Implementation on a high-performance computing platform of a deterministic solver for Double-Gate MOSFETs

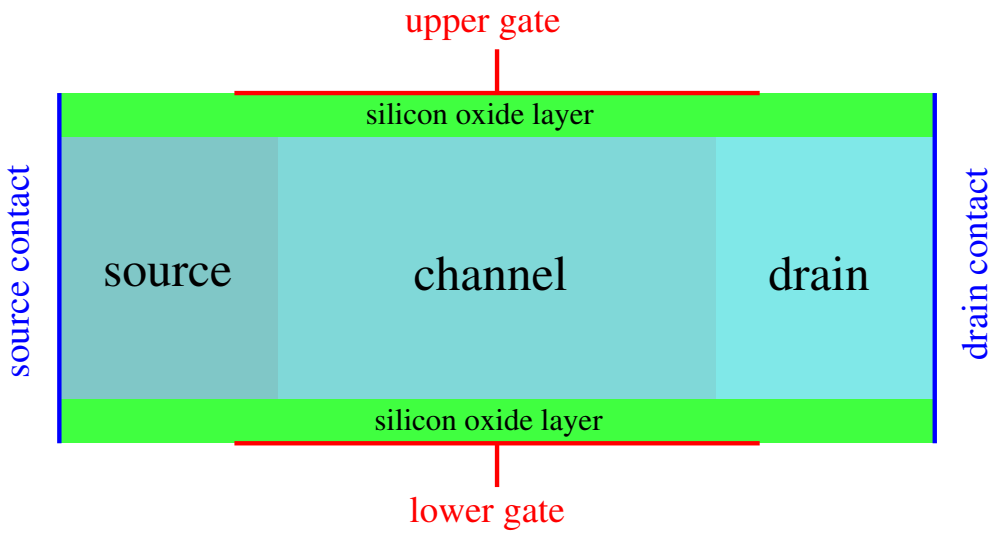
Francesco Vecil, José Miguel Mantas

Santiago de Compostela, ECMI 2016

Outline

Outline

Geometry



The gates

The metallic gates act as a “tap”: they open or close the channel.

The electrons’ confinement

Electrons are confined inside a 3.15 V-deep well along the transversal dimension.

The doping

The $p^+ - p - p^+$ doping attemps to control the conductivity of the device.

Outline

Dimensional coupling

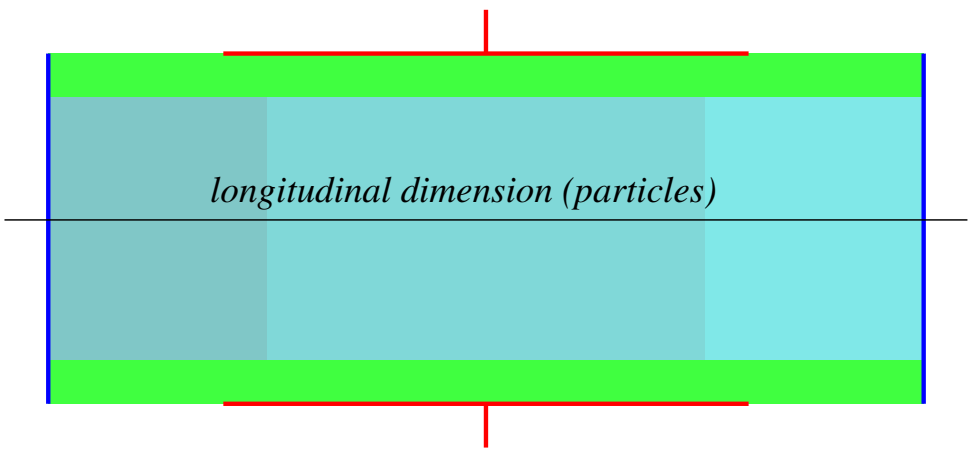
Dimensional coupling

Electrons are **waves** along the transversal dimension, **particles** otherwise.

Description of the longitudinal dimension

One Boltzmann Transport Equation for each pair $(\nu, p) \in \{1, 2, 3\} \times \{1, \ldots, N_{\text{sbn}}\}$

$$\frac{\partial f_{\nu,p}}{\partial t} + v_{x,\nu} \frac{\partial f_{\nu,p}}{\partial x} - \frac{1}{\hbar} \frac{\partial \epsilon_{\nu,p}}{\partial x} \frac{\partial f_{\nu,p}}{\partial k_x} = \mathcal{Q}_{\nu,p}[f]$$



Dimensional coupling

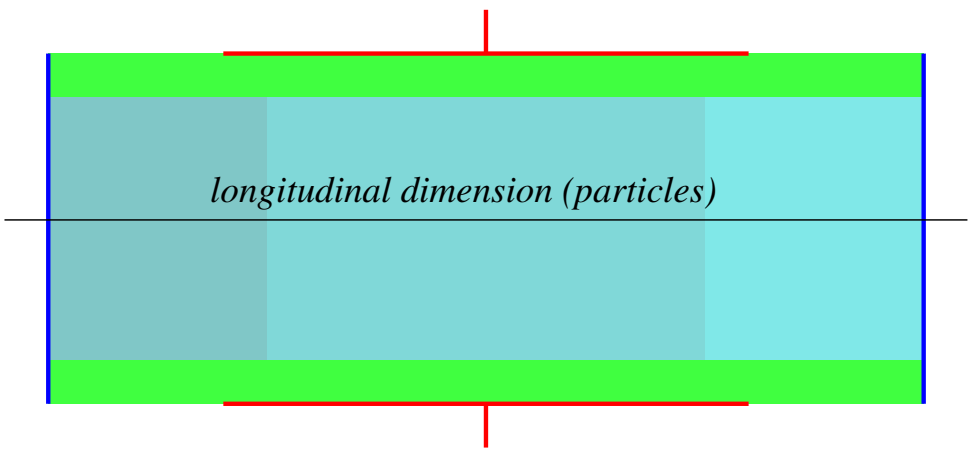
Dimensional coupling

Electrons are **waves** along the transversal dimension, **particles** otherwise.

