Quantum transport models based on NEGF and empirical pseudopotentials for accurate modeling of nanoscale electron devices

Marco Pala, David Esseni

Quantum transport models based on NEGF and empirical pseudopotentials for accurate modelling of nanoscale electron devices

Marco G. Pala[1, a)] and David Esseni[2, b)]
[1)] *Centre de Nanosciences et de Nanotechnologies, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 10 Boulevard Thomas Gobert, 91120 Palaiseau, France*
[2)] *DPIA, University of Udine, Via delle Scienze 206, 33100 Udine, Italy*

(Dated: 21 July 2019)

This paper presents significant new developments concerning the full band, quantum simulation of nanostructured systems and nanoscale electron devices based on an empirical pseudopotential Hamiltonian. We demonstrate that the method is of general applicability, in fact we show results for planar, ultra-thin body FETs and also for three-dimensional, nanowire FETs, we deal with different crystal orientations and account for possible stress/strain conditions in the simulated systems. Some of the simulations reported in this paper have been made computationally viable by the substantial improvements of the numerical efficiency compared to our previous pseudopotentials based methodology.

Most of the methods and algorithms discussed in this paper are not specific to an empirical pseudopotential Hamiltonian, on the contrary they can be applied also to different Hamiltonians described with a plane waves basis, which are frequently employed for *ab-initio*, Density Functional Theory based calculations. The application of the methodologies described in this work may thus be more far reaching than it is illustrated by the case studies explicitly addressed in the present paper.

Keywords: Quantum transport, Green's functions, empirical pseudopotentials, strain, nanowires

I. INTRODUCTION

The active region of modern nanoelectronic and nanophotonic devices frequently consists of semiconductor materials structured at truly nanometric dimensions, either in the form quantum wells and ultra-thin semiconductor film on insulator (UTB-SOI), or in three dimensionally carved architectures such as Fin-FETs, multigate FETs (MuGFETs), and gate-all-around (GAA) nanowire transistors[1]. Quantum mechanical effects are no longer limited to a splitting in subband of the electronics structure of the underlying semiconductors, instead quantum transport phenomena have become important such as source-drain tunnelling in CMOS nanoscale transistors[2–4], and band-to-band-tunnelling (BTBT) in Tunnel-FETs (TFETs)[5,6]. The relevance of quantum mechanical effects in nanoscale FETs is such that CMOS transistors have been recently proposed as a platform for quantum computing[7].

As a matter of fact modern CMOS FETs resemble quantum constrictions connecting the source/drain carrier reservoirs[2], however a full-band quantum transport formalism is theoretically complex and computationally very demanding, so that in the electron devices community it is still popular to use simplified Hamiltonians based either on the effective mass approximation (EMA)[8–10], or on the $\mathbf{k}\cdot\mathbf{p}$ approach[11–13]. Both the EMA and $\mathbf{k}\cdot\mathbf{p}$ methods are known to provide a description of low dimensional systems that is limited to the vicinity of a given symmetry point in the reduced zone of the underlying semiconductor, as well as known to have accuracy limitations in strongly confined systems.

While the empirical tight-binding (TB) method still remains the most mature full-band approach for quantum transport device simulations based on the non-equilibrium Green's function (NEGF) formalism[14,15], an increasing interest has been recently growing for plane-waves based methods. The plane-waves approach provides both the electronic structure and the atomistic wave-function and, moreover, it is the basis most frequently used in Density Functional Theory (DFT) *ab-initio* calculations, albeit at the cost of a quite large basis set that makes its use very challenging for transport calculations.

An Empirical Pseudopotentials (EP) Hamiltonian is an interesting compromise between accuracy and computational burden and has only a few fitting parameters[16–20]. Some recent papers have explored the use of the EP method for full-band quantum transport modelling in carbon nanotubes[21], and more recently in ultra-thin-body FETs[22,23], in nanowire FETs with a body-size of 0.39 nm, in graphene nanoribbon transistors[24–26], and usually by employing some variants of the quantum-transmitting-boundary method.

The authors of this paper have recently made an attempt to use an NEGF based transport model and an EP Hamiltonian[27,28], and reported results for three dimensional systems with no quantum confinement, as well as for ultra-thin body devices. Admittedly, however, the computational burden of EP based transport models remains heavy and, with the approach developed in Ref. 27 and 28, the simulation of nanowire devices still was impractical with conventional computational resources in

[a)] Electronic mail: marco.pala@c2n.upsaclay.fr
[b)] Electronic mail: david.esseni@uniud.it

terms of memory and number of cores.

In this work we substantially extend our previous, brief communication in Ref. 29, and report recent improvements in NEGF simulations based on EP, that allowed us to reduce significantly the size of the blocks of the block tridiagonal Hamiltonian matrix, which is the main figure setting the overall computational burden of the NEGF method. A new approach for the quantum confinement treatment, an adjustment of the discretization scheme, and the development of a new approach for the calculation of the contact self-energies have been synergetically used to achieve a large reduction of the computational complexity, which eventually allowed us to report in this paper simulations for nanowire transistors with technologically relevant geometrical dimensions.

The paper is organized as follows. In Sec. II we first present the formulation of the pseudopotentials and the quantum confinement operator, and then report their application to bandstructure calculations in both bulk materials and nanostructured semiconductors; the inclusion of different strain conditions and crystal orientations is also discussed. In Sec. III we present the model for a self-consistent calculation of carrier densities, terminal currents and electrostatics, and in Sec. IV we illustrate some examples of complete device simulations including an UTB-SOI, p-type Si MOSFET and a nanowire, n-type InAs Tunnel-FET. Our final remarks and outlook are presented in Sec. V.

II. ELECTRONIC STRUCTURE CALCULATIONS BASED ON EMPIRICAL PSEUDOPOTENTIALS

In this work we use the empirical pseudopotential method for band-structure and transport calculations. The Schrödinger equation for a local EP model and a bulk semiconductor crystal can be written as[30]

$$\sum_{\mathbf{G}'} \mathbf{H}_{\mathbf{k}}(\mathbf{G},\mathbf{G}') B_{\mathbf{k}}(\mathbf{G}') = E_b(\mathbf{k})\, B_{\mathbf{k}}(\mathbf{G}) \qquad (1)$$

and the Hamiltonian matrix in $\mathbf{K}$ space is

$$\mathbf{H}_{\mathbf{k}}(\mathbf{G},\mathbf{G}') = T(\mathbf{k}+\mathbf{G})\delta_{\mathbf{G},\mathbf{G}'} + V_L(\mathbf{G}-\mathbf{G}') \qquad (2)$$

where $V_L(\mathbf{G})$ is a spectral component of the local pseudopotential, $T(\mathbf{K})$ is the kinetic energy operator and $E_b(\mathbf{k})$ is the energy dispersion of the bulk semiconductor.

Throughout this paper we use $\mathbf{k}$ to denote a Bloch wave-vector in the reduced zone of the bulk crystal and $\mathbf{K}$=($\mathbf{k}$+$\mathbf{G}$) to indicate a wave-vector in the extended reciprocal space, with $\mathbf{G}$ being a reciprocal lattice vector.

For a diamond semiconductor (e.g. Si or Ge) or a zinc-blende compound (e.g. GaAs, InAs) we have[30,31]

$$V_L(\mathbf{G}) = U_S(|\mathbf{G}|)\cos[\mathbf{G}\cdot\boldsymbol{\tau}] + iU_A(|\mathbf{G}|)\sin[\mathbf{G}\cdot\boldsymbol{\tau}] \qquad (3)$$

where $\boldsymbol{\tau}$=$(1/8)a_0(1,1,1)$ is the atomic basis vector, while $U_S(|\mathbf{G}|)$, $U_A(|\mathbf{G}|)$ are respectively the symmetric and antisymmetric form factors (with $U_A(|\mathbf{G}|)$ being null for diamond materials).

A real space discretization is practically indispensable for transport modelling with the NEGF approach (see also Sec. III and further details in 28), and in this work we use a simple second order, centered difference discretization of the kinetic energy operator given by

$$T(\mathbf{k}+\mathbf{G}) = 2t_0 \sum_{s=x,y,z} \{1-\cos[(k_s+G_s)d]\} \qquad (4)$$

where t_0=$\hbar^2/2m_0d^2$. In all spatial directions s=$\{x,y,z\}$ we employ the same discretization step d=a_0/N_d. Throughout the paper we used $N_d = 30$, which ensues that the bands obtained with either the discretized or the continuous formulation of the kinetic energy operator are practically identical.

The use of a second order discretization is a first difference compared to our previous methodology relying on higher order discretization schemes[28], and a second important difference is the operator used for quantum confinement for nanostructured systems discussed in Sec. II A.

Both these aspects contribute to reduce the size of the blocks of the block tridiagonal Hamiltonian matrix to be used in NEGF transport calculations, as it will be further clarified in Sec. III.

A. A local quantum confinement operator

In order to model a 2D electron gas in a quantum well (QW) or a 1D electron gas in a nanowire (NW), we here propose a new method compared to Ref. 28, because our previous approach resulted in a non local (in real space) confining operator setting a lower limit to the size of the blocks of the Hamiltonian matrix.

