Hybrid implementation of a deterministic solver for DG-MOSFETs

Francesco Vecil (Laboratoire de Mathématiques Blaise Pascal, Université Clermont Auvergne),

José Miguel Mantas (Dpto. de Lenguajes y Sistemas Informáticos, ETS Ingeniería Informática y Telecomunicaciones, Universidad de Granada)

Valencia, 2019/07/18

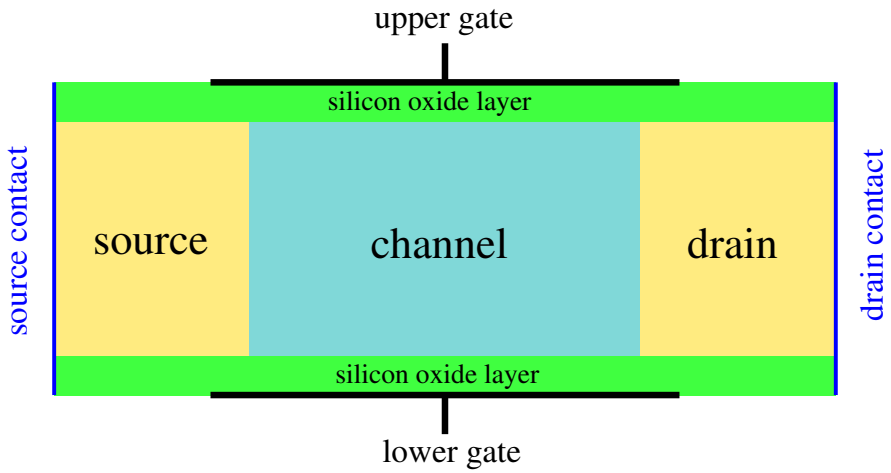
Outline

Publications

- N. Ben Abdallah, M.J. Cáceres, J.A. Carrillo, F. Vecil **(2009)** *A deterministic solver for a hybrid quantum-classical transport model in nanoMOSFETs*, Journal of Computational Physics 228 (17) 6553–6571.
- F. Vecil, J.M. Mantas, M.J. Cáceres, C. Sampedro, A. Godoy, F. Gámiz **(2014)** *A parallel deterministic solver for the Schrödinger–Poisson–Boltzmann system in ultra-short DG-MOSFETs: Comparison with Monte-Carlo*, Computers and Mathematics with Applications 67 (9) 1703–1721.
- F. Vecil, J.M. Mantas *Hybrid openMP-CUDA parallel implementation of a deterministic solver for ultra-short DG-MOSFETs*, International Journal of High Performance Computing Applications, awaiting final decision.

Outline

Geometry



About the scaling

In 1971, the Intel 4004 processor had 1000 transistors, whose channel length was 10000 nm. In 2003 the Intel Pentium IV was 130 nm long. As of 2019, Intel's i3-8121U x86 CPU uses 10-nm long CMOS finFET technology. Though, it seems less efficient than i3-8130U x86 CPU based on 14-nm technology.

Why is it important?

