments.

For simplicity, we assume that the Dirichlet boundary Γ_D of a semiconductor device consists of a finite number of separated ohmic contacts, and the tetrahedral mesh $\mathcal{T}_h$ is generated such that the end-points of any contact are mesh nodes of $\mathcal{T}_h$. Let $S_h \triangleq \text{span}\{\varphi_i\} \subset H^1(\Omega_S)$. If $v_h \in S_h$ satisfies $v_h|_{\Gamma_D} = 0$, then $v_h \in V_h$. For any $\Gamma_C \subset \Gamma_D$, let φ_C be a piecewise constant function satisfying

$$\varphi_C = \begin{cases} 1, & \mathbf{x} \in \Gamma_C, \\ 0, & \mathbf{x} \in \Gamma_D \setminus \Gamma_C. \end{cases}$$

Through multiplying (3.1) with φ_C and integrating by parts, we have

$$0 = \int_{\Omega_S} (\nabla \cdot \mathbf{J}_n) \varphi_C d\Omega_S = \int_{\Gamma_C} \mathbf{J}_n \cdot \mathbf{n} ds - \int_{\Omega_S} \mathbf{J}_n \cdot \nabla \varphi_C d\Omega_S,$$

$$0 = \int_{\Omega_S} (\nabla \cdot \mathbf{J}_p) \varphi_C d\Omega_S = \int_{\Gamma_C} \mathbf{J}_p \cdot \mathbf{n} ds - \int_{\Omega_S} \mathbf{J}_p \cdot \nabla \varphi_C d\Omega_S.$$

Let φ_C^I be the linear interpolant of φ_C, then $\varphi_C^I = \sum_{j=1}^{N_C} \varphi_j$. And, the outflow currents J_{nh}^C and J_{ph}^C through Γ_C are separately defined as

$$J_{nh}^C = \int_{\Gamma_C} \mathbf{J}_{nh} \cdot \mathbf{n} ds = \int_{\Omega_S} \mathbf{J}_{nh} \cdot \nabla \varphi_C^I d\Omega_S = \int_{\Omega_S} qD_n \exp(\psi) \nabla \Phi_{nh} \cdot \nabla \varphi_C^I d\Omega_S$$
$$= \sum_{j=1}^{N_C} \sum_{T \in \mathcal{T}_h} \int_T qD_n E(\psi)_T \nabla \Phi_{nh} \cdot \nabla \varphi_j dT,$$
$$J_{ph}^C = \int_{\Gamma_C} \mathbf{J}_{ph} \cdot \mathbf{n} ds = \int_{\Omega_S} \mathbf{J}_{ph} \cdot \nabla \varphi_C^I d\Omega_S = -\int_{\Omega_S} qD_p \exp(-\psi) \nabla \Phi_{ph} \cdot \nabla \varphi_C^I d\Omega_S$$
$$= -\sum_{j=1}^{N_C} \sum_{T \in \mathcal{T}_h} \int_T qD_p E(-\psi)_T \nabla \Phi_{ph} \cdot \nabla \varphi_j dT.$$

Finally, the total terminal current J_{total}^C flowing out of Γ_C is equal to the sum of the electron and hole currents, namely

$$J_{\text{total}}^C = J_{nh}^C + J_{ph}^C.$$

Proposition 1. *The computed total terminal current flowing out of Γ_D is conservative,*

$$\sum_{\Gamma_C \subset \Gamma_D} J_{total}^C = 0.$$

Proof. Summing J_{nh}^C and J_{ph}^C respectively on all contacts, referring to (3.4)-(3.5), we have

$$\sum_{\Gamma_C \subset \Gamma_D} J_{nh}^C = \sum_{\Gamma_C \subset \Gamma_D} \sum_{T \in \mathcal{T}_h} \int_T qD_n \exp(\psi) \nabla \Phi_{nh} \cdot \nabla \varphi_C^I dT = \sum_{T \in \mathcal{T}_h} \int_T qD_n \exp(\psi) \nabla \Phi_{nh} \cdot \nabla \varphi^I dT$$
$$= \sum_{T \in \mathcal{T}_h} \int_T qD_n \exp(\psi) \nabla \Phi_{nh} \cdot \nabla (\varphi^I - 1) dT = 0,$$

and

$$\sum_{\Gamma_C \subset \Gamma_D} J_{ph}^C = -\sum_{\Gamma_C \subset \Gamma_D} \sum_{T \in \mathcal{T}_h} \int_T qD_p \exp(-\psi) \nabla \Phi_{ph} \cdot \nabla \varphi_C^I dT = -\sum_{T \in \mathcal{T}_h} \int_T qD_p \exp(-\psi) \nabla \Phi_{ph} \cdot \nabla \varphi^I dT$$
$$= -\sum_{T \in \mathcal{T}_h} \int_T qD_p \exp(-\psi) \nabla \Phi_{ph} \cdot \nabla (\varphi^I - 1) dT = 0.$$

Here $\varphi^I = \sum_{\Gamma_C \subset \Gamma_D} \varphi_C^I \in S_h$ and $\varphi^I|_{\Gamma_D} = 1$, thus $\varphi^I - 1 \in V_h$. Then we get

$$\sum_{\Gamma_C \subset \Gamma_D} J_{\text{total}}^C = \sum_{\Gamma_C \subset \Gamma_D} J_{nh}^C + \sum_{\Gamma_C \subset \Gamma_D} J_{ph}^C = 0. \quad \square$$

Proposition 2. *If source terms R_n and R_p exist, conservation of the computed terminal current flowing out of Γ_D can still be maintained, that is,*

$$\sum_{\Gamma_C \subset \Gamma_D} J_{total}^C = \int_{\Omega_S} q(R_n - R_p) d\Omega_s.$$

Proof. At first, the outflow currents J_{nh}^C and J_{ph}^C through Γ_C can be written as

$$J_{nh}^C = \int_{\Gamma_C} \mathbf{J}_{nh} \cdot \mathbf{n} ds = \int_{\Omega_S} \mathbf{J}_{nh} \cdot \nabla \varphi_C^I d\Omega_S + \int_{\Omega_S} qR_n \varphi_C^I d\Omega_S$$
$$= \int_{\Omega_S} qD_n \exp(\psi) \nabla \Phi_{nh} \cdot \nabla \varphi_C^I d\Omega_S + \int_{\Omega_S} qR_n \varphi_C^I d\Omega_S$$
$$= \sum_{j=1}^{N_C} \sum_{T \in \mathcal{T}_h} \int_T qD_n E(\psi)_T \nabla \Phi_{nh} \cdot \nabla \varphi_j dT + \sum_{j=1}^{N_C} \sum_{T \in \mathcal{T}_h} \int_T qR_n \varphi_j dT,$$

$$
\begin{aligned}
J_{ph}^C &= \int\limits_{\Gamma_C} \mathbf{J}_{ph}\cdot\mathbf{n}ds = \int\limits_{\Omega_S} \mathbf{J}_{ph}\cdot\nabla\varphi_C^I d\Omega_S - \int\limits_{\Omega_S} qR_p\varphi_C^I d\Omega_S \\
&= -\int\limits_{\Omega_S} qD_p\exp(-\psi)\nabla\Phi_{ph}\cdot\nabla\varphi_C^I d\Omega_S - \int\limits_{\Omega_S} qR_p\varphi_C^I d\Omega_S \\
&= -\sum_{j=1}^{N_C}\sum_{T\in\mathcal{T}_h}\int\limits_T qD_pE(-\psi)_T\nabla\Phi_{ph}\cdot\nabla\varphi_j dT - \sum_{j=1}^{N_C}\sum_{T\in\mathcal{T}_h}\int\limits_T qR_p\varphi_j dT.
\end{aligned}
$$