More specifically, in order to describe quantum confinement as a local operator in real space, we here introduce a pseudo-oxide region, whose only purpose is to setup a conduction and valence band discontinuity with respect to the semiconductor, such that carriers are effectively confined in the semiconductor region.

If we let $V_{\rm sct}(\mathbf{r})$ and $V_{\rm ox}(\mathbf{r})$ denote the pseudopotentials describing respectively the actual semiconductor and the pseudo-oxide, then we write the overall pseudopotential $V_{2D}(\mathbf{r})$ for a QW as

$$V_{2D}(\mathbf{r}) = V_{\rm sct}(\mathbf{r}) + V_{\rm cnf}(\mathbf{r})\,\theta_{2D}(z)\,, \qquad (5)$$

where z is the confinement direction, the potential $V_{\rm cnf}(\mathbf{r})$ is defined as $V_{\rm cnf}(\mathbf{r})$=$[V_{\rm ox}(\mathbf{r})$−$V_{\rm sct}(\mathbf{r})]$, and $\theta_{2D}(z)$ is a box function such that $\theta_{2D}(z) = 0$ for $|z|\leq T_{\rm sct}/2$ and 1 otherwise, where $T_{\rm sct}$ is the thickness of the semiconductor film.

It is important to notice that we here assume that the pseudo-oxide has the same lattice constant a_0 as the corresponding semiconductor, so that both the direct and the reciprocal space of the two materials coincide. Such a pseudo-oxide is essentially a computational tool, whose

EP parameters can be adjusted to obtain the desired values for the conduction and valence band discontinuity with the semiconductor. As an example, Fig. 1(a) illustrates the energy dispersion of silicon (in black) and the pseudo-oxide material (in red), showing that the pseudo-oxide has a direct bandgap of about 9eV, and it results in a conduction and valence band discontinuity with silicon of respectively 3 eV and 4.7 eV, which are values representative of the Si-SiO_2 system[32]. Similarly, Fig. 1(b) shows the energy dispersion of InAs (in blue) and the associated pseudo-oxide (in magenta), with discontinuities of the conduction and valence band of 3.8 eV and 4.5 eV, respectively. Tab. I reports the parameters of the empirical pseudopotential model for silicon, InAs and for the two corresponding pseudo-oxides used for the calculations in Fig. 1.

An important feature of the $V_{2D}(\mathbf{r})$ defined in Eq. (5) is that it is by definition local in real space, so that the V_{2D} in $\mathbf{K}$ space can be readily written as

$$V_{2D}(\mathbf{K}-\mathbf{K}') = V_{\rm sct}(\mathbf{G}-\mathbf{G}')\delta_{\mathbf{k},\mathbf{k}'} + \sum_{G_z''} V_{\rm cnf}(\mathbf{G}_{xy}-\mathbf{G}'_{xy}, G_z - G'_z - G''_z)\theta_{2D}(K_z - K'_z + G''_z)\delta_{\mathbf{k}_{xy},\mathbf{k}'_{xy}} \quad (6)$$

$$V_{2D}(\mathbf{K}-\mathbf{K}') = V_{\rm sct}(\mathbf{G}-\mathbf{G}')\delta_{\mathbf{k},\mathbf{k}'} + \sum_{G_z''} V_{\rm cnf}(\mathbf{G}_{xy}-\mathbf{G}'_{xy}, G_z - G'_z - G''_z) \times\theta_{2D}(K_z - K'_z + G''_z)\delta_{\mathbf{k}_{xy},\mathbf{k}'_{xy}} \quad (7)$$

where vectors are defined as $\mathbf{K}$=[$(\mathbf{k}_{xy}, k_z)$+$\mathbf{G}$], $\mathbf{K}'$=[$(\mathbf{k}'_{xy}, k'_z)$+$\mathbf{G}'$], $\mathbf{G}$=$(\mathbf{G}_{xy}, G_z)$, and we have exploited the fact that the real space product $V_{\rm cnf}(\mathbf{r})\theta_{2D}(z)$ in Eq. (5) transforms into a convolution in reciprocal space. Hence for a 2D electron gas having z as the quantization direction, the overall Hamiltonian in $\mathbf{K}$ space reads

$$\mathbf{H}_{\mathbf{k}_{xy}}(\mathbf{K},\mathbf{K}') = T(\mathbf{k}+\mathbf{G})\delta_{\mathbf{G},\mathbf{G}'}\delta_{k_z,k'_z} + V_{2D}(\mathbf{K}-\mathbf{K}') \quad (8)$$

and the energy dispersion is obtained by solving the eigenvalue problem associated to $\mathbf{H}_{\mathbf{k}_{xy}}(\mathbf{K},\mathbf{K}')$ with $\mathbf{k}_{xy}$ varying in the 2D reduced zone.

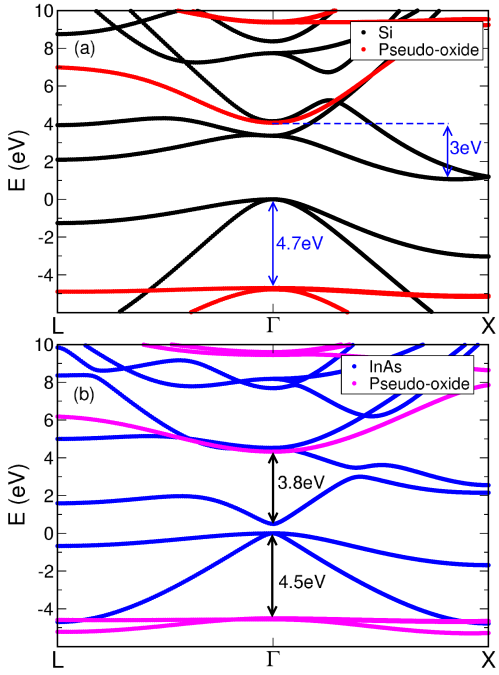


FIG. 1: (a) Bandstructure of the pseudo-oxide (red) compared to that of bulk silicon (black). (b) Bandstructure of the pseudo-oxide (magenta) compared to that of bulk InAs (blue).

	U_{S3}	U_{S8}	U_{S11}	U_{A3}	U_{A4}	U_{A11}
Si	-0.224	0.055	0.072	0	0	0
pseudo-ox. on Si	-0.84	0.09	0.19	0	0	0
InAs	-0.22	0	0.05	0.08	0.05	0.03
pseudo-ox. on InAs	-0.64	0	0.14	0.225	0.14	0.08

TABLE I: EP parameters (in Ry) for Si and the corresponding pseudo-oxide (both having a 0.543 nm lattice constant), as well as for InAs and the corresponding pseudo-oxide on InAs (having a 0.608 nm lattice constant).

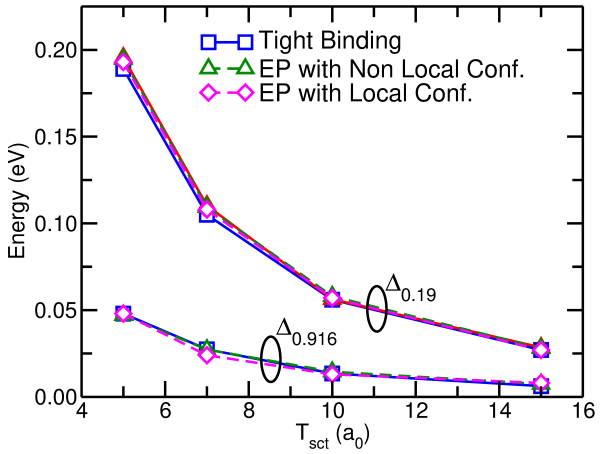


FIG. 2: Lowest ($\Delta_{0.916}$) and second lowest conduction subband minimum ($\Delta_{0.19}$) located respectively at (k_x,k_y)=(0,0) and (k_x,k_y)=(0.85,0) (in units of $2\pi/a_0$) for an ultra-thin silicon film versus the film thickness $T_{\rm sct}$. Results obtained with either the local confinement operator of this work or the non local confinement operator used in Ref. 28 are in good agreement, and they agree also with tight-binding results with parameters from[33]. Reproduced with permission from Ref. 29. Copyright 2018 IEEE.

As an illustration and a validation of the quantum confinement model described above, Fig. 2 shows the conduction band minima of an ultra-thin silicon quantum well versus the thickness of the silicon film. As it can be seen the EP calculations obtained with the new, local pseudopotential $V_{2D}(\mathbf{r})$ in Eqs. (5) and (6) agree well with the results of the non local confinement operator previously used by these authors[27], as well as with tight-binding calculations.