Description of the longitudinal dimension

One Boltzmann Transport Equation for each pair $(\nu, p) \in \{1, 2, 3\} \times \{1, \ldots, N_{\text{sbn}}\}$

$$\frac{\partial f_{\nu,p}}{\partial t} + v_{x,\nu} \frac{\partial f_{\nu,p}}{\partial x} - \frac{1}{\hbar} \frac{\partial \epsilon_{\nu,p}}{\partial x} \frac{\partial f_{\nu,p}}{\partial k_x} = \mathcal{Q}_{\nu,p}[\boldsymbol{f}]$$

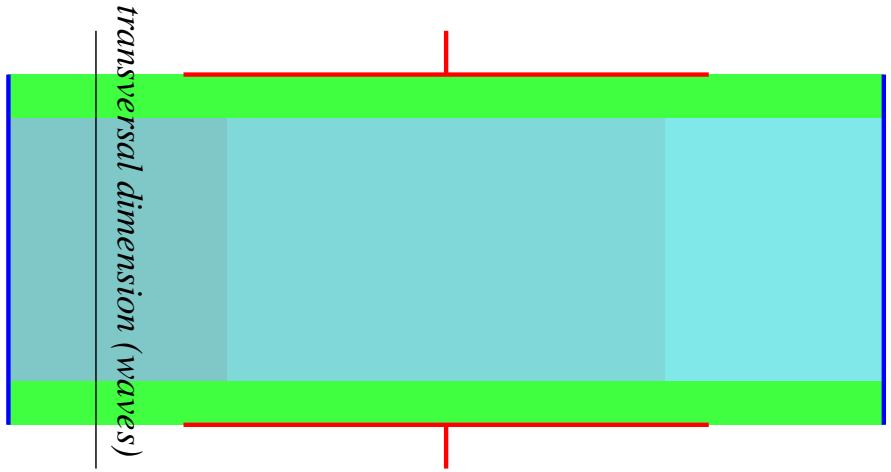


Dimensional coupling

Description of the transversal dimension

A set of 1D Schrödinger eigenvalue problems:

$$-\frac{\hbar^2}{2}\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{1}{m_{z,\nu}}\frac{\mathrm{d}\psi_{\nu,p}}{\mathrm{d}z}\right]-q\left(V+V_c\right)\psi_{\nu,p}=\epsilon_{\nu,p}\,\psi_{\nu,p}.$$

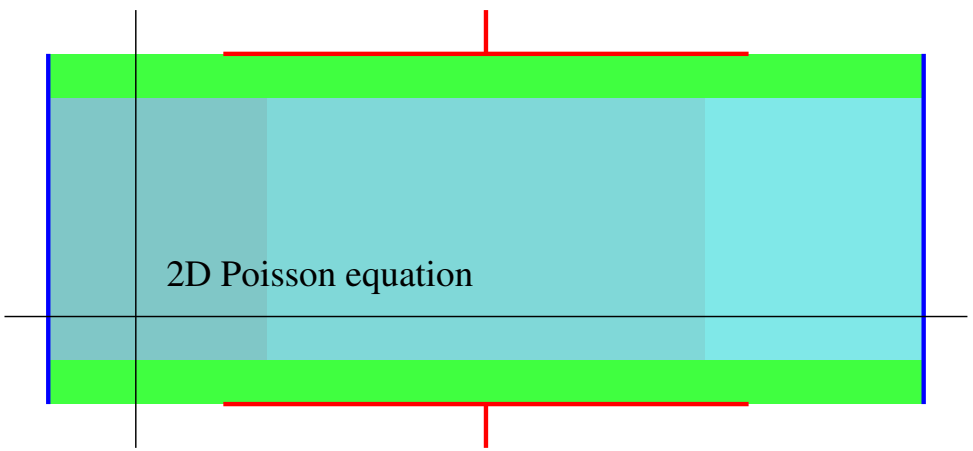


Dimensional coupling

Electrostatic field

The 2D Poisson equation couples the "classical" and the "quantum" dimensions:

$$\mathrm{div}_{x,z}\left[\varepsilon_R \nabla_{x,z} V\right] = \frac{q}{\varepsilon_0}\left(2\sum_{\nu=1}^{3}\sum_{p=1}^{N_{\mathrm{sbn}}}\int_{\mathbb{R}^2} f_{\nu,p}\,\mathrm{d}\boldsymbol{k}\;|\psi_{\nu,p}|^2 - N_D\right).$$



Outline

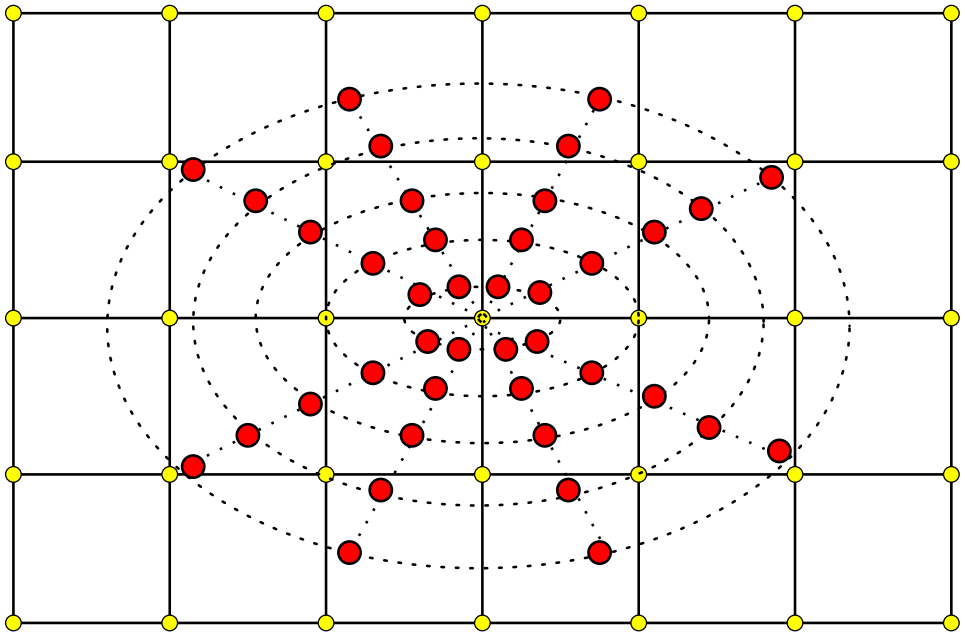
Transport part

BTE's

After changing variables from Cartesian to ellipsoidal for $\boldsymbol{k}$, the set of BTE's rewrite

$$\frac{\partial \Phi_{\nu,p}}{\partial t} + \frac{\partial}{\partial x}\left[a^1_{\nu}\,\Phi_{\nu,p}\right] + \frac{\partial}{\partial w}\left[a^2_{\nu,p}\,\Phi_{\nu,p}\right] + \frac{\partial}{\partial \phi}\left[a^3_{\nu,p}\,\Phi_{\nu,p}\right] = \mathcal{Q}_{\nu,p}[\boldsymbol{\Phi}]\,s_{\nu}(w),$$

with $a^2_{\nu,p}$ and $a^3_{\nu,p}$ representing the electrostatic force.