And the equations (3.4)-(3.5) have the following forms when the source terms exist:

$$
\begin{aligned}
-\sum_{T\in\mathcal{T}_h}\int\limits_T R_n v_h dT &= \sum_{T\in\mathcal{T}_h}\int\limits_T D_n\exp(\psi)\nabla\Phi_{nh}\cdot\nabla v_h\, dT \approx \sum_{T\in\mathcal{T}_h} D_nE(\psi)_T\int\limits_T \nabla\Phi_{nh}\cdot\nabla v_h\, dT \\
&= \sum_{T\in\mathcal{T}_h} D_nE(\psi)_T\sum_{q_i\in T}\Phi_{nh}(q_i)\int\limits_T \nabla\varphi_i\cdot\nabla v_h\, dT, \quad \forall v_h\in V_h,
\end{aligned} \tag{4.1}
$$

and

$$
\begin{aligned}
-\sum_{T\in\mathcal{T}_h}\int\limits_T R_p v_h dT &= \sum_{T\in\mathcal{T}_h}\int\limits_T D_p\exp(-\psi)\nabla\Phi_{ph}\cdot\nabla v_h\, dT \approx \sum_{T\in\mathcal{T}_h} D_pE(-\psi)_T\int\limits_T \nabla\Phi_{ph}\cdot\nabla v_h\, dT \\
&= \sum_{T\in\mathcal{T}_h} D_pE(-\psi)_T\sum_{q_i\in T}\Phi_{ph}(q_i)\int\limits_T \nabla\varphi_i\cdot\nabla v_h\, dT, \quad \forall v_h\in V_h.
\end{aligned} \tag{4.2}
$$

By referring to (4.1)-(4.2), we get the respective sum of J_{nh}^C and J_{ph}^C on all contacts:

$$
\begin{aligned}
\sum_{\Gamma_C\subset\Gamma_D} J_{nh}^C &= \sum_{\Gamma_C\subset\Gamma_D}\sum_{T\in\mathcal{T}_h}\int\limits_T qD_n\exp(\psi)\nabla\Phi_{nh}\cdot\nabla\varphi_C^I dT + \sum_{\Gamma_C\subset\Gamma_D}\sum_{T\in\mathcal{T}_h}\int\limits_T qR_n\varphi_C^I dT \\
&= \sum_{T\in\mathcal{T}_h}\int\limits_T qD_n\exp(\psi)\nabla\Phi_{nh}\cdot\nabla\varphi^I dT + \sum_{T\in\mathcal{T}_h}\int\limits_T qR_n\varphi^I dT \\
&= \sum_{T\in\mathcal{T}_h}\int\limits_T qD_n\exp(\psi)\nabla\Phi_{nh}\cdot\nabla(\varphi^I-1)dT + \sum_{T\in\mathcal{T}_h}\int\limits_T qR_n(\varphi^I-1)dT + \sum_{T\in\mathcal{T}_h}\int\limits_T qR_n dT, \\
&= \sum_{T\in\mathcal{T}_h}\int\limits_T qR_n dT, \\
\sum_{\Gamma_C\subset\Gamma_D} J_{ph}^C &= -\sum_{\Gamma_C\subset\Gamma_D}\sum_{T\in\mathcal{T}_h}\int\limits_T qD_p\exp(-\psi)\nabla\Phi_{ph}\cdot\nabla\varphi_C^I dT - \sum_{\Gamma_C\subset\Gamma_D}\sum_{T\in\mathcal{T}_h}\int\limits_T qR_p\varphi_C^I dT \\
&= -\sum_{T\in\mathcal{T}_h}\int\limits_T qD_p\exp(-\psi)\nabla\Phi_{ph}\cdot\nabla\varphi^I dT - \sum_{T\in\mathcal{T}_h}\int\limits_T qR_p\varphi^I dT \\
&= -\sum_{T\in\mathcal{T}_h}\int\limits_T qD_p\exp(-\psi)\nabla\Phi_{ph}\cdot\nabla(\varphi^I-1)dT - \sum_{T\in\mathcal{T}_h}\int\limits_T qR_p(\varphi^I-1)dT - \sum_{T\in\mathcal{T}_h}\int\limits_T qR_p dT \\
&= -\sum_{T\in\mathcal{T}_h}\int\limits_T qR_p dT,
\end{aligned}
$$

where $\varphi^I = \sum_{\Gamma_C\subset\Gamma_D}\varphi_C^I$. Since $\varphi^I\in S_h$ and $\varphi^I|_{\Gamma_D}=1$, we have $\varphi^I-1\in V_h$. Therefore,

$$
\sum_{\Gamma_C\subset\Gamma_D} J_{\text{total}}^C = \sum_{\Gamma_C\subset\Gamma_D} J_{nh}^C + \sum_{\Gamma_C\subset\Gamma_D} J_{ph}^C = \int\limits_{\Omega_S} q(R_n-R_p)d\Omega_S. \quad \square
$$

5. Numerical experiments

In this section, a simple cube test is carried out to check the accuracy of our schemes. Then we evaluate the effectiveness of our methods by applying them to simulating two realistic three-dimensional semiconductor devices: a $p-n$ junction and an n-channel MOSFET. Their respective physical properties have been discussed in [44].

In the numerical experiments, the DD model (2.7) is decoupled with the Gummel iterative method [45] and solved using nonlinear block Gauss-Seidel iterations based on the successive solution of the nonlinear Poisson equation and two continuity equations. We also use piecewise linear finite element basis functions to discretize the variational form of the nonlinear Poisson equation:

$$\int_{\Omega} \epsilon \nabla \psi \cdot \nabla v d\Omega - \int_{\Gamma_N} \epsilon \frac{\partial \psi}{\partial \mathbf{n}} v ds = \int_{\Omega} \frac{q^2}{k_B T} \left[n_{\text{ie}} \left(\exp(\phi_p - \psi) - \exp(\psi - \phi_n) \right) + C \right] v d\Omega, \quad \forall v \in H_D^1(\Omega).$$

And the resulting nonlinear algebraic systems are solved by Newton's method.

Our methods are implemented based on the open-source finite element toolbox Parallel Hierarchical Grid (PHG) [46]. The computations were done on the high performance computers of State Key Laboratory of Scientific and Engineering Computing, Chinese Academy of Sciences.

5.1. A cube test

Before applying our methods to semiconductor devices, we use a cube $[-10nm, 10nm]^3$ to test their accuracy. In this cube test, the doping profile is not considered. Then the DD model degenerates into a Poisson-Nernst-Planck model [15]. An aqueous solution of KCl is considered in this cube, and the dielectric constant ϵ is 80. For simplicity, we use p and n to denote K^+ and Cl^- concentrations only in this cube test. Their diffusion coefficients are $D_{K^+} = 1.96 \times 10^{-9}$ m^2/s and $D_{Cl^-} = 2.03 \times 10^{-9}$ m^2/s. The top and bottom of the cube, denoted with Γ_t and Γ_b, are set as Dirichlet boundaries:

$$\begin{cases} \psi\big|_{\Gamma_b} = 0, \quad \psi\big|_{\Gamma_t} = \dfrac{q}{k_B T} V_{\text{ext}}, \\ p\big|_{\Gamma_b} = c^b, \quad p\big|_{\Gamma_t} = c^b, \\ n\big|_{\Gamma_b} = c^b, \quad n\big|_{\Gamma_t} = c^b. \end{cases}$$

And the sides of the cube are set as Neumann boundaries:

$$\begin{cases} \left.\dfrac{\partial \psi}{\partial \mathbf{n}}\right|_{\Gamma_N} = 0, \\ \mathbf{J}_n \cdot \mathbf{n}\big|_{\Gamma_N} = \mathbf{J}_p \cdot \mathbf{n}\big|_{\Gamma_N} = 0. \end{cases}$$

Hence, the total current flowing out of Γ_t or Γ_b is respectively calculated as

$$\begin{aligned} J^t_{\text{total}} &= J^t_{nh} + J^t_{ph} = -D_n q c^b \frac{q}{k_B T} \frac{V_{\text{ext}}}{H} S - D_p q c^b \frac{q}{k_B T} \frac{V_{\text{ext}}}{H} S, \\ J^b_{\text{total}} &= -J^t_{\text{total}} = D_n q c^b \frac{q}{k_B T} \frac{V_{\text{ext}}}{H} S + D_p q c^b \frac{q}{k_B T} \frac{V_{\text{ext}}}{H} S, \end{aligned} \tag{5.1}$$

where H is the height of the cube, and S is the area of the top/bottom of the cube. Let $c^b = 0.01M$, and we make comparisons between our four methods with different external voltage values. Numerical results in Table 1 show that all of our methods can guarantee the conservation of total currents, and scheme A4 provides more accurate numerical results than the other methods referring to the theoretical currents calculated with (5.1). As the external voltage increases, the gaps between the currents calculated with schemes A1-A3 and the theoretical currents become larger and larger.

Our previous numerical experience shows that the standard FEM and the SUPG method cannot work for simulating semiconductor devices of tens of microns. Therefore, in the following numerical experiments, we employ our schemes to simulate the realistic three-dimensional semiconductor devices.