A completely similar approach can be used for a 1D electron gas in a NW having x as the unconstrained direction. In this case we can define an overall pseudopotential $V_{1D}(\mathbf{r})$

$$V_{1D}(\mathbf{r}) = V_{\rm sct}(\mathbf{r}) + V_{\rm cnf}(\mathbf{r})\,\theta_{1D}(y,z)\,, \tag{9}$$

where (y,z) is the confinement plane and $\theta_{1D}(y,z)$ is a box function such that $\theta_{1D}(y,z)=1$ in the pseudo-oxide region and 0 in the semiconductor. The $\mathbf{K}$ space confining operator for a 1D gas reads

$$V_{1D}(\mathbf{K}-\mathbf{K}') = V_{\rm sct}(\mathbf{G}-\mathbf{G}')\delta_{\mathbf{k},\mathbf{k}'} + \sum_{\mathbf{G}''_{yz}} V_{\rm cnf}(G_x-G'_x, \mathbf{G}_{yz}-\mathbf{G}'_{yz}-\mathbf{G}''_{yz})\theta_{1D}(\mathbf{K}_{yz}-\mathbf{K}'_{yz}+\mathbf{G}''_{yz})\delta_{k_x,k'_x}\,, \tag{10}$$

$$\begin{aligned} V_{1D}(\mathbf{K}-\mathbf{K}') &= V_{\rm sct}(\mathbf{G}-\mathbf{G}')\delta_{\mathbf{k},\mathbf{k}'} + \\ &\sum_{\mathbf{G}''_{yz}} V_{\rm cnf}(G_x-G'_x, \mathbf{G}_{yz}-\mathbf{G}'_{yz}-\mathbf{G}''_{yz}) \\ &\times\theta_{1D}(\mathbf{K}_{yz}-\mathbf{K}'_{yz}+\mathbf{G}''_{yz})\delta_{k_x,k'_x}\,, \end{aligned} \tag{11}$$

where $\mathbf{K}$=[$(k_x,\mathbf{k}_{yz})$+$\mathbf{G}$], $\mathbf{K}'$=[$(k'_x,\mathbf{k}'_{yz})$+$\mathbf{G}'$], $\mathbf{G}$=[$(G_x,\mathbf{G}_{yz})$ and the energy dispersion is described by an overall Hamiltonian similar to Eq. (8) that has to be solved by varying k_x in the 1D reduced zone.

B. Crystal orientation and strain

In all equations discussed so far we have implicitly assumed that transport and confining directions are aligned with the ⟨100⟩ directions of the underlying semiconductor[27,28], and the semiconductor is free of strain. However in electron devices simulations the Device Coordinate System (DCS) is frequently different from the Crystal Coordinate System (CCS) (see, for example, Fig. 6), and strain is an important engineering option, so that in this sub-section we briefly explain how these aspects have been included in our modelling framework.

Because the confinement and transport directions are defined in the DCS, if the DCS is other than the CCS then the reciprocal lattice vectors must be expressed in the DCS as $\mathbf{G}$=$\mathbf{R}_{CD}\mathbf{G}_c$, where $\mathbf{R}_{CD}$ is a 3×3 rotation matrix from CCS to DCS, and $\mathbf{G}_c$ are the well known lattice vectors in the CCS; the atomic basis vector $\boldsymbol{\tau}$ similarly transforms as $\boldsymbol{\tau}$=$\mathbf{R}_{CD}\boldsymbol{\tau}_c$. It is here worth to notice that the $\mathbf{G}$ vectors in the DCS set the reduced zone of the bulk semiconductor to be used for transport calculations, and that in our methodology the reduced zone has the shape of a prism[19,28]. For a 2D electron gas in the UTB FET of Fig. 6, for example, if we let $\mathbf{G}_{xx}$, $\mathbf{G}_{zz}$ denote the smallest $\mathbf{G}$ vectors aligned with respectively the x and z direction in the DCS, then the k_x range of the reduced zone is $-|\mathbf{G}_{xx}|/2\leq k_x<|\mathbf{G}_{xx}|/2$, which ensures that, for any $\mathbf{K}_{yz}$=[(k_y,k_z)+(G_y,G_z)], the corresponding K_x=k_x+G_x components cover with no voids the entire extended K_x range[28]. Then the $|\mathbf{G}_{zz}|$ sets the k_z range of the reduced zone in the confinement direction as $-|\mathbf{G}_{zz}|/2\leq k_z<|\mathbf{G}_{zz}|/2$, while the k_y range of the reduced zone is finally established by the fact that the volume of the reduced zone must be $4(2\pi/a_0)^3$ for the unstrained lattice. A few examples of reduced zones are illustrated in the table of Fig. 6. A more detailed discussion about the reduced zone of the underlying semiconductor crystal that should be used in NEGF simulations based on an EP Hamiltonian can be found in Sec. V of Ref. 28.

The EP method can naturally include the effects of strain[34]. If we denote with $\boldsymbol{\varepsilon}_c$ the 3×3 strain matrix in the CCS, the direct lattice vectors, $\mathbf{a}$, reciprocal lattice vectors, $\mathbf{b}$, and unit cell volume, Ω, of the strained lattice are given by[32]

$$\mathbf{a} = (\boldsymbol{I}_3 + \boldsymbol{\varepsilon}_c)\mathbf{a}^0; \quad \Omega = \Omega_0(1 + \varepsilon_{c,xx} + \varepsilon_{c,yy} + \varepsilon_{c,zz}); \quad \mathbf{b}_1 = \frac{2\pi}{\Omega}(\mathbf{a}_2 \times \mathbf{a}_3);\ \mathbf{b}_2 = \frac{2\pi}{\Omega}(\mathbf{a}_3 \times \mathbf{a}_1);\ \mathbf{b}_3 = \frac{2\pi}{\Omega}(\mathbf{a}_1 \times \mathbf{a}_2) \quad (12)$$

where $\mathbf{a}^0$ and Ω_0 are lattice vectors and unit cell volume of the unstrained lattice, and $\boldsymbol{I}_3$ is the 3×3 identity matrix.

Strain affects also the atomic basis vector $\boldsymbol{\tau}_c$=$[\boldsymbol{I}_3$+$\boldsymbol{\varepsilon}_c]\boldsymbol{\tau}_c^0$, where $\boldsymbol{\tau}_c^0$=$(1/8)a_0(1,1,1)$ is the basis vector of the unstrained lattice. While the deformation of the unit cell can be determined from macroscopic strain, the possible atomic rearrangement inside the unit cell requires additional information from *ab-initio* calculations and, particularly in the presence of shear strain, some adjustments to $\boldsymbol{\tau}$ have been proposed[35]. However we verified that such corrections have a practically negligible effect in the cases considered in this paper.

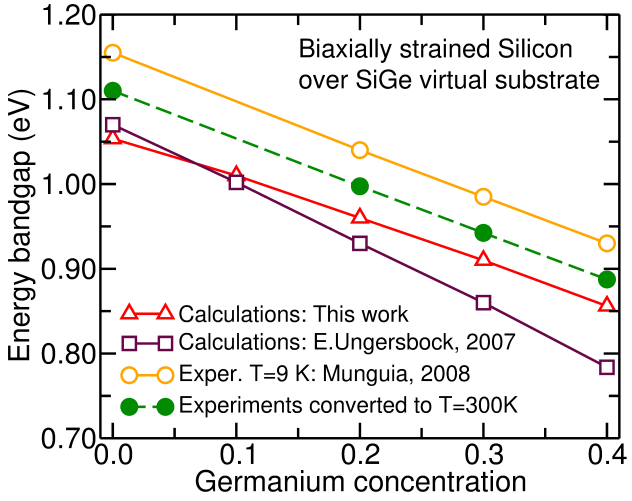


FIG. 3: Energy bandgap for biaxially strained silicon on a SiGe virtual substrate. Calculations of this work are in good agreement with experiments from[36], and with pseudopotential calculations from[37]. Experiments were converted from T=9 K to T=300 K by using $E_G(T)$=$E_G(T$=0$) - \alpha T^2/(T+\beta)$, with α =4.7·10^{-4} eV/K, β =655 K[38]. Reproduced with permission from Ref. 29. Copyright 2018 IEEE.

For an unstrained crystal the calculations based on empirical pseudopotentials can be carried out by using only three non null $U_L(|\mathbf{G}|)$ components for $|\mathbf{G}|$=$\sqrt{3}$, $\sqrt{8}$, $\sqrt{11}$. On the contrary in a strained lattice the $|\mathbf{G}|$ vectors take also different values and the form factors $U_L(Q)$ need to be interpolated between the values at Q=$\sqrt{3}$, $\sqrt{8}$, $\sqrt{11}$; in this work a used a cubic spline interpolation and we set $U_L(0)$=0 and $U_L(Q)$=0 for Q>$\sqrt{12}$.

As a validation of the methodology used to include strain effects in our model, Fig. 3 compares our calculations for the energy bandgap of biaxially strained silicon versus the Ge content of the underlying virtual substrate: a good agreement is obtained with experiments[36], as well as with previous calculations[37,39].