Transport part

Time discretization

Third-order Total Variation Diminishing (TVD) Runge-Kutta scheme:

$$\begin{aligned}
\boldsymbol{\Phi}^1 &= \boldsymbol{\Phi}^n + \Delta t\, \boldsymbol{H}(\boldsymbol{\Phi}^n), \\
\boldsymbol{\Phi}^2 &= \frac{3}{4}\,\boldsymbol{\Phi}^n + \frac{\boldsymbol{\Phi}^1}{4} + \frac{\Delta t}{4}\boldsymbol{H}(\boldsymbol{\Phi}^1), \\
\boldsymbol{\Phi}^{n+1} &= \frac{\boldsymbol{\Phi}^n}{3} + \frac{2}{3}\,\boldsymbol{\Phi}^2 + \frac{2}{3}\Delta t\, \boldsymbol{H}(\boldsymbol{\Phi}^2)
\end{aligned}$$

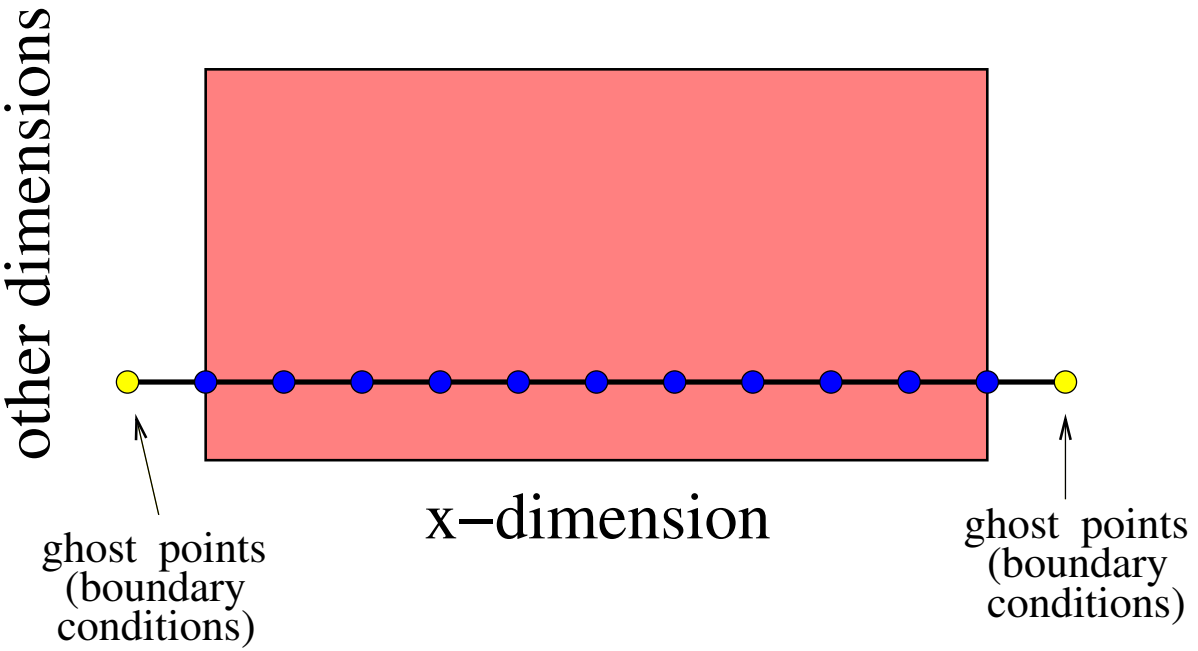
where $\boldsymbol{H}$ are the right-hand sides of the BTE's:

$$H_{\nu,p}(\boldsymbol{\Phi}) := -\frac{\partial}{\partial x}\left[a^1_\nu\, \Phi_{\nu,p}\right] - \frac{\partial}{\partial w}\left[a^2_{\nu,p}\, \Phi_{\nu,p}\right] - \frac{\partial}{\partial \phi}\left[a^3_{\nu,p}\, \Phi_{\nu,p}\right] + \mathcal{Q}_{\nu,p}[\boldsymbol{\Phi}]\, s_\nu(w).$$

Transport part

Partial derivatives

The three partial derivatives $\frac{\partial}{\partial x}\left[a^1_\nu\,\Phi_{\nu,p}\right]$, $\frac{\partial}{\partial w}\left[a^2_{\nu,p}\,\Phi_{\nu,p}\right]$ and $\frac{\partial}{\partial \phi}\left[a^3_{\nu,p}\,\Phi_{\nu,p}\right]$ are approximated *line by line* using a WENO routine for finite differences.



Transport part

Scattering operator

Explicit formulae for the scattering operator are (neglecting many constants):

$$\mathcal{Q}_{\nu,p}[\Phi]\, s_\nu(w) = s_\nu(w) \sum_{\nu',p'} \frac{\int \left[\mathbb{I}_{\Gamma_-}\, \Phi_{\nu',p'}(\Gamma_-,\phi') + \mathbb{I}_{\Gamma_+}\, \Phi_{\nu',p'}(\Gamma_+,\phi')\right] \mathrm{d}\phi'}{W_{(\nu',p')\leftrightarrow(\nu,p)}}$$

$$- 2\pi\, \Phi_{\nu,p}(w,\phi) \sum_{\nu',p'} \frac{\mathbb{I}_{\Gamma_+}\, s_{\nu'}\,(\Gamma_+) + \mathbb{I}_{\Gamma_-}\, s_{\nu'}\,(\Gamma_-)}{W_{(\nu,p)\leftrightarrow(\nu',p')}}$$

with W being the *overlap integral*

$$\frac{1}{W_{(\nu,p)\leftrightarrow(\nu',p')}} = \int |\psi_{\nu,p}|^2 |\psi_{\nu',p'}|^2 \,\mathrm{d}z.$$