5.2. A silicon $p-n$ junction

The $p-n$ junction plays a significant role in understanding other semiconductor devices, and it is an important block for the bipolar junction transistor (BJT) and the MOSFET. We choose the silicon $p-n$ junction depicted in Fig. 1 as a benchmark to test the effectiveness of our methods and only consider the abrupt junction — the doping profile is

$$C = N_D - N_A = \begin{cases} -10^{17}, & \text{in p-type region,} \\ 10^{17}, & \text{in n-type region.} \end{cases}$$

Table 1
Total currents flowing out of Γ_t and Γ_b for different external voltage values.

	$V_{\text{ext}}[V]$	0.2	0.4	0.6	0.8	1.0
Absolute values of theoretical currents $[A]$		5.956475e-10	1.191295e-09	1.786943e-09	2.382590e-09	2.978238e-09
A1	$J^t_{\text{total}}[A]$	-6.003695e-10	-1.218452e-09	-1.844623e-09	×	×
	$J^b_{\text{total}}[A]$	6.003695e-10	1.218452e-09	1.844623e-09	×	×
A2	$J^t_{\text{total}}[A]$	-6.191300e-10	-1.385850e-09	-2.482934e-09	-4.171907e-09	-6.855393e-09
	$J^b_{\text{total}}[A]$	6.191300e-10	1.385850e-09	2.482934e-09	4.171907e-09	6.855393e-09
A3	$J^t_{\text{total}}[A]$	-5.843078e-10	-1.106388e-09	-1.527970e-09	-1.841442e-09	-2.060518e-09
	$J^b_{\text{total}}[A]$	5.843078e-10	1.106388e-09	1.527970e-09	1.841442e-09	2.060518e-09
A4	$J^t_{\text{total}}[A]$	-5.956475e-10	-1.191295e-09	-1.786943e-09	-2.382590e-09	-2.978238e-09
	$J^b_{\text{total}}[A]$	5.956475e-10	1.191295e-09	1.786943e-09	2.382590e-09	2.978238e-09

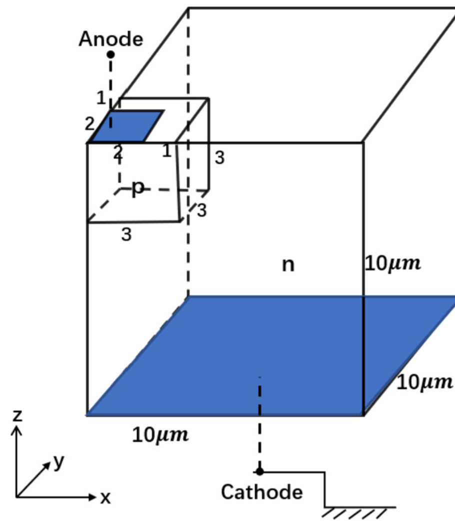


Fig. 1. A $p-n$ junction with the ohmic contacts shaded.

Table 2
Terminal currents for different forward biases.

	Bias [V]	0.2	0.4	0.6	0.8	1.0
A1	$J_A[A]$	-1.297674e-14	-2.964586e-11	-6.747598e-08	-9.360842e-05	-2.235388e-03
	$J_C[A]$	1.295691e-14	2.964264e-11	6.746864e-08	9.358315e-05	2.235333e-03
A2	$J_A[A]$	-1.328226e-14	-3.037920e-11	-6.942833e-08	-9.647769e-05	-2.370094e-03
	$J_C[A]$	1.328455e-14	3.037920e-11	6.942833e-08	9.647769e-05	2.370094e-03
A3	$J_A[A]$	-1.207576e-14	-2.774850e-11	-6.405003e-08	-9.359628e-05	-1.499261e-03
	$J_C[A]$	1.208389e-14	2.774850e-11	6.405003e-08	9.359628e-05	1.499261e-03
A4	$J_A[A]$	-1.297674e-14	-2.964586e-11	-6.747598e-08	-9.360842e-05	-2.235388e-03
	$J_C[A]$	1.295691e-14	2.964264e-11	6.746864e-08	9.358315e-05	2.235333e-03

We always let the cathode be grounded, namely the bias on the cathode is zero.

At first, we verify that our methods can guarantee the conservation of the computed terminal currents, see Table 2, which gives the computed terminal currents flowing in the anode ($J_A[A]$) and out from the cathode ($J_C[A]$) for different forward biases — positive biases on the anode. The negative sign represents the inflow of the current. From Table 2, we can see the conservation of the computed terminal currents even though there are some insignificant differences between the computed terminal currents of the two electrodes, which may be caused by rounding errors in the floating-point operations.

Then we study different performances of our four schemes when a forward bias is applied, and the results are shown in Fig. 2. They show that all schemes perform well when the forward bias is less than 1 V. Fig. 2(a) also illustrates that

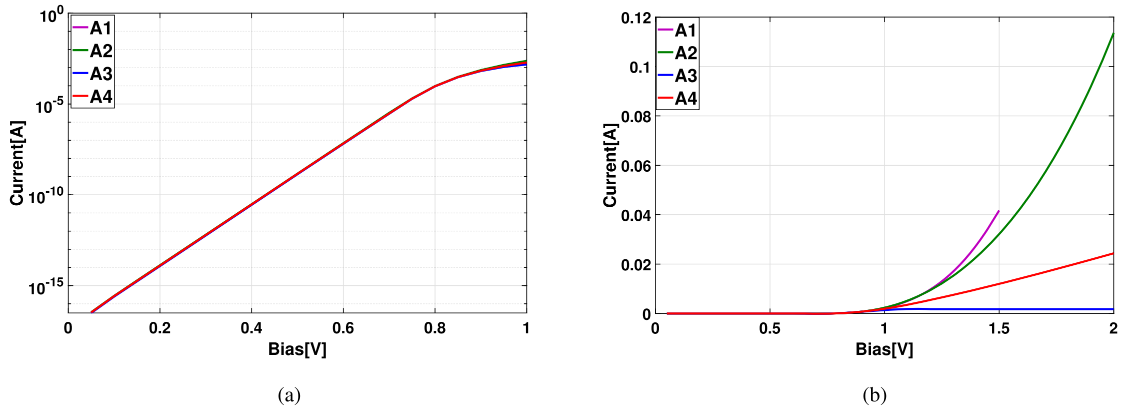


Fig. 2. Current-voltage characteristics of the $p-n$ junction: (a) Semilog plot (b) Cartesian plot.

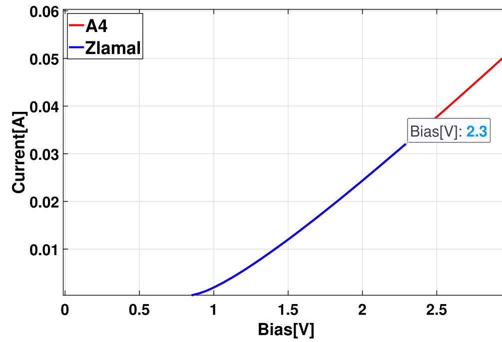


Fig. 3. Current-voltage characteristics of the $p-n$ junction.

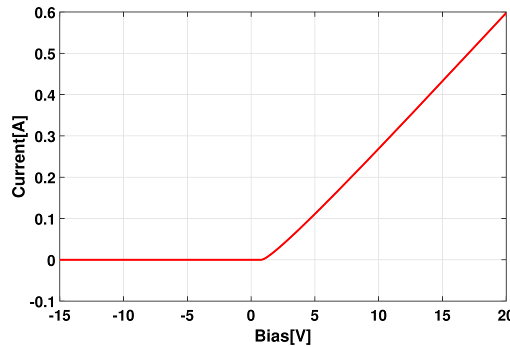


Fig. 4. Ideal current-voltage characteristic of the $p-n$ junction calculated by scheme A4.

the cut-in voltage is around 0.8 V, and the current increases exponentially as the bias increases before the bias reaches the cut-in voltage. But when the bias is higher than 1 V, currents calculated by schemes A1-A4 are apparently different. For scheme A4, the current increases linearly as the bias increases after the bias exceeds the cut-in voltage, which is in good accordance with the ideal current-voltage (I-V) characteristics of the silicon $p-n$ junction. For scheme A3, the current is smaller than that obtained from scheme A4, and it seems to reach a saturation state when the bias is higher than 1 V. For scheme A2, the current is larger than that derived from scheme A4, and it increases faster than the linear speed. For scheme A1, the current increases so rapidly as the bias increases, and the scheme does not work when the bias is higher than 1.5 V. Therefore, only scheme A4 can produce an I-V curve meeting the ideal physical characteristics of the $p-n$ junction well when the bias voltage is high.