III. TRANSPORT FORMALISM BASED ON THE NEGF METHOD

The transport model employed in this paper relies on the Non Equilibrium Green's Function (NEGF) method formulated in a hybrid basis consisting of real space in the transport direction x and plane waves in the (y,z) directions. Such a basis will be hereafter indicated as $(x,\mathbf{K}_{yz})$. In the $(x,\mathbf{K}_{yz})$ basis the pseudopotential $V_L(x_i,(\mathbf{G}_{yz}-\mathbf{G}'_{yz}))$ is periodic of a_0 along x and, in a periodicity interval that with no loss of generality we took as x_i=$0, d, 2d\cdots(a_0-d)$, it can be written

$$V_L(x_i, \mathbf{G}_{yz} - \mathbf{G}'_{yz}) = \frac{2}{N_d} \sum_{(G_x, G'_x)} V_L(G_x - G'_x, \mathbf{G}_{yz} - \mathbf{G}'_{yz}) \exp[i(G_x - G'_x)\, x_i] \quad (13)$$

where $\mathbf{G}$=$(G_x,\mathbf{G}_{yz})$, $\mathbf{G}'$=$(G'_x,\mathbf{G}'_{yz})$ are reciprocal lattice vectors, $V_L(\mathbf{G})$ is given by Eq. (3), and $N_d/2$ is the num-

$$V_L(x_i, \mathbf{G}_{yz} - \mathbf{G}'_{yz}) = \frac{2}{N_d} \sum_{(G_x, G'_x)} V_L(G_x - G'_x, \mathbf{G}_{yz} - \mathbf{G}'_{yz}) \times \exp[i(G_x - G'_x)\, x_i] \quad (14)$$

ber of G_x components in the expansion volume.

A. Block tridiagonal structure of the Hamiltonian matrix

In the hybrid basis the Hamiltonian matrix featuring closed boundary conditions along the transport direction x can be written as[28]

$$[\mathbf{H}_{x\mathbf{K}_{yz}}] = \begin{bmatrix} \mathbf{H}_{1,1} & \mathbf{H}_{0,1} & 0 & 0 & \cdots & 0 \\ \mathbf{H}_{0,1}^{\dagger} & \mathbf{H}_{2,2} & \mathbf{H}_{0,1} & 0 & \cdots & 0 \\ \cdots & & \cdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \mathbf{H}_{0,1}^{\dagger} & \mathbf{H}_{N_b,N_b} \end{bmatrix} \quad (15)$$

where N_b is the number of blocks along the transport direction. The number of x discretization points in the blocks $\mathbf{H}_{l,l}$, $\mathbf{H}_{0,1}$ is set by the non locality in real space of the Hamiltonian operator. Because the quantum confinement operator described in Sec. II A is local, the only non local part of the Hamiltonian stems from the discretization of the kinetic energy operator, whose non locality has been minimized by opting for the second order discretization scheme summarized by Eq. (4). Consequently, the methodology developed in this work allows us to have $\mathbf{H}_{l,l}$, $\mathbf{H}_{0,1}$ blocks corresponding to a single discretization point along the transport direction x.

The disadvantage of a second order compared to a higher-order discretization scheme is that the second order requires a smaller step d in Eq. (4) in order to attain a given discretization accuracy, which in turn results in a larger number of blocks N_b. However, such a drawback is more than compensated by the significant reduction of the block size, which is the most relevant scaling parameter describing the computational burden.

Figure 4 provides a graphical illustration of how the blocks of the Hamiltonian matrix are connected in the block tridiagonal structure of Eq. (15). In particular, Fig. 4 recalls that in our previous formulation[28] each Hamiltonian block consisted of N_d discretization points and had a length equal to the lattice constant $a_0 = N_d\, d$, whereas in the new approach of this work each block consists of a single discretization point and it is only d long.

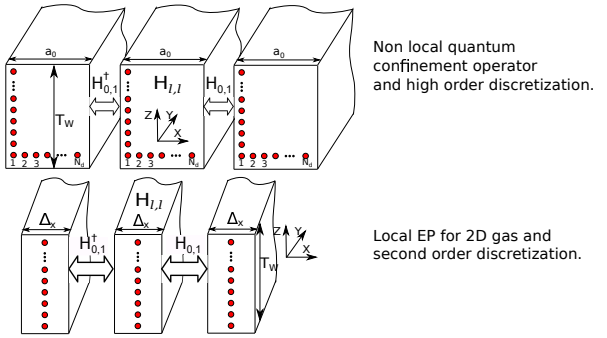


FIG. 4: Pictorial illustration of the size of and coupling between the blocks of the block tridiagonal Hamiltonian matrix. Top: formulation based on a non local (in real space) quantum confinement operator where each block includes N_d discretization points[28]. Bottom: new formulation of this work where each block includes a single discretization point. Reproduced with permission from Ref. 29. Copyright 2018 IEEE.

While the form of the Hamiltonian matrix in Eq. (15) holds for any electron gas dimensionality, the size of the blocks increases when we move from a 3D to a 2D and then to a 1D electron gas [28]. In order to discuss the size of the problem we here denote with N_G the number of $\mathbf{G}$ vectors used for the pseudopotential description, with N_{k_z} the number of k_z in the reduced zone of the underlying semiconductor crystal necessary for the simulation of a 2D electron gas, and with N_{k_y}, N_{k_z} the corresponding numbers for a 1D electron gas. As discussed in details in Ref. 28, the $\mathbf{G}$ vectors belong to a cubic volume expansion with maximum $\mathbf{G}$ components set by the condition

$$|G_s| \leq \frac{N_d}{2}\frac{2\pi}{a_0} \quad s = x, y, z\,, \quad (16)$$

so that N_G is proportional to N_d^3.

Moreover, for a 2D electron gas and a [001] quantization direction, for example, N_{k_z} is equal to $2N_{cz}$, with N_{cz} being the number of unit cells along z in the simulation domain. For a 1D gas and a transport direction along $\langle 100 \rangle$, N_{k_y}, N_{k_z} are similarly given by either the number or twice the number of unit cells in the confinement directions, depending on the shape of the bulk crystal reduced zone used for the calculations. With the above definitions, the size of the Hamiltonian blocks for a 3D electron gas is $M_{3D} = 2N_G/N_d$, for a 2D gas it is $M_{2D} = (2N_G/N_d)N_{k_z}$, and for a 1D gas it is finally $M_{1D} = (2N_G/N_d)N_{k_z}N_{k_y}$, with $2N_G/N_d$ being the number of $\mathbf{G}_{yz}$ vectors in the plane orthogonal to the transport. Hence M_{2D} and M_{1D} increase quadratically with N_d and proportionally to the number of unit cells in the confinement directions.

One last important observation concerning the computational complexity is that, in NEGF based calculations, a further reduction of the size of the block tridiagonal Hamiltonian matrix can be achieved by employing a mode-space transformation[8], and then by keeping only the lowest energy transverse modes, which are the most relevant for transport calculations. The mode-space

Hamiltonian is obtained by means of a unitary transformation for each section of the system along x, namely for a single x discretization point for the methodology of this work (see Fig. 4). The unitary matrix is given by $\mathbf{U}^{(l)} = \left[\xi_1^{(l)} \cdots \xi_{N_{\rm mod}}^{(l)}\right]$, where $\xi_n^{(l)}$ is the eigenvector of the eigenvalue problem

$$\left[\mathbf{H}_{l,l} + \mathbf{H}_{0,1} + \mathbf{H}_{0,1}^{\dagger}\right] \xi_n^{(l)} = E_n^{(l)} \xi_n^{(l)} \,. \qquad (17)$$

We found that for the methodology of this work the mode space approximation works well and helps reduce significantly the size of Hamiltonian block for a 2D and a 1D electron gas. This is not surprising because the off-diagonal blocks of the Hamiltonian $\mathbf{H}_{0,1}$ are in turn diagonal matrices with a constant term t_0 on the diagonal, consequently the transverse modes $\left[\xi_n^{(l)}\right]$ obtained by Eq. (17) are also eigenfunctions of the diagonal blocks $\mathbf{H}_{l,l}$.

We verified that, thanks to the mode space approach, the size of the Hamiltonian blocks can be reduced to $M_{3D} = N_{\rm mod}$ for a 3D gas, and to approximately $M_{2D} = N_{\rm mod} N_{k_z}$ and $M_{1D} = N_{\rm mod} N_{k_z} N_{k_y}$ for respectively a 2D and a 1D gas, where, for the materials and devices analyzed in this paper, an $N_{\rm mod}$ of about 12 is sufficient to have an agreement within a few percent between the mode space results and the results obtained without introducing the mode space reduction.

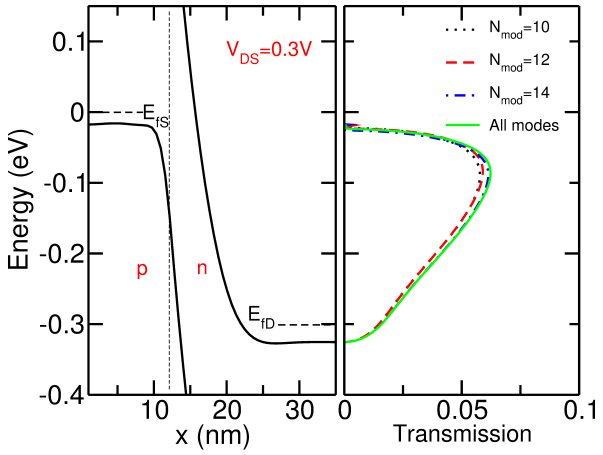


FIG. 5: (Left) Valence and conduction band profile along the transport direction [100] and (right) the corresponding transmission for an InAs Esaki diode with an applied bias of V_{DS}=0.3 V. Doping concentration is $N_A = 5 \times 10^{19}$ cm^{-3} in the p-doped region and $N_D = 10^{19}$ cm^{-3} in the n-doped region. The source Fermi level E_{fS}=0 eV is taken as the energy reference. The transmission is calculated either by using all the 512 transverse modes (green solid line), or by using different numbers of modes corresponding to a substantial mode space reduction.