The confined dimension

Schrödinger-Poisson block

$$-\frac{\hbar^2}{2}\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{1}{m_{z,\nu}}\frac{\mathrm{d}\psi_{\nu,p}}{\mathrm{d}z}\right] - q\left(V + V_c\right)\psi_{\nu,p} = \epsilon_{\nu,p}\,\psi_{\nu,p}$$

$$\mathrm{div}_{x,z}\left[\varepsilon_R \nabla_{x,z} V\right] = \frac{q}{\varepsilon_0}\left(2\sum_{\nu=1}^{3}\sum_{p=1}^{N_{\mathrm{sbn}}} \varrho_{\nu,p}\ |\psi_{\nu,p}|^2 - N_D\right).$$

Input/output

surf. densities $\varrho(x)$ $\longrightarrow$ subbands $\epsilon(x)$, wave functions $\psi(x,z)$, potential $V(x,z)$.

The confined dimension

Schrödinger-Poisson block

$$-\frac{\hbar^2}{2}\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{1}{m_{z,\nu}}\frac{\mathrm{d}\psi_{\nu,p}}{\mathrm{d}z}\right] - q\left(V + V_c\right)\psi_{\nu,p} = \epsilon_{\nu,p}\,\psi_{\nu,p}$$

$$\mathrm{div}_{x,z}\left[\varepsilon_R \nabla_{x,z} V\right] = \frac{q}{\varepsilon_0}\left(2\sum_{\nu=1}^{3}\sum_{p=1}^{N_{\mathrm{sbn}}} \varrho_{\nu,p}\ |\psi_{\nu,p}|^2 - N_D\right).$$

Input/output

surf. densities $\boldsymbol{\varrho}(x)$ $\longrightarrow$ subbands $\boldsymbol{\epsilon}(x)$, wave functions $\boldsymbol{\psi}(x,z)$, potential $V(x,z)$.

The confined dimension

Reformulating the problem

The Schrödinger-Poisson problem (here written extremely simplified) is solved by a Newton-Raphson iterative scheme seeking for the zero of

$$P[V] = -\Delta V + \sum_{\nu,p} \varrho_{\nu,p} \, |\psi_{\nu,p}|^2 \qquad \text{under} \qquad \left(\frac{\mathrm{d}^2}{\mathrm{d}z^2} - V \right) \psi_{\nu,p} = \epsilon_{\nu,p} \, \psi_{\nu,p}.$$

The iterative scheme

We are led to a Poisson-like equation to update the guess for the potential V

$$-\Delta V^{\text{new}} + \int \mathcal{A}(z,\zeta) \, V^{\text{new}}(\zeta) \mathrm{d}\zeta = -\sum_{\nu,p} \varrho_{\nu,p} \left| \psi^{\text{old}}_{\nu,p} \right|^2 + \int \mathcal{A}(z,\zeta) \, V^{\text{old}}(\zeta) \mathrm{d}\zeta.$$

The confined dimension

Reformulating the problem

The Schrödinger-Poisson problem (here written extremely simplified) is solved by a Newton-Raphson iterative scheme seeking for the zero of

$$P[V] = -\Delta V + \sum_{\nu,p} \varrho_{\nu,p} \, |\psi_{\nu,p}|^2 \qquad \text{under} \qquad \left(\frac{\mathrm{d}^2}{\mathrm{d}z^2} - V \right) \psi_{\nu,p} = \epsilon_{\nu,p} \, \psi_{\nu,p}.$$

The iterative scheme

We are led to a Poisson-like equation to update the guess for the potential V

$$-\Delta V^{\text{new}} + \int \mathcal{A}(z,\zeta) \, V^{\text{new}}(\zeta) \mathrm{d}\zeta = -\sum_{\nu,p} \varrho_{\nu,p} \left| \psi^{\text{old}}_{\nu,p} \right|^2 + \int \mathcal{A}(z,\zeta) \, V^{\text{old}}(\zeta) \mathrm{d}\zeta.$$

The confined dimension

The Newton-Raphson kernels

$$\mathcal{A}(z,\zeta) = 2 \sum_{\nu,p} \sum_{p' \neq p} \frac{\varrho_{\nu,p} - \varrho_{\nu,p'}}{\epsilon^{\text{old}}_{\nu,p'} - \epsilon^{\text{old}}_{\nu,p}} \psi^{\text{old}}_{\nu,p}(\zeta)\, \psi^{\text{old}}_{\nu,p'}(\zeta)\, \psi^{\text{old}}_{\nu,p'}(z)\, \psi^{\text{old}}_{\nu,p}(z).$$

Solver for the Schrödinger equation

Each of the $3 \times N_x$ eigenvalue problems $\left(\frac{\mathrm{d}^2}{\mathrm{d}z^2} - V \right) \psi_{\nu,p} = \epsilon_{\nu,p}\, \psi_{\nu,p}$ is independently solved by LAPACK's **DSTEQR**.

Solver for the Poisson-like equation

The linear system on V^{new}

$$-\Delta V^{\text{new}} + \int \mathcal{A}(z,\zeta)\, V^{\text{new}}(\zeta)\mathrm{d}\zeta = -\sum_{\nu,p} \varrho_{\nu,p} \left| \psi^{\text{old}}_{\nu,p} \right|^2 + \int \mathcal{A}(z,\zeta)\, V^{\text{old}}(\zeta)\mathrm{d}\zeta.$$

is preconditioned using **ILUC**, then solved by the **IDR** method.

The confined dimension

The Newton-Raphson kernels

$$\mathcal{A}(z,\zeta) = 2 \sum_{\nu,p} \sum_{p' \neq p} \frac{\varrho_{\nu,p} - \varrho_{\nu,p'}}{\epsilon^{\text{old}}_{\nu,p'} - \epsilon^{\text{old}}_{\nu,p}} \psi^{\text{old}}_{\nu,p}(\zeta)\, \psi^{\text{old}}_{\nu,p'}(\zeta)\, \psi^{\text{old}}_{\nu,p'}(z)\, \psi^{\text{old}}_{\nu,p}(z).$$

Solver for the Schrödinger equation

Each of the $3 \times N_x$ eigenvalue problems $\left(\frac{\mathrm{d}^2}{\mathrm{d}z^2} - V \right) \psi_{\nu,p} = \epsilon_{\nu,p}\, \psi_{\nu,p}$ is independently solved by LAPACK's **DSTEQR**.