To further check scheme A4, we compare it with the Zlámal finite element method [1], which can solve the three-dimensional DD model directly on tetrahedral elements rather than on the control volumes. Fig. 3 shows that the current values calculated by these two methods are consistent, and scheme A4 can work with a higher voltage range than the Zlámal method, as shown in Fig. 4. In this numerical example, the Zlámal finite element method cannot work if the bias is

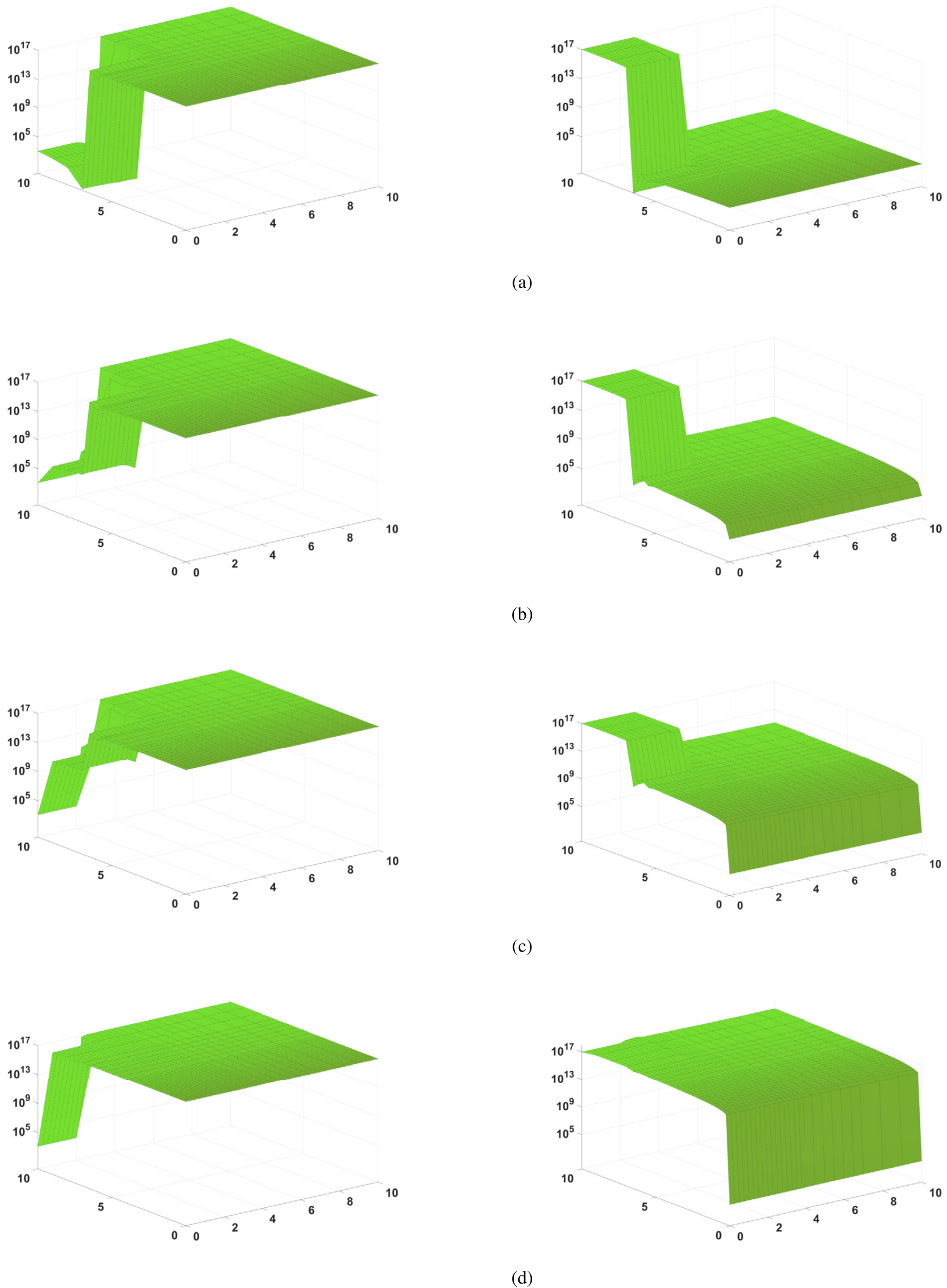


Fig. 5. Electron and hole concentration distributions computed with the scheme A4 on the cross-section $y = 0$ μm for different biases: (a) $V_{\text{ext|anode}} = -1.0$ V (b) $V_{\text{ext|anode}} = 0.2$ V (c) $V_{\text{ext|anode}} = 0.5$ V (d) $V_{\text{ext|anode}} = 1.0$ V.

higher than 2.3 V, whereas scheme A4 can still work under as high as 20 V bias. Fig. 4 also illustrates that no virtual current flows initially when we apply a reverse bias. Electron and hole concentration distributions computed with scheme A4 on the cross-section $y = 0$ μm for different biases are plotted in Fig. 5. From these plots we find that the numerical results produced by scheme A4 don't have spurious oscillations even though they display sharp interior layers in the neighborhood

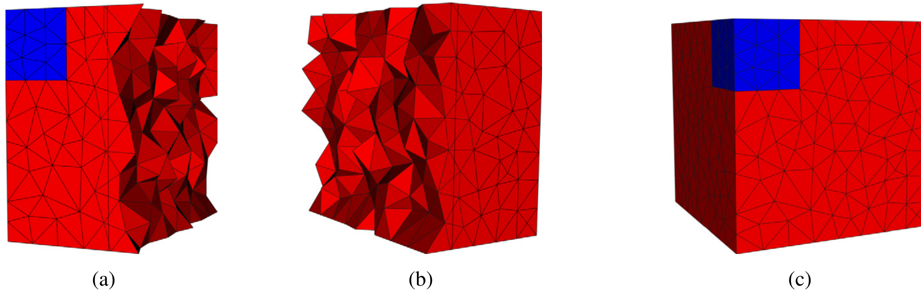


Fig. 6. A poor-quality grid for the $p-n$ junction with 3839 tetrahedral elements.

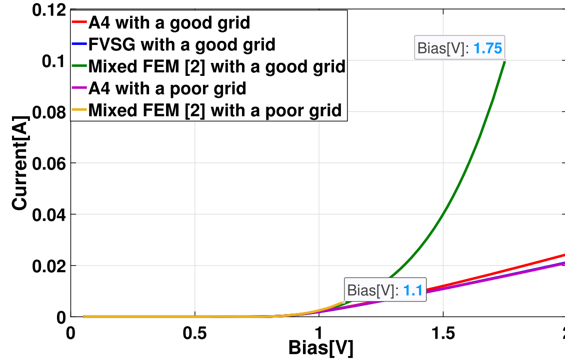


Fig. 7. Current-voltage characteristics of the $p-n$ junction.

of the $p-n$ junction. This illustrates that our scheme can solve the instability problem of the standard FEM as effectively as the stable SUPG method or the even more robust SUPG-IP (interior penalty) method [15]. Moreover, Fig. 5 presents that the depletion layer (interior layer) width under a forward bias condition is obviously smaller than that under a reverse bias condition, and the depletion layer narrows as the forward bias increases.

At last, we test the robustness of our scheme A4 on a poor-quality grid shown in Fig. 6, which is an unstructured tetrahedral grid generated by Tetgen [47]. On this grid, the FVSG method [4] cannot solve the DD model even at the equilibrium state, and a tetrahedral mixed FEM [2] can solve the DD model with this grid only when the forward bias is no higher than 1.1 V, see Fig. 7. On a good grid, the terminal current produced by the tetrahedral mixed FEM is not accurate with the bias higher than 1.0 V. The current increases so rapidly as the bias increases, and the tetrahedral mixed FEM can only work for a bias lower than 1.75 V. Fig. 7 also shows that scheme A4 can still produce an I-V curve satisfying ideal physical properties of the $p-n$ junction on this poor-quality grid, which is very close to that calculated with the FVSG method on a good grid.

Based on the numerical experience gained from this example, we next use scheme A4 to study the following semiconductor device: an n-channel MOSFET.