The rapid convergence of the transmission as a function of $N_{\rm mod}$ is illustrated in Fig. 5 showing the transmission across an InAs Esaki diode under a bias of V_{DS}=0.3 V. As it can be seen, the calculated transmission becomes rapidly independent of $N_{\rm mod}$ for $N_{\rm mod}$ larger than about 12 or 14, which enables a drastic reduction of the Hamiltonian blocks in Eq. (15), that have to be manipulated in the Green's function calculations. The idea to fasten the calculations by exploiting reduced basis sets, as also shown in Ref. 26, seems to be a viable method to enable the use of empirical pseudopotential Hamiltonians in quantum transport problems.

B. Charge, current and self-consistent calculations

Charge and current density can be expressed in terms of the retarded, $[\mathbf{G}_{x\mathbf{K}_{yz}}]$, and lesser-than Green's functions, $[\mathbf{G}^{<}_{x\mathbf{K}_{yz}}]$, which, at a given energy E, are defined as

$$[\mathbf{G}_{x\mathbf{K}_{yz}}(E)] = \left[E\mathbf{I} - [\mathbf{H}_{x\mathbf{K}_{yz}}] - [\mathbf{\Sigma}(E)]\right]^{-1} \qquad (18)$$

and

$$[\mathbf{G}^{<}_{x\mathbf{K}_{yz}}(E)] = [\mathbf{G}_{x\mathbf{K}_{yz}}(E)][\mathbf{\Sigma}^{<}(E)][\mathbf{G}_{x\mathbf{K}_{yz}}(E)]^{\dagger} \qquad (19)$$

where $[\mathbf{\Sigma}]$=$[\mathbf{\Sigma}_L]$+$[\mathbf{\Sigma}_R]$+$[\mathbf{\Sigma}_{\rm ph}]$ and $[\mathbf{\Sigma}^{<}]$=$[\mathbf{\Sigma}^{<}_L]$+$[\mathbf{\Sigma}^{<}_R]$+$[\mathbf{\Sigma}^{<}_{\rm ph}]$ are the retarded and the lesser-than self-energies describing the connection to contacts (i.e. left lead, L, and right lead, R), or possible interaction with photons or phonons[40]. The inclusion of inelastic scattering (not addressed in the present paper) would couple non-linearly Eqs. (18) and (19), therefore requiring to solve them by means of a self-consistent loop.

For transport calculations it is not necessary to directly solve Eqs. (18) and (19), because only the the blocks of the principal diagonal of $[\mathbf{G}_{x\mathbf{K}_{yz}}]$, $[\mathbf{G}^{<}_{x\mathbf{K}_{yz}}(E)]$ are needed to calculate the charge, and only the blocks of the first sub-diagonal are necessary for the current. More precisely, the electron concentration is computed in terms of the real-space, lesser-than Green's function $[\mathbf{G}^{<}_{\mathbf{r}}(E)]$ as

$$n(\mathbf{r}) = \frac{-ig_{\rm s}}{d^3} \int_{E_0(x_i)}^{\infty} \frac{dE}{2\pi} \mathbf{G}^{<}_{\mathbf{r}}(\mathbf{r},\mathbf{r};E) \,, \qquad (20)$$

where $g_{\rm s}$ is the spin degeneracy, $E_0(x_i)$ is the neutrality energy level at the abscissa x_i, here assumed to be in the center of the energy gap. A similar equation holds for hole concentration

$$p(\mathbf{r}) = \frac{ig_{\rm s}}{d^3} \int_{-\infty}^{E_0(x_i)} \frac{dE}{2\pi} \mathbf{G}^{>}_{\mathbf{r}}(\mathbf{r},\mathbf{r};E) \,, \qquad (21)$$

where $[\mathbf{G}^{>}_{\mathbf{r}}(E)]$ is the real-space, greater-than Green's function defined as $[\mathbf{G}^{>}_{\mathbf{r}}(E)]$ = $[\mathbf{G}^{<}_{\mathbf{r}}(E)]$+$[\mathbf{G}_{\mathbf{r}}(E)]$−$[\mathbf{G}_{\mathbf{r}}(E)]^{\dagger}$. The real space Green's functions can be computed from the hybrid basis Green's functions by using a unitary transformation from (K_y,K_z) to (y,z) in each device section[28].

The spatial distribution of the current along the transport direction is expressed in the hybrid basis as

$$I_{x_l \to x_{l+1}} = \frac{g_s e}{\hbar} \int dE \;\; {\rm tr}\left\{\mathbf{H}_{0,1}\mathbf{G}^{<}_{l+1,l} - \mathbf{G}^{<}_{l,l+1}\mathbf{H}^{\dagger}_{0,1}\right\} \qquad (22)$$

where tr$\{\cdots\}$ denotes the trace of a matrix, and $\mathbf{G}^<_{l+1,l}$, $\mathbf{G}^<_{l,l+1}$ are the blocks of $[\mathbf{G}^<_{x\mathbf{K}_{yz}}]$ placed respectively above and below the main diagonal. For the calculation of those blocks we took advantage of the tri-diagonal block shape of the Hamiltonian matrix in Eq. (15), which allowed us to use the recursive Green's functions algorithms based on Dysons's equations[41], which is the ultimate reason why the size of the blocks is a crucial figure for the computational burden.

As for the contact self energies $[\mathbf{\Sigma}_{L,R}]$, these could be obtained by means of one of the approaches already described in the literature such as the Sancho-Rubio iterative scheme[42] or the eigenvalue method[43]. However, these methods compute the surface Green's function of a semi-infinite chain of periodic blocks by taking advantage of the crystal periodicity, and thus they provide the Green's function related to the whole periodic block. In our approach this would be very inefficient because along the transport direction we have a large number N_d of discretization points in the unit cell, even if the coupling terms connect only the first and the last sections of two adjacent unit cells.

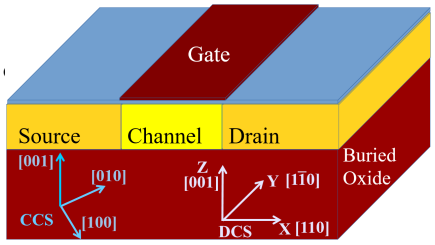
In order to speed up the calculation of the contact self-energy, we developed an entirely new algorithm that takes advantage of the local nature of both the pseudopotential and the confinement operator. More details about the novel method to calculate the contact self-energies are reported in the Appendix.

Finally, self-consistent simulations are obtained by coupling the solution of Eqs. (20-21) with the electrostatic potential arising from the 3D Poisson equation

$$\nabla \cdot [\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})] = -e\,[p(\mathbf{r}) - n(\mathbf{r}) + N_D(\mathbf{r}) - N_A(\mathbf{r})] \tag{23}$$

where $\phi(\mathbf{r})$ is the electrostatic potential, $\epsilon(\mathbf{r})$ is the material-dependent permittivity, and $N_A(\mathbf{r})$, $N_D(\mathbf{r})$ are the acceptor and donor concentration, respectively. When solving Eq. (23), electron and hole concentrations were first computed according to Eqs. (20-21) and using the fine discretization grid d, then, because the electrostatic potential presents slow spatial variations on the scale of the lattice constant, the carrier concentrations were interpolated on a coarser mesh with a discretization step $d_c = a_0/2$.

For all the simulated devices a rapid convergence was achieved with a number of iterations varying approximately between four to eight depending on the specific bias point; no damping of the potential updates was necessary to



(Confinement)/[Transp.]	k_x $(2\pi/a_0)$	k_y $(2\pi/a_0)$	k_z $(2\pi/a_0)$
(001)/[100]	[-1,1[	[-0.5,0.5[	[-1,1[
(001)/[110]	$[-\sqrt{2},\sqrt{2}[$	$[-\frac{0.5}{\sqrt{2}},\frac{0.5}{\sqrt{2}}[$	[−1, 1[

FIG. 6: Device (DCS) and Crystal Coordinate System (CCS) for an ultra-thin body (UTB) FET. The table reports examples of bulk silicon reduced **k** zone for different DCS. (001)/[100] corresponds to DCS≡CCS: x=[100[, y=[010[, z=[001[. (001)/[110] corresponds to x=[110[, y=[1$\bar{1}$0[, z=[001]. Reproduced with permission from Ref. 29. Copyright 2018 IEEE.