Solver for the Poisson-like equation

The linear system on V^{new}

$$-\Delta V^{\text{new}} + \int \mathcal{A}(z,\zeta)\, V^{\text{new}}(\zeta)\mathrm{d}\zeta = -\sum_{\nu,p} \varrho_{\nu,p} \left| \psi^{\text{old}}_{\nu,p} \right|^2 + \int \mathcal{A}(z,\zeta)\, V^{\text{old}}(\zeta)\mathrm{d}\zeta.$$

is preconditioned using **ILUC**, then solved by the **IDR** method.

The confined dimension

The Newton-Raphson kernels

$$\mathcal{A}(z,\zeta) = 2 \sum_{\nu,p} \sum_{p' \neq p} \frac{\varrho_{\nu,p} - \varrho_{\nu,p'}}{\epsilon^{\text{old}}_{\nu,p'} - \epsilon^{\text{old}}_{\nu,p}} \psi^{\text{old}}_{\nu,p}(\zeta)\, \psi^{\text{old}}_{\nu,p'}(\zeta)\, \psi^{\text{old}}_{\nu,p'}(z)\, \psi^{\text{old}}_{\nu,p}(z).$$

Solver for the Schrödinger equation

Each of the $3 \times N_x$ eigenvalue problems $\left(\frac{\mathrm{d}^2}{\mathrm{d}z^2} - V \right) \psi_{\nu,p} = \epsilon_{\nu,p}\, \psi_{\nu,p}$ is independently solved by LAPACK's **DSTEQR**.

Solver for the Poisson-like equation

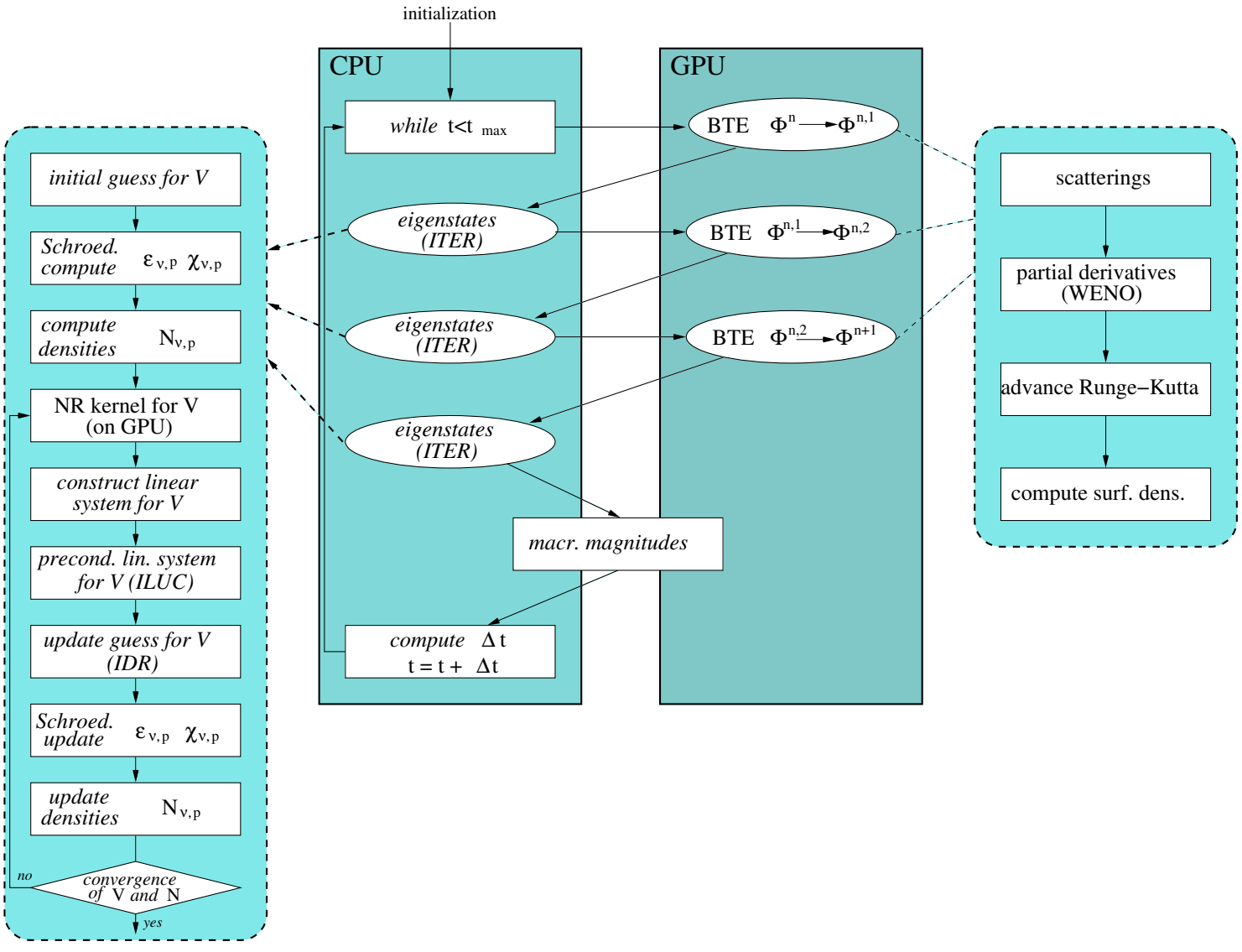
The linear system on V^{new}

$$-\Delta V^{\text{new}} + \int \mathcal{A}(z,\zeta)\, V^{\text{new}}(\zeta)\mathrm{d}\zeta = -\sum_{\nu,p} \varrho_{\nu,p} \left| \psi^{\text{old}}_{\nu,p} \right|^2 + \int \mathcal{A}(z,\zeta)\, V^{\text{old}}(\zeta)\mathrm{d}\zeta.$$

is preconditioned using **ILUC**, then solved by the **IDR** method.