5.3. An n-channel MOSFET

The basic MOSFET is composed of a silicon semiconductor part Ω_S and a thin insulating oxide layer Ω_O placed immediately adjacent to Ω_S. The device geometry considered in this numerical example is depicted in Fig. 8, where the ohmic contacts and the gate are respectively shaded. We still consider the abrupt junction, and the doping profile:

$$C = N_D - N_A = \begin{cases} -10^{16}, & \text{in p-type region}, \\ 10^{18}, & \text{in } n^{+}\text{-type region}. \end{cases}$$

The substrate and source are grounded, which means biases on them are zero. Scheme A4 is employed in this numerical example.

We first study transfer characteristics and output characteristics of an ideal n-channel MOSFET. For an ideal MOSFET, the work function difference ϕ_{ms} is set to be zero. In addition, we don't consider the effect of charges in the oxide and traps at the oxide-semiconductor interface. The transfer characteristic, i.e., the drain current I_D versus the gate bias V_G, with a fixed drain bias $V_D = 0.5$ V is plotted in Fig. 9, in which we note that a positive gate bias higher than the threshold voltage V_T, i.e., $V_T = 0.85$ V in this ideal situation, must be applied before a substantial drain current flows. This type of MOSFET is called the normally-off (enhancement) n-channel MOSFET. Electron and hole concentration distributions corresponding to

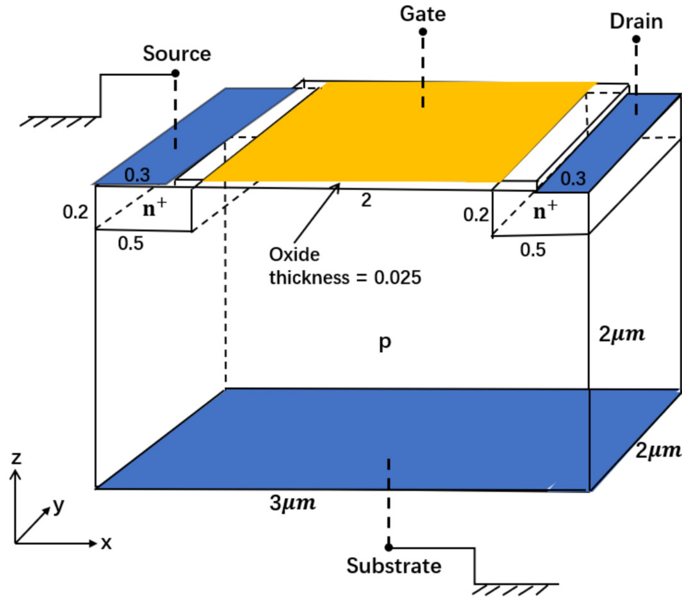


Fig. 8. An n-channel MOSFET, where the ohmic contacts and the gate are respectively shaded.

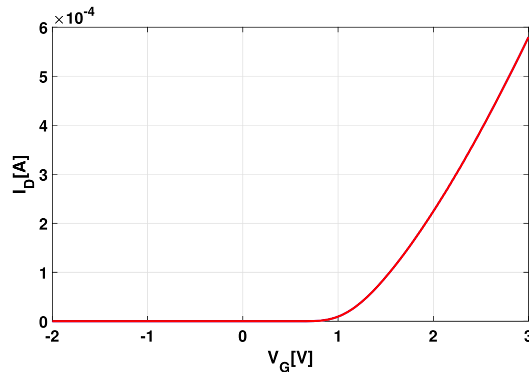


Fig. 9. The transfer characteristic of an ideal n-channel MOSFET with $V_D = 0.5$ V.

the transfer characteristic on the cross-section $y = 0$ μm are shown in Fig. 10. From these plots we find that a negative gate bias ($V_G < 0$) will make holes accumulate near the oxide-semiconductor interface. And in this accumulation case, no current flows in the device. If a positive voltage ($V_G > 0$) is applied to the gate, electrons start to accumulate near the oxide-semiconductor interface. When the number of electrons at the surface is greater than the number of holes, the surface is inverted, see Fig. 10(c) (weak inversion case) and Fig. 10(d) (strong inversion case).

Next, we consider the subthreshold characteristic still with the drain bias V_D fixed at 0.5 V, see Fig. 11. The above transfer characteristic illustrates that the threshold voltage $V_T = 0.85$ V. Fig. 11 shows the exponential dependence of I_D on $(V_G - V_T)$ for $V_G < V_T$, and the subthreshold swing S, defined as $[\partial(lgI_D)/\partial V_G]^{-1}$, is around $88mV$.

At last, we study the output characteristics, i.e., the drain current I_D versus the drain bias V_D, of the ideal n-channel MOSFET for different gate biases, which are plotted in Fig. 12. We consider the situation that the gate bias V_G is higher than the threshold voltage V_T. Therefore, an inversion is caused at the oxide-semiconductor interface. Initially, the drain bias V_D is small, then the drain current I_D is proportional to V_D, which is called the linear region. As the drain bias V_D increases, it eventually reaches $V_{\text{Dsat}} = V_G - V_T$. The drain current I_D gradually reaches the saturation region where I_D is a constant regardless of an increase in the drain bias. But in Fig. 12, we notice that the drain currents calculated by our method still increase slightly in the saturation region, especially with high fixed gate biases, which implies that our scheme may not be perfect.

In Fig. 13 we plot the electron concentration distributions corresponding to the output characteristic with a fixed gate bias $V_G = 3.4$ V on the cross-section $y = 0$ μm. From these plots, we find that the depletion region near the drain becomes wider as the drain bias increases.