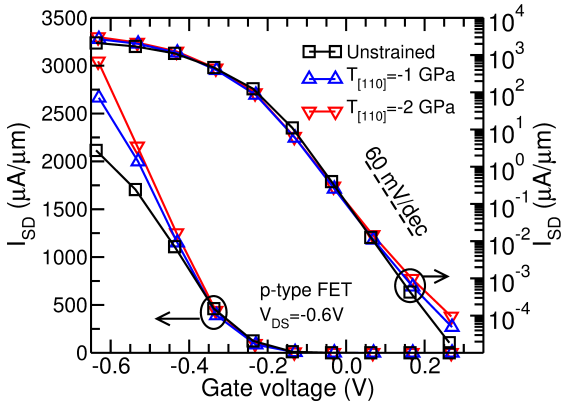


FIG. 7: Simulated I_{DS} versus V_{GS} characteristics at V_{DS}=−0.6 V for a p-type, Si FET with gate length $L_G \simeq 13$ nm and T_{sct}= $7a_0 \simeq 3.8$ nm. Quantization and transport directions are [001] and [110] (see the DCS in Fig. 6). Results for unstrained Si and for compressive uniaxial stress in the [110] transport direction. Gate workfunction is about 4.92 eV for the unstrained FET and it has been adjusted in strained FETs so as to have approximately the same $I_{\rm off}$=0.1μA/μm at V_{GS}=0 V for all devices. Reproduced with permission from Ref. 29. Copyright 2018 IEEE.

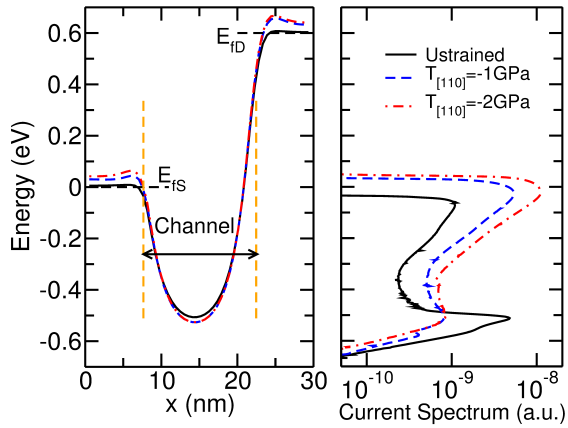


FIG. 8: Valence subband profile (left) and corresponding current density plot (right) for the p-type silicon FET of Fig. 7 and at a bias of V_{GS}=+0.27 V, V_{DS}=−0.6 V, corresponding to the subthreshold region of operation. The source Fermi level E_{fS}=0 eV is taken as the energy reference. Reproduced with permission from Ref. 29. Copyright 2018 IEEE.

IV. SELF-CONSISTENT DEVICE SIMULATIONS

A. Ultra-thin body silicon MOSFET

Figure 6 shows a sketch of the UTB-SOI silicon FET simulated in this section, featuring a body thickness T_w=$7a_0$, namely 3.8 nm in silicon, an equivalent oxide thickness of 0.56 nm and a gate length, L_G, of approximately 13 nm, which is a device structure inspired to ITRS projections for year 2021. For this device the rank of the Hamiltonian blocks of Eq. (15) equals M_{2D}=216. The (001)/[110] orientation corresponds to the DCS shown in Fig. 6, where transport direction is along [110] and confinement direction is along [001]. As already mentioned in Sec. II B, the table in Fig. 6 shows the extension along k_x, k_y, k_z of the bulk crystal reduced zone used for transport calculations. All simulations were run at room temperature, T=300K, if not otherwise stated.

Figure 7 shows the I_{DS} versus V_{GS} curves for an UTB-SOI p-FET for unstrained Si and for different values of compressive uniaxial stress along the channel direction. The large I_{DS} values are due to the fact that neither scattering nor series resistance are included. As can be seen, the stress improves I_{DS} in the on-state at fixed off-current, which is due to a reduction of the effective mass in the [110] transport direction. Strained FETs, however, also have degraded sub-threshold swing, SS, for I_{DS} below approximately $01\mu A/\mu m$. This behavior is explained in Fig. 8, reporting the subbands profile and the current spectral density $J_D(E)$. As it can be seen, the reduction of the transport effective mass implies an increase of source-to-drain tunnelling.

B. InAs nanowire Tunnel-FET

The structure of the square cross-section InAs nanowire Tunnel-FETs addressed in this section is depicted in Fig. 9.

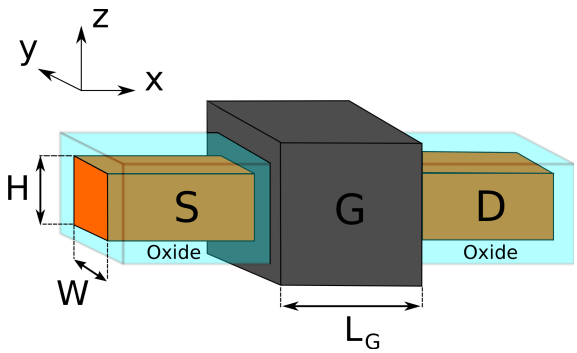


FIG. 9: Sketch of the simulated gate-all-around nanowire FETs, where x is the transport direction and (y, z) the plane of quantum confinement.

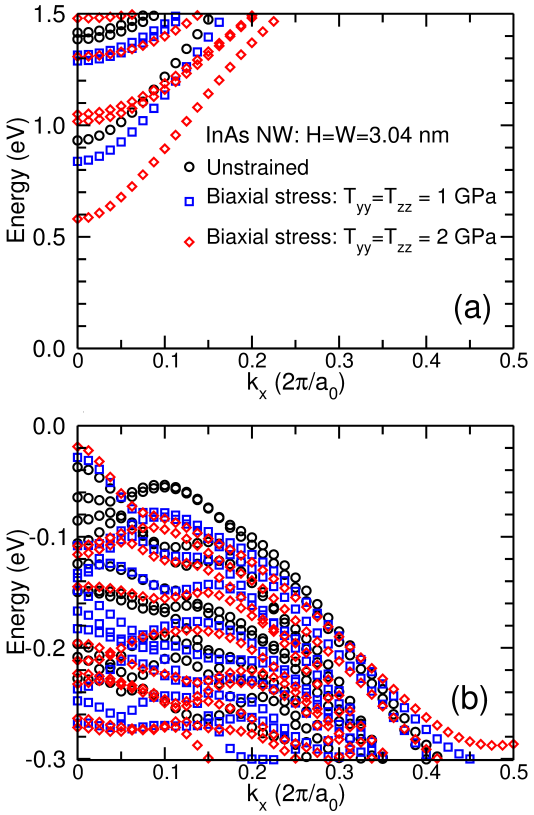


FIG. 10: Band-structure about the Γ point for an InAs nanowire with a semiconductor cross section of H=W=$5a_0$≃=3.04 nm (see Fig. 9): (a) conduction band states; (b) valence band states. The biaxial tensile stress along the plane orthogonal to the transport direction tends to significantly reduce the energy gap. The energy reference is the top of the valence band of unstrained bulk InAs.

Figure 10 illustrates the bandstructure for an InAs nanowire either relaxed or subject to a tensile biaxial stress and having a square cross-section with a 3.04 nm and an equivalent oxide thickness of 0.608 nm. For this device the rank of the Hamiltonian blocks of Eq. (15) equals M_{1D}=1176. As expected, the biaxial strain results in a large reduction of the energy gap[12], that for the unstrained system is approximately 0.97 eV.

Figure 11 illustrates the I_{DS} versus V_{GS} characteristics of the InAs nanowire Tunnel-FET obtained with self-consistent NEGF simulations based on the EP Hamiltonian. The metal gate workfunction was adjusted so as to have approximately the same off current $I_{\rm off}$=I_{DS}[V_{GS}=0]=10pA/μm for all stress conditions. The biaxial tensile stress improves the on state I_{DS} at fixed $I_{\rm off}$, with no sizeable change of the sub-threshold swing in the explored V_{GS} range. Moreover, Fig. 12 shows the subband profiles along the device channel and the current spectra for the simulations in Fig. 11 at V_{GS}≈V_{DS}=0.3 V. As it can be seen the biaxial stress greatly increases the transmission across the channel region and consequently the on current of the Tunnel-FET.

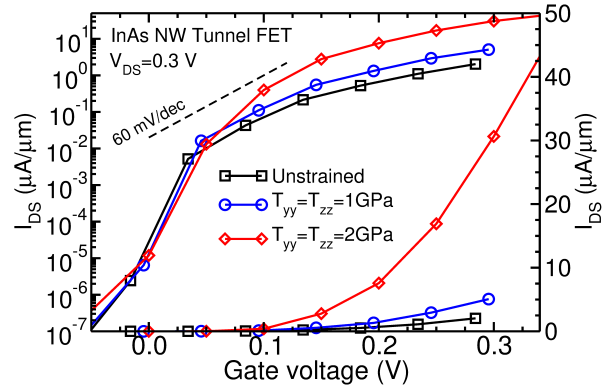


FIG. 11: Simulated drain current versus gate voltage characteristics for an n-type, InAs Tunnel-FET at V_{DS}=0.3 V with gate length $L_G \simeq 17$ nm and $H = T = 5a_0 \simeq 3.04$ nm. Gate workfunction is about 4.384 eV for the unstrained FET and it has been adjusted in strained FETs so as to have approximately the same $I_{\rm off}$=10pA/μm at V_{GS}=0 V for all devices.