Outline

Overall design of the solver



Main computational phases

The **scatterings** phase (on the GPU)

$$\mathcal{Q}_{\nu,p}[\Phi]\, s_\nu(w) = s_\nu(w) \sum_{\nu',p'} \frac{\int \left[\mathbb{I}_{\Gamma_-}\, \Phi_{\nu',p'}(\Gamma_-,\phi') + \mathbb{I}_{\Gamma_+}\, \Phi_{\nu',p'}(\Gamma_+,\phi')\right] \mathrm{d}\phi'}{W_{(\nu',p')\leftrightarrow(\nu,p)}}$$
$$- 2\pi\, \Phi_{\nu,p}(w,\phi) \sum_{\nu',p'} \frac{\mathbb{I}_{\Gamma_+}\, s_{\nu'}\,(\Gamma_+) + \mathbb{I}_{\Gamma_-}\, s_{\nu'}\,(\Gamma_-)}{W_{(\nu,p)\leftrightarrow(\nu',p')}}.$$

- $W_{(\nu,p)\leftrightarrow(\nu',p')}$ is precomputed.
- The operator requires $7 \times 3 \times N_{\text{sbn}} \times N_x \times N_w \times N_\phi$ independent ϕ-integrations on total ($\approx 59,000,000$ for the "standard" meshes).
- Loop on the scattering phenomenon, and $3 \times N_{\text{sbn}} \times N_x \times N_w \times N_\phi$ threads created at each time.

The **WENO** phase (GPU)

The computation of $\dfrac{\partial}{\partial x}\left[a^1_\nu\, \Phi_{\nu,p}\right]$ is performed on the GPU, using $3 \times N_{\text{sbn}} \times N_w \times N_\phi$ threads (no N_x). Similarly for the other dimensions.

Main computational phases

The **scatterings** phase (on the GPU)

$$\mathcal{Q}_{\nu,p}[\Phi]\, s_\nu(w) = s_\nu(w) \sum_{\nu',p'} \frac{\int \left[\mathbb{I}_{\Gamma_-}\, \Phi_{\nu',p'}(\Gamma_-,\phi') + \mathbb{I}_{\Gamma_+}\, \Phi_{\nu',p'}(\Gamma_+,\phi')\right] \mathrm{d}\phi'}{W_{(\nu',p')\leftrightarrow(\nu,p)}}$$
$$- 2\pi\, \Phi_{\nu,p}(w,\phi) \sum_{\nu',p'} \frac{\mathbb{I}_{\Gamma_+}\, s_{\nu'}\,(\Gamma_+) + \mathbb{I}_{\Gamma_-}\, s_{\nu'}\,(\Gamma_-)}{W_{(\nu,p)\leftrightarrow(\nu',p')}}.$$

- $W_{(\nu,p)\leftrightarrow(\nu',p')}$ is precomputed.
- The operator requires $7 \times 3 \times N_{\text{sbn}} \times N_x \times N_w \times N_\phi$ independent ϕ-integrations on total ($\approx 59,000,000$ for the "standard" meshes).
- Loop on the scattering phenomenon, and $3 \times N_{\text{sbn}} \times N_x \times N_w \times N_\phi$ threads created at each time.

The **WENO** phase (GPU)

The computation of $\dfrac{\partial}{\partial x}\left[a^1_\nu\, \Phi_{\nu,p}\right]$ is performed on the GPU, using $3 \times N_{\text{sbn}} \times N_w \times N_\phi$ threads (no N_x). Similarly for the other dimensions.

Main computational phases

The **update V** phase (hybrid)

- The Newton-Raphson kernel is computed on the GPU as (x-dep. omitted)

$$\mathcal{A}(z_j, z_{j'}) = 2 \sum_{\nu,p} \psi^{\text{old}}_{\nu,p}(z_{j'})\, \psi^{\text{old}}_{\nu,p}(z_j) \sum_{p' \neq p} \underbrace{\frac{\varrho_{\nu,p} - \varrho_{\nu,p'}}{\epsilon^{\text{old}}_{\nu,p'} - \epsilon^{\text{old}}_{\nu,p}}}_{\text{precomputed}} \psi^{\text{old}}_{\nu,p'}(z_{j'})\, \psi^{\text{old}}_{\nu,p'}(z_j).$$

  using shared memory, so that threads in the same block share the same x.

- The linear system is constructed, preconditioned and solved on the CPU

$$-\Delta V^{\text{new}} + \int \mathcal{A}(z,\zeta)\, V^{\text{new}}(\zeta)\mathrm{d}\zeta = -\sum_{\nu,p} \varrho_{\nu,p} \left|\psi^{\text{old}}_{\nu,p}\right|^2 + \int \mathcal{A}(z,\zeta)\, V^{\text{old}}(\zeta)\mathrm{d}\zeta.$$

The **Schrödinger** phase (CPU via OpenMP)

We parallelize via OpenMP the diagonalization of the Schrödinger matrices

$$-\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{1}{m_{z,\nu}}\frac{\mathrm{d}\psi_{\nu,p}}{\mathrm{d}z}\right) - V\psi_{\nu,p} = \epsilon_{\nu,p}\,\psi_{\nu,p}.$$

Main computational phases

The **update V** phase (hybrid)

- The Newton-Raphson kernel is computed on the GPU as (x-dep. omitted)

$$\mathcal{A}(z_j, z_{j'}) = 2 \sum_{\nu,p} \psi^{\text{old}}_{\nu,p}(z_{j'})\, \psi^{\text{old}}_{\nu,p}(z_j) \sum_{p' \neq p} \underbrace{\frac{\varrho_{\nu,p} - \varrho_{\nu,p'}}{\epsilon^{\text{old}}_{\nu,p'} - \epsilon^{\text{old}}_{\nu,p}}}_{\text{precomputed}} \psi^{\text{old}}_{\nu,p'}(z_{j'})\, \psi^{\text{old}}_{\nu,p'}(z_j).$$

  using shared memory, so that threads in the same block share the same x.