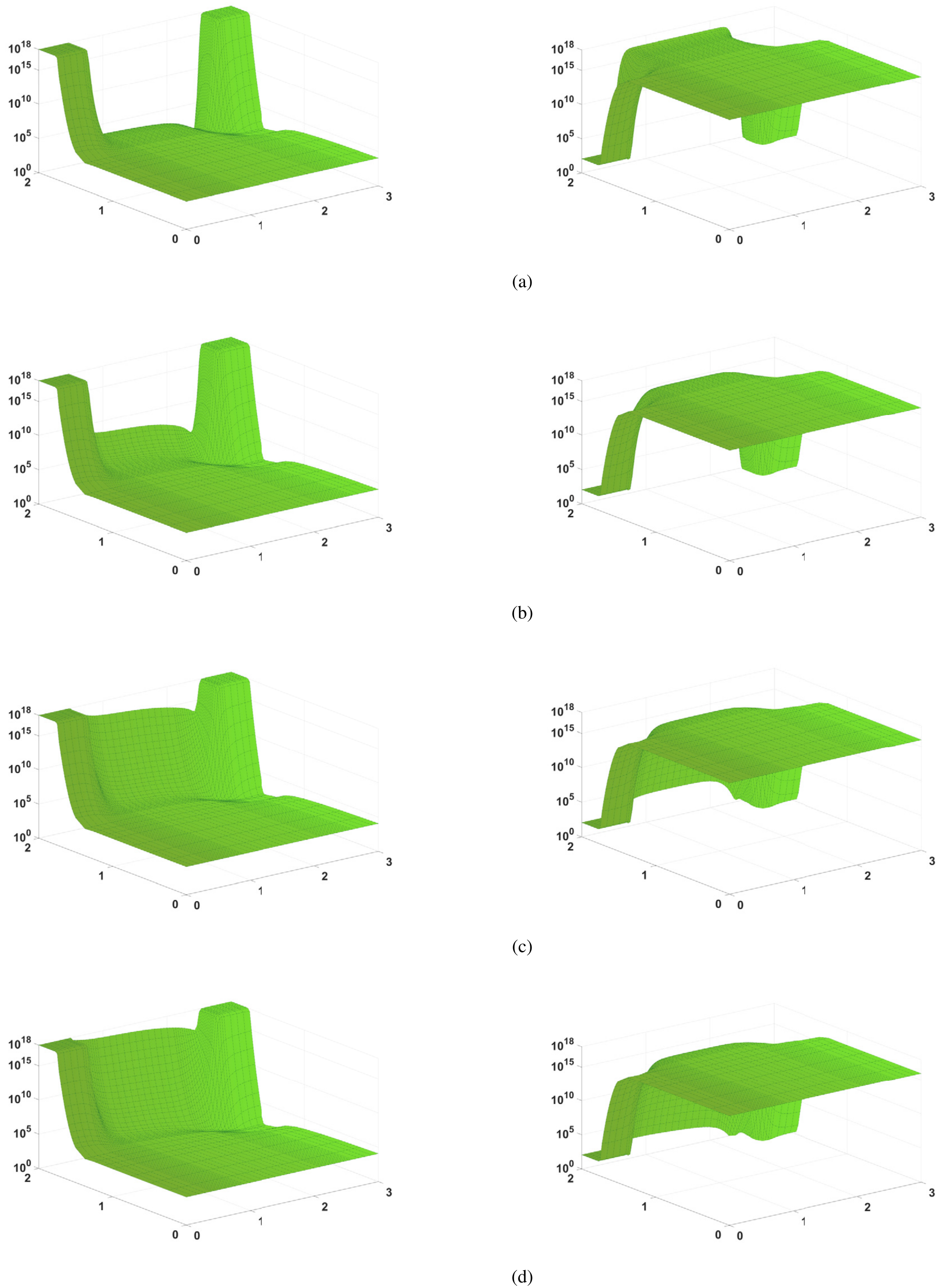


Fig. 10. Electron and hole concentration distributions on the cross-section $y = 0$ μm with a fixed $V_D = 0.5$ V for different gate biases V_G: (a) $V_G = -0.5$ V (b) $V_G = 0.0$ V (c) $V_G = 0.7$ V (d) $V_G = V_T = 0.85$ V.

For a general $SiO_2 - Si$ MOSFET, the work function difference ϕ_{ms}, $SiO_2 - Si$ interface traps, and oxide charges will affect the value of the threshold voltage V_T. The work function difference is related to the material of the metal placed on the gate. Without loss of generality, in the following part of this numerical example, we can still take $\phi_{\text{ms}} = 0$ and only consider the effect of the fixed charges located within approximately $3nm$ from the $SiO_2 - Si$ interface, assuming that the interface traps,

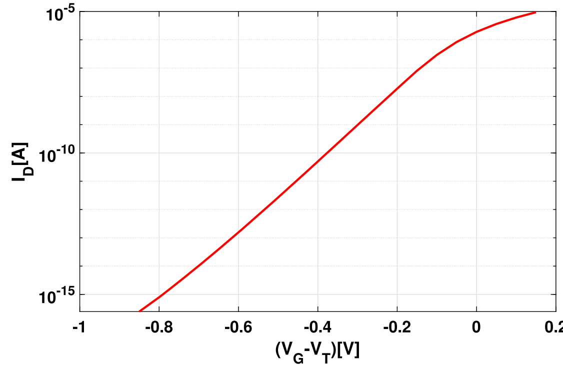


Fig. 11. The subthreshold characteristic of an ideal n-channel MOSFET with $V_D = 0.5$ V.

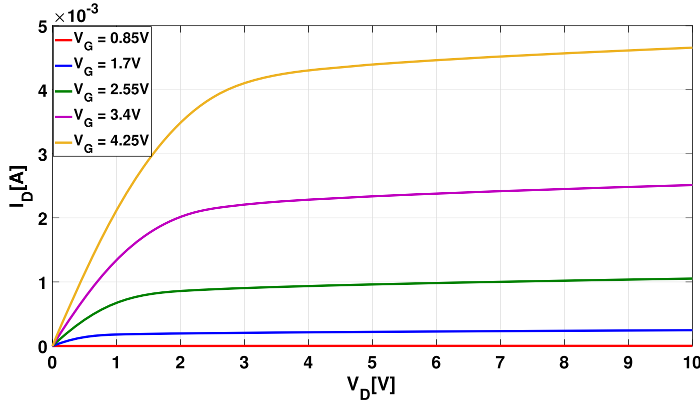


Fig. 12. The output characteristics of an ideal n-channel MOSFET.

oxide traps, and mobile ionic charges in the oxide are all negligible. Fig. 14 shows the transfer characteristics for different fixed charge densities Q_f with a fixed drain bias $V_D = 0.5$ V. It is obvious that the fixed charges within the oxide can lower the threshold voltage V_T. And when the fixed charge density reaches a certain value, such as $Q_f/q = 5.0e11$ cm^{-2}, a substantial drain current can flow at zero gate bias $V_G = 0$ V. This type of MOSFET is called the normally-on (depletion) n-channel MOSFET corresponding to the normally-off (enhancement) n-channel MOSFET introduced at the beginning of this numerical example. In Fig. 15, we plot the electron and hole concentration distributions corresponding to the transfer characteristic with the fixed charge density $Q_f/q = 5.0e11$ cm^{-2} on the cross-section $y = 0$ μm. From these plots, we see that an n-channel exists at zero gate bias, even at a negative gate bias, which once again verifies that a substantial current may flow at $V_G = 0$. The output characteristics of this normally-on n-channel MOSFET with the fixed charge density $Q_f/q = 5.0e11$ cm^{-2} are presented in Fig. 16. The slight increase of the drain current in the saturation region may be caused by our numerical method and we are trying to find a way to solve this problem.

5.4. Devices of different sizes

To further evaluate the robustness of scheme A4, we apply it to devices of different sizes, a 100 nm $p-n$ junction and a 300 μm n-channel MOSFET. We first scale down the $p-n$ junction depicted in Fig. 1 by a factor of 100 and set the doping profile as follows:

$$C = N_D - N_A = \begin{cases} -10^{20}, & \text{in p-type region,} \\ 10^{20}, & \text{in n-type region.} \end{cases}$$

Fig. 17 shows the ideal current-voltage characteristic of the 100 nm $p-n$ junction, and its trend is the same as that in Fig. 4.

Then we scale the n-channel MOSFET depicted in Fig. 8 by a factor of 100 and set the doping profile as

$$C = N_D - N_A = \begin{cases} -10^{13}, & \text{in p-type region,} \\ 10^{15}, & \text{in } n^+\text{-type region.} \end{cases}$$

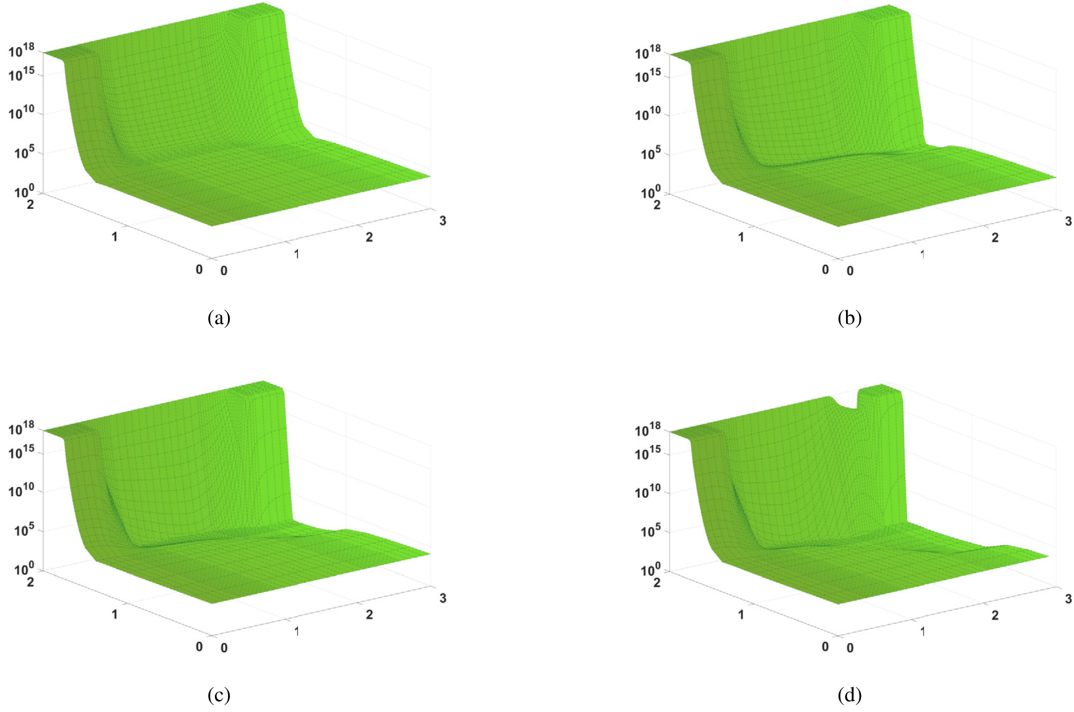


Fig. 13. Electron concentration distributions on the cross-section $y = 0$ μm with a fixed $V_G = 3.4$ V for different drain biases V_D: (a) $V_D = 0.05$ V (b) $V_D = 0.5$ V (c) $V_D = V_{\text{Dsat}} = 2.55$ V (d) $V_D = 10.0$ V.