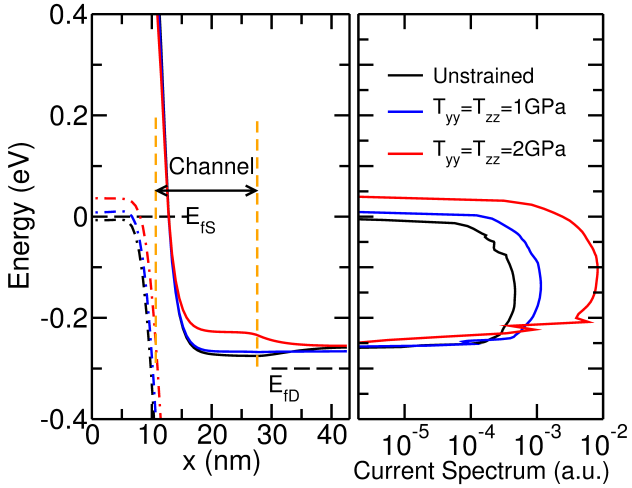


FIG. 12: Conduction (solid lines) and valence (dashed lines) subband profile (left) and corresponding current density plot (right) for the InAs Tunnel-FETs of Fig. 11 at V_{GS}=V_{DS}=0.3 V, and for different stress conditions. The source Fermi level E_{fS}=0 is taken as the energy reference.

V. CONCLUSIONS

In this paper we have presented substantial new developments for full quantum transport simulations obtained with a pseudopotential Hamiltonian and the NEGF method. The proposed approach is fairly general and flexible, in fact it applies to any electron carrier dimensionality, it can be used for arbitrary crystal orientations and it accounts for the effect of a possible strain in the underlying semiconductor crystals.

The numerical efficiency has been greatly improved compared to our previous approach[28], which made it possible to obtain self-consistent simulations even for nanowire MOSFETs and Tunnel-FETs, where the two-dimensional quantum confinement in the plane normal to the transport direction makes the problem the most challenging from the computational burden perspective.

While all the results shown in this paper were obtained by considering coherent transport, our formalism is in principle capable of dealing with incoherent transport and, for example, with electron-phonon interactions. We have not yet addressed incoherent transport, but we see this as an interesting hint for future work.

We conclude by arguing that, besides its application to an empirical pseudopotential Hamiltonian, our approach can be directly applied to plane-waves *ab-initio* Hamiltonian operators utilized in many Density Functional Theory calculations. Our formalism may thus provide an interesting alternative to the methods based on maximally localized Wannier functions, although the theoretical and computational viability of this application remains to be explored, which is left as a stimulating hint for further investigations.

Appendix: Contact self-energy

Here, we present a new algorithm to efficiently compute the contact self-energy of devices described by full-band Hamiltonians in the hybrid basis consisting of either real-space/plane-waves or real-space/transverse modes.

This algorithm exploits the fact that in our formalism the pseudopotential is local in real space and the kinetic energy term in Eq. (4) couples only adjacent slices separated by the discretization step d. Consequently, the Hamiltonian describing the unit cell of length a_0 in Eq. (15) consists of N_d diagonal blocks $\mathbf{H}_{l,l}$ corresponding to a single discretization point along the transport direction, and of the sub-diagonal blocks $\mathbf{H}_{0,1}$ coupling only adjacents slices separated by d. Therefore, only the first and last discretization points of each unit cell are actually connected to the adjacent unit cells, hence we do not need to compute the Green's function (GF) of the entire unit cell of length a_0, which is the outcome of standard algorithms such as the Sancho-Rubio iterative scheme[42], but we need only the blocks corresponding to the first and last discretization point. More precisely, the non-null components of the retarded self-energies for the left (L) and right (R) contacts are defined as

$$\Sigma_L = \mathbf{H}_{0,1}^{\dagger} \mathbf{G}_{N,N} \mathbf{H}_{0,1} \qquad \text{(A.1)}$$

$$\Sigma_R = \mathbf{H}_{0,1} \mathbf{G}_{1,1} \mathbf{H}_{0,1}^{\dagger} \,, \qquad \text{(A.2)}$$

where $\mathbf{G}_{N,N}$ and $\mathbf{G}_{1,1}$ are the surface GFs of the so-called *lead*, defined as a chain of N_T identical unit cells with $N = N_d N_T$ discretization points. In our approach this is much more efficient than using the Sancho-Rubio algorithm, because this would imply to manipulate Hamiltonian blocks describing all the N_d slices of the unit cell, where N_d has to be very large (we used N_d=30) due to the accuracy requirements on the kinetic energy term.

In order to compute the two surface GFs in Eqs. (A.1-A.2), we proceed as follows. The preliminary step, illustrated by the n=0 case in Fig. 13, consists in using the Dyson equation[40] to compute a few GF blocks of the unit cell, namely the $\mathbf{G}^{(0)}_{1,1}$ corresponding to the first discretization point of the single unit cell, the $\mathbf{G}^{(0)}_{N_d,N_d}$ corresponding to the N_d-th discretization point, as well as the $\mathbf{G}^{(0)}_{1,N_d}$ and $\mathbf{G}^{(0)}_{N_d,1}$ linking the first and last point in the single unit cell. Once this is accomplished, the first step consists in computing similar GFs for the system composed by two identical cells and having a constant coupling matrix given by $-\mathbf{H}_{0,1}$ (see the case n=1 in Fig. 13).

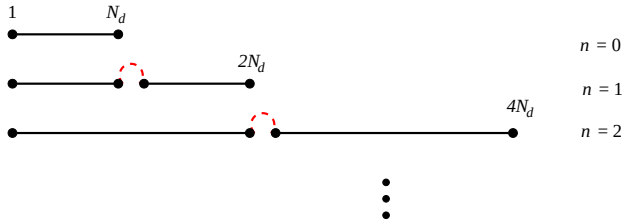


FIG. 13: Scheme of the iterative procedure used to obtain the contact self energy.

Now we can now exploit again the Dyson equations to obtain the GFs for the system consisting of two cells by using the following equations:

$$\mathbf{G}^{(1)}_{1,1} = \mathbf{G}^{(0)}_{1,1} - \mathbf{G}^{(0)}_{1,N_d}\mathbf{H}_{0,1}\mathbf{G}^{(1)}_{N_d+1,1} \tag{A.3}$$

with

$$\mathbf{G}^{(1)}_{N_d+1,1} = -\left(\mathbf{I} - \mathbf{G}^{(0)}_{1,1}\mathbf{H}^{\dagger}_{0,1}\mathbf{G}^{(0)}_{N_d,N_d}\mathbf{H}_{0,1}\right)^{-1} \times \mathbf{G}^{(0)}_{1,1}\mathbf{H}^{\dagger}_{0,1}\mathbf{G}^{(0)}_{N_d,1} \tag{A.4}$$

and

$$\mathbf{G}^{(1)}_{2N_d,1} = -\mathbf{G}^{(0)}_{N_d,1}\mathbf{H}^{\dagger}_{0,1}\mathbf{G}^{(1)}_{N_d,1} \tag{A.5}$$

with

$$\mathbf{G}^{(1)}_{N_d,1} = \mathbf{G}^{(0)}_{N_d,1} - \mathbf{G}^{(0)}_{N_d,N_d}\mathbf{H}_{0,1}\mathbf{G}^{(1)}_{N_d+1,1}\,, \tag{A.6}$$

while similar equations can be used to compute $\mathbf{G}^{(1)}_{2N_d,2N_d}$ and $\mathbf{G}^{(1)}_{1,2N_d}$.

The idea behind this recursive method is to increase the length of the *lead* by iteratively connecting two identical stubs whose GFs were obtained in the previous step. Since the generic n-th step of this iteration scheme consists in connecting two identical sections composed of $2^{n-1}N_d$ slices, it is convenient to use a simplified notation and to define GFs related to the left (L) section and to the right (R) section. By using this notation we have that $\mathbf{G}^{(n)}_{1,1} = \mathbf{G}^{LL}_{1,1}$, $\mathbf{G}^{(n)}_{2^nN_d,2^nN_d} = \mathbf{G}^{RR}_{N,N}$, $\mathbf{G}^{(n)}_{2^nN_d,1} = \mathbf{G}^{RL}_{N,1}$ and $\mathbf{G}^{(n)}_{1,2^nN_d} = \mathbf{G}^{LR}_{1,N}$, whereas the GFs related to the step $n-1$ are written as $\mathbf{G}^{(n-1)}_{1,1} = \mathbf{g}_{1,1}$, $\mathbf{G}^{(n-1)}_{2^{n-1}N_d,2^{n-1}N_d} = \mathbf{g}_{N,N}$, $\mathbf{G}^{(n-1)}_{2^{n-1}N_d,1} = \mathbf{g}_{N,1}$, $\mathbf{G}^{(n-1)}_{1,2^{n-1}N_d} = \mathbf{g}_{1,N}$.