- The linear system is constructed, preconditioned and solved on the CPU

$$-\Delta V^{\text{new}} + \int \mathcal{A}(z,\zeta)\, V^{\text{new}}(\zeta) \mathrm{d}\zeta = -\sum_{\nu,p} \varrho_{\nu,p} \left|\psi^{\text{old}}_{\nu,p}\right|^2 + \int \mathcal{A}(z,\zeta)\, V^{\text{old}}(\zeta) \mathrm{d}\zeta.$$

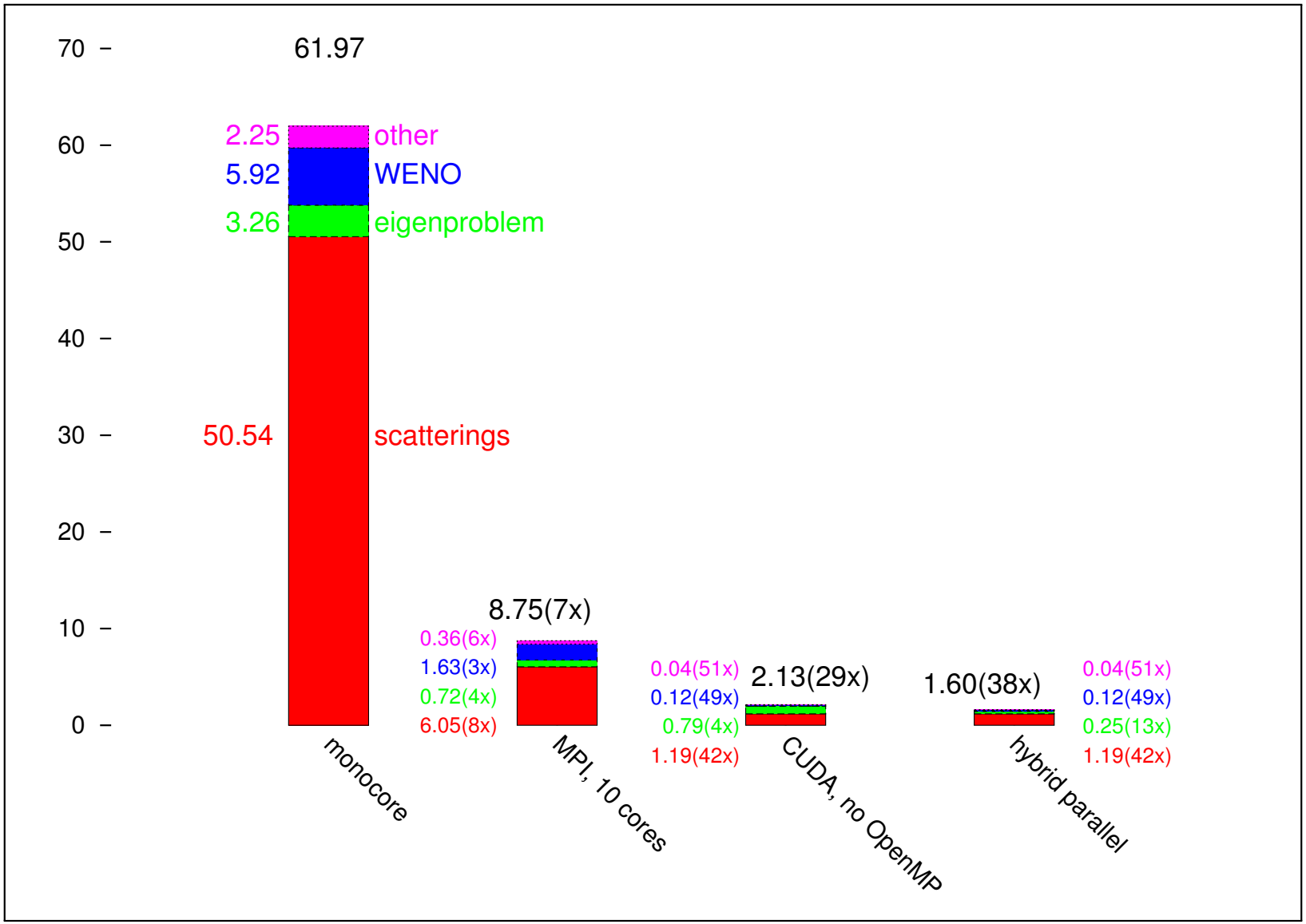
The **Schrödinger** phase (CPU via OpenMP)

We parallelize via OpenMP the diagonalization of the Schrödinger matrices

$$-\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{1}{m_{z,\nu}}\frac{\mathrm{d}\psi_{\nu,p}}{\mathrm{d}z}\right) - V\psi_{\nu,p} = \epsilon_{\nu,p}\,\psi_{\nu,p}.$$