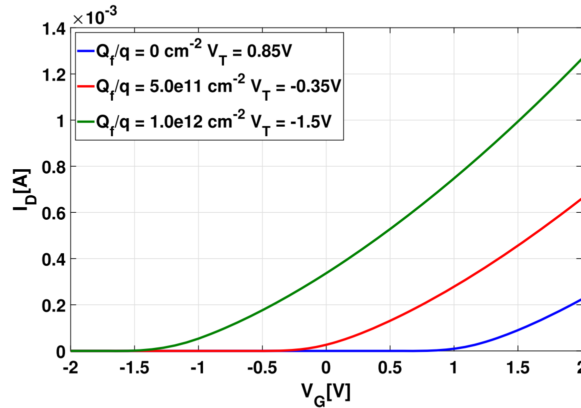


Fig. 14. The transfer characteristics for different fixed charge densities Q_f with $V_D = 0.5$ V.

We only consider the ideal 300 μm n-channel MOSFET. Its transfer characteristic with a fixed drain bias $V_D = 0.5$ V and output characteristics with different fixed gate biases are separately shown in Fig. 18 and Fig. 19. From Fig. 18, we note that the threshold voltage V_T of this ideal device is 1.0 V. Fig. 19 shows that the drain current I_D is proportional to the drain bias V_D in the linear region. When V_D reaches $V_{\text{Dsat}} = V_G - V_T$, the drain current I_D gradually reaches a saturation region. In the saturation region, the drain current increases slightly. The trends of the output characteristics are the same as that in Fig. 12.

The above numerical experiments illustrate that scheme A4 can produce rational numerical results for devices of different sizes.

6. Conclusion

To solve the three-dimensional convection-dominated continuity equations in the DD model, we propose a series of finite element discretization methods. These methods firstly transform the continuity equations into self-adjoint equations with exponentially behaved coefficients (e^{ψ} or $e^{-\psi}$) by employing the Slotboom variables. Because the flux density varies

(a)

(b)

(c)

Fig. 15. Electron and hole concentration distributions on the cross-section $y = 0$ μm with the fixed charge density $Q_f/q = 5.0e11$ cm^{-2} and the drain bias $V_D = 0.5$ V for different gate biases V_G: (a) $V_G = -0.35$ V (b) $V_G = 0.0$ V (c) $V_G = 0.35$ V.

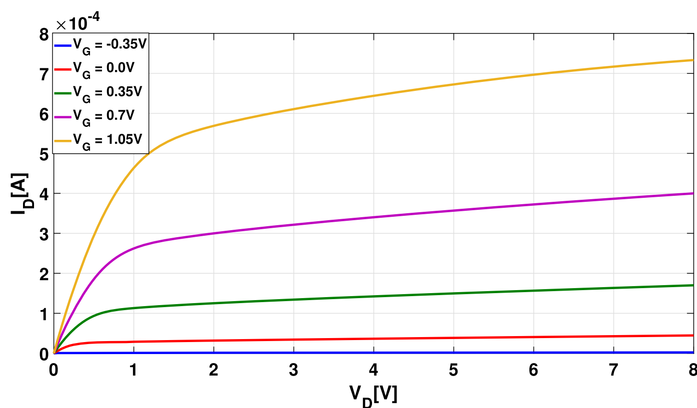


Fig. 16. The output characteristics of a normally-on *n*-channel MOSFET with the fixed charge density $Q_f/q = 5.0e11$ cm^{-2}.

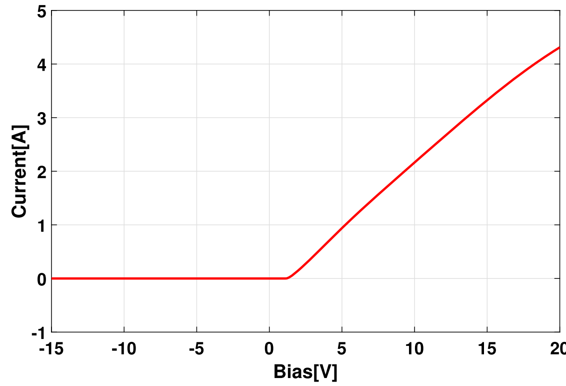


Fig. 17. Ideal current-voltage characteristic of the 100 nm $p-n$ junction.

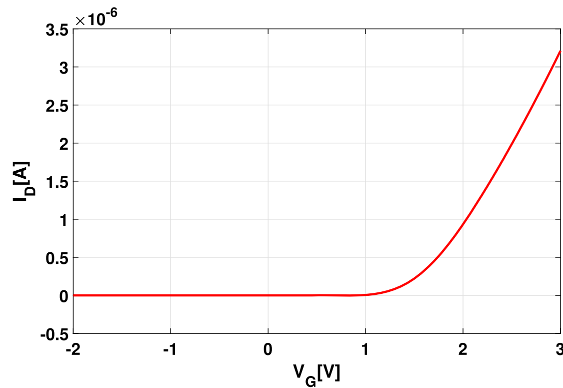


Fig. 18. The transfer characteristic of an ideal 300 μm n-channel MOSFET with $V_D = 0.5$ V.

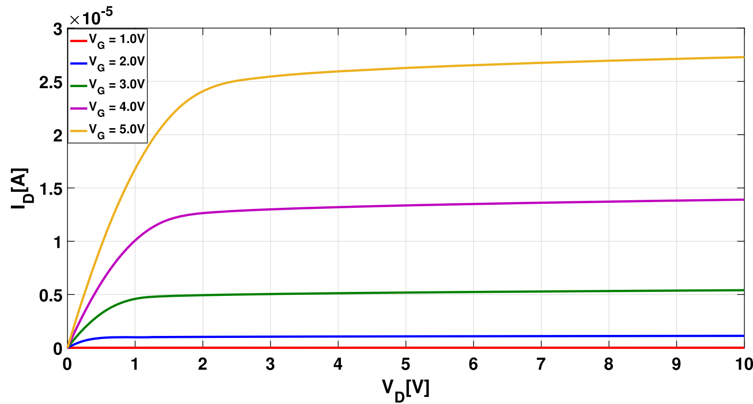


Fig. 19. The output characteristics of an ideal 300 μm n-channel MOSFET.

moderately in the whole domain, we approximate it with a constant vector on each tetrahedral element. Correspondingly, the exponential coefficient is also approximated with a constant on the whole tetrahedron or each edge of the tetrahedron. In this work, we propose four different schemes based on different averaging techniques, denoted as A1-A4, to obtain the average of the exponential coefficient. The first one calculates the harmonic average directly on a whole tetrahedral element. The other three schemes calculate the average of the exponential coefficient respectively on the edge of the tetrahedral element. These methods can deal with the layer oscillation problems and guarantee the conservation of the computed terminal currents with our terminal current evaluation approach. In addition, derivation of our methods is simple and doesn't require the construction of the dual Voronoi grid, which makes their parallel implementation easy. We first use a simple cube test to check the accuracy of our methods. Simulations of two realistic three-dimensional semiconductor devices, a $p-n$ junction and an n-channel MOSFET, further validate the accuracy and stability of the methods. According

to the numerical results and simple analysis, we conclude that scheme A4 can produce more accurate numerical solutions than schemes A1-A3, particularly under high bias conditions. Our numerical results also show that scheme A4 can work with a larger bias range than the Zlámal finite element method, and it performs better than the FVSG method and a type of tetrahedral mixed FEM [2] on poor-quality grids. Numerical experiments about the n-channel MOSFET indicate that scheme A4 can simulate rich physical properties of this device well. In the future, we will further improve our methods and apply them to other numerical simulation fields, such as nanomaterial simulations.

CRediT authorship contribution statement

Qianru Zhang: Conceptualization, Methodology, Software, Writing – original draft. **Qin Wang:** Conceptualization, Software, Writing – review & editing. **Linbo Zhang:** Funding acquisition, Conceptualization, Software, Writing – review & editing. **Benzhuo Lu:** Funding acquisition, Supervision, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Sheng Gui and Yuhui Ni for their help in producing grids for semiconductor devices.
Funding: This work was supported by the National Key R & D Program of China (Grant Nos. 2019YFA0709600 and 2019YFA0709601) and the National Natural Science Foundation of China (Grant Nos. 11771435 and 22073110).