At the step n the equations to be solved for the left section are therefore

$$\mathbf{G}^{RL}_{1,1} = -\left(\mathbf{I} - \mathbf{g}_{1,1}\mathbf{H}^{\dagger}_{0,1}\mathbf{g}_{N,N}\mathbf{H}_{0,1}\right)^{-1} \times \mathbf{g}_{1,1}\mathbf{H}^{\dagger}_{0,1}\mathbf{g}_{N,1} \tag{A.7}$$

$$\mathbf{G}^{LL}_{1,1} = \mathbf{g}_{1,1} - \mathbf{g}_{1,N}\mathbf{H}_{0,1}\mathbf{G}^{RL}_{1,1} \tag{A.8}$$

$$\mathbf{G}^{LL}_{N,1} = \mathbf{g}_{N,1} - \mathbf{g}_{N,N}\mathbf{H}_{0,1}\mathbf{G}^{RL}_{1,1} \tag{A.9}$$

$$\mathbf{G}^{RL}_{N,1} = -\mathbf{g}_{N,1}\mathbf{H}^{\dagger}_{0,1}\mathbf{G}^{LL}_{N,1}\,, \tag{A.10}$$

while, for the right section they are

$$\mathbf{G}^{LR}_{N,N} = -\left(\mathbf{I} - \mathbf{g}_{N,N}\mathbf{H}_{0,1}\mathbf{g}_{1,1}\mathbf{H}^{\dagger}_{0,1}\right)^{-1} \times \mathbf{g}_{N,N}\mathbf{H}_{0,1}\mathbf{g}_{1,N} \tag{A.11}$$

$$\mathbf{G}^{RR}_{N,N} = \mathbf{g}_{N,N} - \mathbf{g}_{N,1}\mathbf{H}^{\dagger}_{0,1}\mathbf{G}^{LR}_{N,N} \tag{A.12}$$

$$\mathbf{G}^{RR}_{1,N} = \mathbf{g}_{1,N} - \mathbf{g}_{1,1}\mathbf{H}^{\dagger}_{0,1}\mathbf{G}^{LR}_{N,N} \tag{A.13}$$

$$\mathbf{G}^{LR}_{1,N} = -\mathbf{g}_{1,N}\mathbf{H}_{0,1}\mathbf{G}^{RR}_{1,N}\,. \tag{A.14}$$

The convergence is reached once that the largest element of $\left(\left|\mathbf{G}^{LL}_{1,1} - \mathbf{g}_{1,1}\right| + \left|\mathbf{G}^{RR}_{N,N} - \mathbf{g}_{N,N}\right|\right)$ is smaller than a predefined tolerance, making possible the identification of $\mathbf{G}^{RR}_{N,N}$ and $\mathbf{G}^{LL}_{1,1}$ with the surface GF in Eqs. (A.1) and (A.2). A satisfactory convergence can be obtained in a few steps (typically 10 to 15) thanks to the fact that the length of the *lead* increases as $2^{n+1}a_0$.

[1] "The International Technology Roadmap for Semiconductors (ITRS)," (2015).
[2] R. Kim, U. E. Avci, and I. A. Young, IEEE Trans. on Electron Devices **62**, 713 (2015).
[3] D. Esseni, M. Pala, and T. Rollo, IEEE Trans. on Electron Devices **62**, 3084-3091 (2015).
[4] C. Grillet, D. Logoteta, A. Cresti, and M. G. Pala, IEEE Trans. on Electron Devices **64**, 2425 (2017).
[5] A. Seabaugh and Q. Zhang, Proceedings of the IEEE **98**, 2095 (2010).
[6] D. Esseni, M. Pala, P. Palestri, C. Alper, and T. Rollo, Semiconductor Science Technology **32**, 083005 (2017).
[7] S.De Franceschi, L. Hutin, R. Maurand, L. Bourdet, H. Bohuslavskyi, A. Corna, D. Kotekar-Patil, S. Barraud, X. Jehl, M. S. Y.-M. Niquet, and M. Vinet, in IEEE International Electron Devices Meeting , 339 (2016).
[8] R. Venugopal, Z. Ren, S. Datta, and M. Lundstrom, J. Appl. Phys. **92**, 3730 (2002).
[9] J.Wang, E.Polizzi, and M. Lundstrom, J. Appl. Phys. **96**, 2192 (2004).
[10] S. Poli, M. G. Pala, T. Poiroux, S. Deleonibus, and G. Baccarani, IEEE Transactions on Electron Devices **55**, 2968 (2008).
[11] M. Shin, Journal of Applied Physics **106**, 054505 (2009).
[12] F. Conzatti, M. Pala, D. Esseni, E. Bano, and L. Selmi, Electron Devices, IEEE Transactions on **59**, 2085 (2012).
[13] E. Baravelli, E. Gnani, R. Grassi, A. Gnudi, S. Reggiani, and G. Baccarani, Electron Devices, IEEE Transactions on **61**, 178 (2014).
[14] M. Luisier, A. Schenk, and W. Fichtner, Phys. Rev. B **74**, 205323 (2006).
[15] G. Klimeck, S.S. Ahmed, H. Bae, N. Kharche, R. Rahman, S. Clark, B. Haley, S. Lee, M. Naumov, H. Ryu, F. Saied, M. Prada, M. Korkusinski, and T.B. Boykin, IEEE Trans. on Electron Devices **54**, 2079 (2007).

[16]L.-W. Wang, A. Franceschetti, and A. Zunger, Phys. Rev. Lett. **78**, 2819 (1997).
[17]L.-W. Wang and A. Zunger, Phys. Rev. B **59**, 15806 (1999).
[18]F. Chirico, A. Di Carlo, and P. Lugli, Phys. Rev. B **64**, 045314 (2001).
[19]D. Esseni and P. Palestri, Phys. Rev. B **72**, 165342.1 (2005).
[20]J.-L. van der Steen, D. Esseni, P. Palestri, L. Selmi, and R.J.E. Hueting, IEEE Trans. on Electron Devices **54**, 1843 (2007).
[21]H. J. Choi and J. Ihm, Phys. Rev. B **59**, 2267 (1999).
[22]Xiang-Wei Jiang, Shu-Shen Li, Jian-Bai Xia, and Lin-Wang Wang, Journal of Applied Physics **109**, 054503 (2011).
[23]A. Garcia-Lekue, M. Vergniory, X. Jiang, and L. Wang, Progress in Surface Science **90**, 292 (2015).
[24]J. Fang, W. G. Vandenberghe, B. Fu, and M. V. Fischetti, Journal of Applied Physics **119**, 035701 (2016).
[25]J. Fang, S. Chen, W. G. Vandenberghe, B. Fu, and M. V. Fischetti, Electron Devices, IEEE Transactions on **64**, 2758 (2017).
[26]M. L. Van de Put, M. V. Fischetti, and W. G. Vandenberghe, arXiv preprint arXiv:1903.00548 (2019).
[27]M.Pala, O.Badami, and D. Esseni, in IEEE International Electron Devices Meeting , 35.1.1 (2017).
[28]M. G. Pala and D. Esseni, Phys. Rev. B **97**, 125310 (2018).
[29]M.Pala, O.Badami, and D. Esseni, in IEEE International Electron Devices Meeting , 33.1.1 (2018).
[30]J. R. Chelikowsky and M. L. Cohen, Phys. Rev. B **10**, 5095 (1974).
[31]M. L. Cohen and J. R. Chelikowsky, "Electron structure and optical properties of semiconductors," (Springer Series in Solid-State Sciences. Springer-Verlag Berlin Heidelberg New York London Tokyo, 1988).
[32]D. Esseni, P. Palestri, and L. Selmi, *"Nanoscale MOS Transistors - Semi-Classical Transport and Applications"*, 1st ed. (Cambridge University Press., 2011).
[33]T. B. Boykin, G. Klimeck, and F. Oyafuso, Phys. Rev. B **69**, 115201 (2004).
[34]M.V. Fischetti and S.E. Laux, Journal of Applied Physics **80**, 2234 (1996).
[35]Q. M. Ma, K. L. Wang, and J. N. Schulman, Phys. Rev. B **47**, 1936 (1993).
[36]J. Munguía, G. Bremond, J. M. Bluet, J. M. Hartmann, and M. Mermoux, Applied Physics Letters **93** (2008).
[37]S. E. Ungersboeck, *Advanced modeling of strained CMOS technology*, Ph.D. thesis, Technischen Universität Wien, Wien, Austria (2007).
[38]S. M. Sze and K. K. Ng, *"Physics of Semiconductor Devices"* (John Wiley & Sons, 2006).
[39]E. Ungersboeck, S. Dhar, G. Karlowatz, V. Sverdlov, H. Kosina, and S. Selberherr, IEEE Trans. on Electron Devices , 2183 (2007).
[40]G.D. Mahan, "Many-Particle Physics," (Plenum Press, New York, 1988).
[41]M. P. Anantram, M. S. Lundstrom, and D. E. Nikonov, Proceedings of the IEEE **96**, 1511 (2008).
[42]M. P. L. Sancho, J. M. L. Sancho, and J. Rubio, Journal of Physics F: Metal Physics **14**, 1205 (1984).
[43]D. H. Lee and J. D. Joannopoulos, Phys. Rev. B **23**, 4988 (1981).