Outline

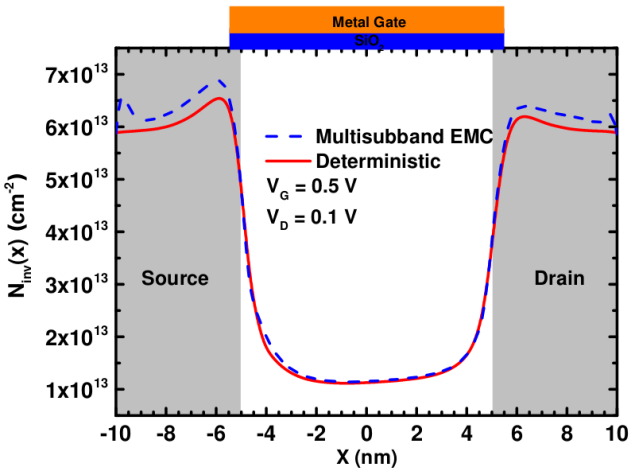
Speedup obtained



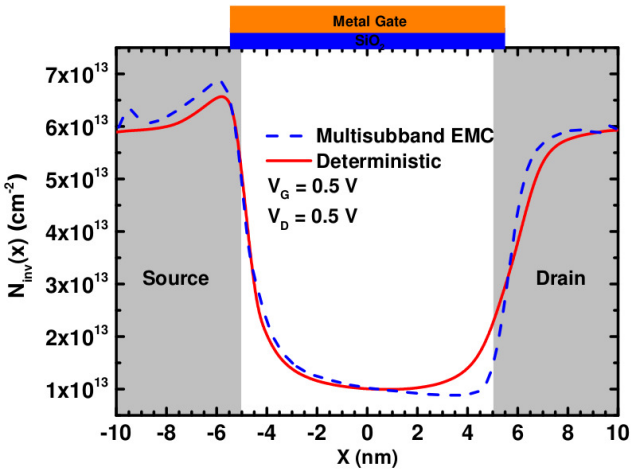
Performances OpenMP/CUDA

Outline

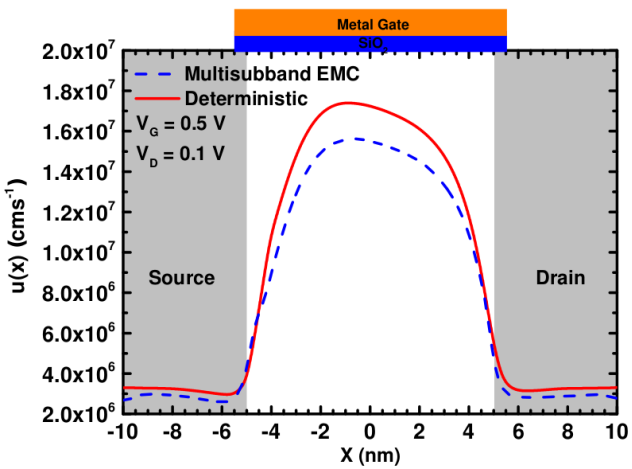
Comparison to Monte-Carlo



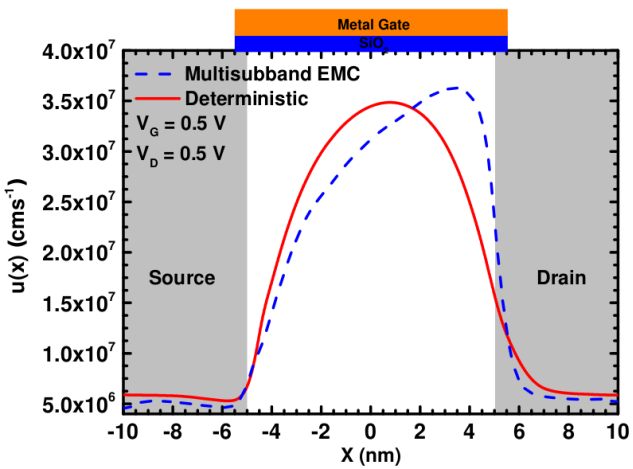
(a) For a $V_{\mathrm{D}} = 0.1$ V bias



(b) For a $V_{\mathrm{D}} = 0.5$ V bias

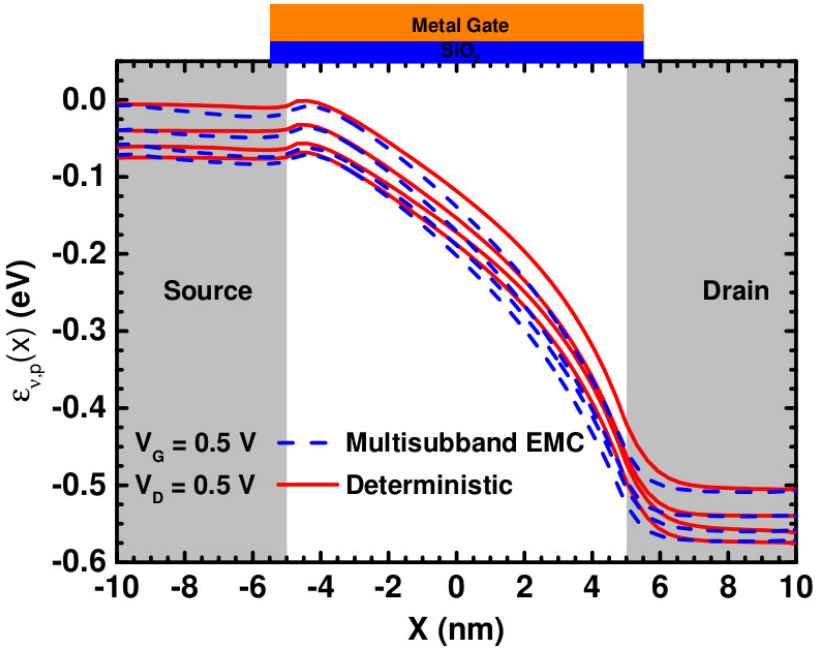


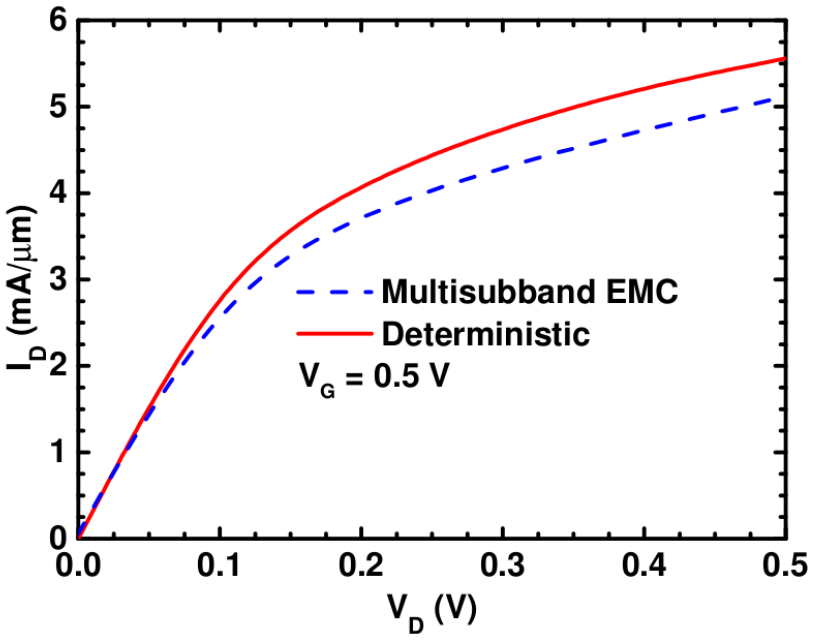
(c) For a $V_{\mathrm{D}} = 0.1$ V bias



(d) For a $V_{\mathrm{D}} = 0.5$ V bias

Comparison to Monte-Carlo





¡GRACIAS!

Se agradecen los proyectos **MTM2011-27739-C04-02** y **MTM2014-52056-P** financiados por el Ministerio de Economía y Competitividad, y el Fondo Europeo de Desarrollo Regional.

francesco.vecil@gmail.com