References

[1] M. Zlámal, Finite element solution of the fundamental equations of semiconductor devices. I, Math. Comput. 46 (1986) 27–43.
[2] J. Miller, S. Wang, A tetrahedral mixed finite element method for the stationary semiconductor continuity equations, SIAM J. Numer. Anal. 31 (1994) 196–216.
[3] W. Van Roosbroeck, Theory of flow of electrons and holes in germanium and other semiconductors, Bell Syst. Tech. J. 29 (1950) 560–607.
[4] D.L. Scharfetter, H.K. Gummel, Large-signal analysis of a silicon read diode oscillator, IEEE Trans. Electron Devices 16 (1969) 64–77.
[5] R.E. Bank, W. Coughran Jr, L.C. Cowsar, The finite volume Scharfetter-Gummel method for steady convection diffusion equations, Comput. Vis. Sci. 1 (1998) 123–136.
[6] S.-W. Cheng, T.K. Dey, J. Shewchuk, Delaunay Mesh Generation, CRC Press, 2012.
[7] R.E. Bank, J.F. Bürgler, W. Fichtner, R.K. Smith, Some upwinding techniques for finite element approximations of convection-diffusion equations, Numer. Math. 58 (1990) 185–202.
[8] A.N. Brooks, T.J. Hughes, Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations, Comput. Methods Appl. Mech. Eng. 32 (1982) 199–259.
[9] G. Carey, M. Sharma, Semiconductor device modeling using flux upwind finite elements, Compel (1989).
[10] T.J. Hughes, Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods, Comput. Methods Appl. Mech. Eng. 127 (1995) 387–401.
[11] C. Johnson, Numerical Solution of Partial Differential Equations by the Finite Element Method, Courier Corporation, 2012.
[12] C. Johnson, U. Navert, J. Pitkaranta, Finite element methods for linear hyperbolic problems, Comput. Methods Appl. Mech. Eng. 45 (1984) 285–312.
[13] S. Micheletti, Stabilized finite elements for semiconductor device simulation, Comput. Vis. Sci. 3 (2001) 177–183.
[14] M. Sharma, G.F. Carey, Semiconductor device simulation using adaptive refinement and flux upwinding, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 8 (1989) 590–598.
[15] Q. Wang, H. Li, L. Zhang, B. Lu, A stabilized finite element method for the Poisson–Nernst–Planck equations in three-dimensional ion channel simulations, Appl. Math. Lett. 111 (2021) 106652.
[16] J. Xu, L. Zikatanov, A monotone finite element scheme for convection-diffusion equations, Math. Comput. 68 (1999) 1429–1446.
[17] F. Liu, J.J. Miller, Inverse average type tetrahedral finite-element schemes for the stationary semiconductor device equations, J. Comput. Appl. Math. 44 (1992) 77–94.
[18] P.A. Markowich, M.A. Zlámal, Inverse-average-type finite element discretizations of selfadjoint second-order elliptic problems, Math. Comput. 51 (1988) 431–449.
[19] J. Miller, W. Schilders, S. Wang, Application of finite element methods to the simulation of semiconductor devices, Rep. Prog. Phys. 62 (1999) 277.
[20] Shuonan Wu, Jinchao Xu, Simplex-averaged finite element methods for H (grad), H (curl), and H (div) convection-diffusion problems, SIAM J. Numer. Anal. 58 (2020) 884–906.
[21] L. Angermann, S. Wang, Three-dimensional exponentially fitted conforming tetrahedral finite elements for the semiconductor continuity equations, Appl. Numer. Math. 46 (2003) 19–43.
[22] F. Brezzi, L.D. Marini, P. Pietra, Two-dimensional exponential fitting and applications to semiconductor device equations, Ist., Consiglio, 1987.
[23] F. Brezzi, L.D. Marini, P. Pietra, Numerical simulation of semiconductor devices, Comput. Methods Appl. Mech. Eng. 75 (1989) 493–514.
[24] L.D. Marini, P. Pietra, New mixed finite element schemes for current continuity equations, Compel (1990).
[25] Pavel Bochev, Kara Peterson, A parameter-free stabilized finite element method for scalar advection-diffusion problems, Open Math. 11 (2013) 1458–1477.
[26] Pavel Bochev, Kara Peterson, Xujiao Gao, A new control volume finite element method for the stable and accurate solution of the drift–diffusion equations on general unstructured grids, Comput. Methods Appl. Mech. Eng. 254 (2013) 126–145.
[27] Pavel Bochev, Mauro Perego, Kara Peterson, Formulation and analysis of a parameter-free stabilized finite element method, SIAM J. Numer. Anal. 53 (2015) 2363–2388.
[28] R. Sacco, E. Gatti, L. Gotusso, The patch test as a validation of a new finite element for the solution of convection-diffusion equations, Comput. Methods Appl. Mech. Eng. 124 (1995) 113–124.

[29] R. Sacco, M. Stynes, et al., Finite element methods for convection-diffusion problems using exponential splines on triangles, Comput. Math. Appl. 35 (1998) 35–45.
[30] S. Wang, A novel exponentially fitted triangular finite element method for an advection–diffusion problem with boundary layers, J. Comput. Phys. 134 (1997) 253–260.
[31] S. Wang, A new exponentially fitted triangular finite element method for the continuity equations in the drift-diffusion model of semiconductor devices, ESAIM: Math. Model. Numer. Anal. 33 (1999) 99–112.
[32] F. Brezzi, L. Marini, P. Markowich, P. Pietra, On some numerical problems in semiconductor device simulation, in: Mathematical Aspects of Fluid and Plasma Dynamics, Springer, 1991, pp. 31–42.
[33] F. Brezzi, L. Marini, S. Micheletti, P. Pietra, R. Sacco, S. Wang, Discretization of semiconductor device problems (I), Handb. Numer. Anal. 13 (2005) 317–441.
[34] F. Brezzi, L.D. Marini, S. Micheletti, P. Pietra, R. Sacco, Stability and error analysis of mixed finite-volume methods for advection dominated problems, Comput. Math. Appl. 51 (2006) 681–696.
[35] J. Miller, S. Wang, A triangular mixed finite element method for the stationary semiconductor device equations, ESAIM: Math. Model. Numer. Anal. 25 (1991) 441–463.
[36] V. John, T. Mitkova, M. Roland, K. Sundmacher, L. Tobiska, A. Voigt, Simulations of population balance systems with one internal coordinate using finite element methods, Chem. Eng. Sci. 64 (2009) 733–741.
[37] V. John, E. Schmeyer, Finite element methods for time-dependent convection–diffusion–reaction equations with small diffusion, Comput. Methods Appl. Mech. Eng. 198 (2008) 475–494.
[38] 王芹, 马召灿, 白石阳, 张林波, 卢本卓, 李鸿亮, 三维半导体器件漂移扩散模型的并行有限元方法研究, 数值计算与计算机应用 41 (2020) 85.
[39] Matteo Patriarca, Patricio Farrell, Jürgen Fuhrmann, Thomas Koprucki, Highly accurate quadrature-based Scharfetter–Gummel schemes for charge transport in degenerate semiconductors, Comput. Phys. Commun. 235 (2019) 40–49.
[40] J.W. Slotboom, Computer-aided two-dimensional analysis of bipolar transistors, IEEE Trans. Electron Devices 20 (1973) 669–679.
[41] D. Klaassen, A unified mobility model for device simulation—I. Model equations and concentration dependence, Solid-State Electron. 35 (1992) 953–959.
[42] I. Babuška, J.E. Osborn, Generalized finite element methods: their performance and their relation to mixed methods, SIAM J. Numer. Anal. 20 (1983) 510–536.
[43] F. Brezzi, M. Fortin, Mixed and Hybrid Finite Element Methods, vol. 15, Springer Science & Business Media, 2012.
[44] S.M. Sze, Semiconductor Devices: Physics and Technology, John Wiley & Sons, 2008.
[45] H. Gummel, A self-consistent iterative scheme for one-dimensional steady state transistor calculations, IEEE Trans. Electron Devices 11 (1964) 455–465.
[46] L.-B. Zhang, et al., A parallel algorithm for adaptive local refinement of tetrahedral meshes using bisection, Numer. Math., Theory Methods Appl. 2 (2009) 65–89.
[47] H. Si, TetGen, a Delaunay-based quality tetrahedral mesh generator, ACM Trans. Math. Softw. 41 (2015) 11.