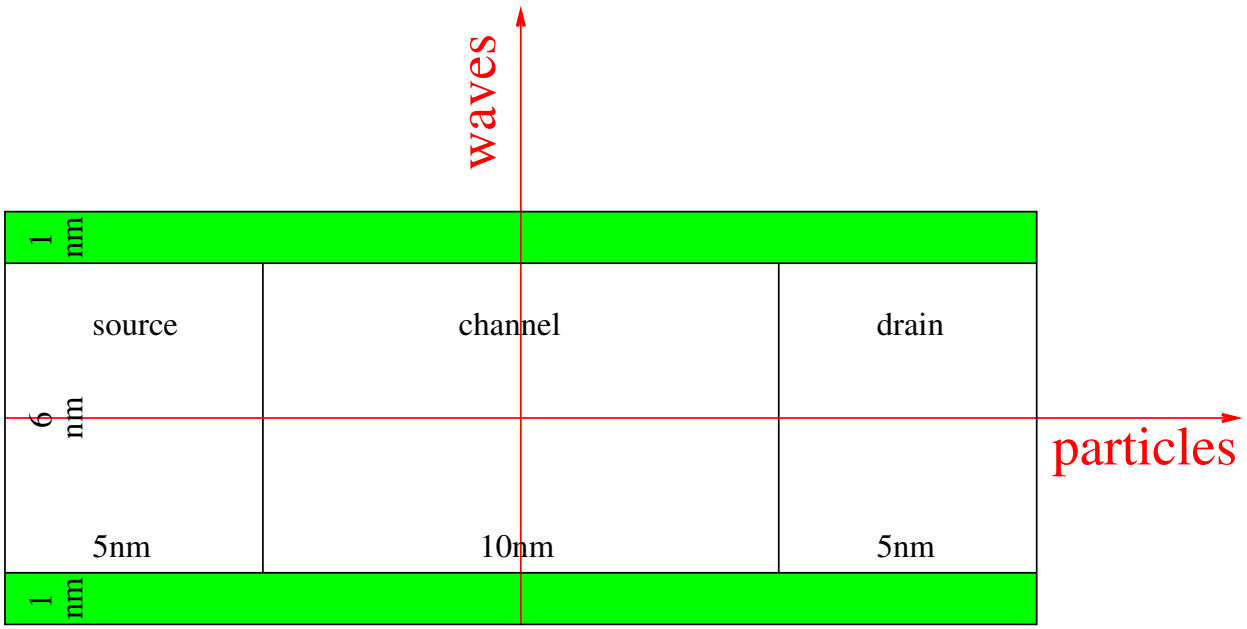
Smaller MOSFETs allow for the construction of smaller devices with better performances; moreover, they allow silicon and energy saving, due to the lower voltages needed to switch on or off the transistor.

Outline

The confinement

Dimensional coupling

Electrons are **particles** along the x-dimension (longitudinal, transport), **waves** along the z-dimension (transversal, quantum).



The confinement

Description of the confinement

A set of 1D steady-state Schrödinger eigenproblems describe z-dimension for each valley $\nu \in \{1, 2, 3\}$ (valleys are physical properties of the semiconductor).

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

input: V
output: $\{\epsilon_{\nu,p}, \psi_{\nu,p}\}_{p\geq 1}$

Effects of the confinement

The confinement produces a discretization of the electrons' energy levels, and a split of the electron population.

Quantized magnitudes

The eigenvalues $\{\epsilon_{\nu,p}(x)\}_{p\geq 1}$ represent the *energy levels*.
The eigenfunctions $\{\psi_{\nu,p}(x,z)\}_{p\geq 1}$ are called *wave functions* in physics.

The confinement

Description of the confinement

A set of 1D steady-state Schrödinger eigenproblems describe z-dimension for each valley $\nu \in \{1, 2, 3\}$ (valleys are physical properties of the semiconductor).

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

input: V
output: $\{\epsilon_{\nu,p}, \psi_{\nu,p}\}_{p\geq 1}$

Effects of the confinement

The confinement produces a discretization of the electrons' energy levels, and a split of the electron population.

Quantized magnitudes

The eigenvalues $\{\epsilon_{\nu,p}(x)\}_{p\geq 1}$ represent the *energy levels*.
The eigenfunctions $\{\psi_{\nu,p}(x,z)\}_{p\geq 1}$ are called *wave functions* in physics.

The confinement

Description of the confinement

A set of 1D steady-state Schrödinger eigenproblems describe z-dimension for each valley $\nu \in \{1, 2, 3\}$ (valleys are physical properties of the semiconductor).

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

input: V
output: $\{\epsilon_{\nu,p}, \psi_{\nu,p}\}_{p\geq 1}$

Effects of the confinement

The confinement produces a discretization of the electrons' energy levels, and a split of the electron population.

Quantized magnitudes

The eigenvalues $\{\epsilon_{\nu,p}(x)\}_{p\geq 1}$ represent the *energy levels*.
The eigenfunctions $\{\psi_{\nu,p}(x,z)\}_{p\geq 1}$ are called *wave functions* in physics.

The unconfined dimension

Classical transport along x

$$\left\{ \frac{\partial f_{\nu,p}}{\partial t} + \overbrace{\frac{1}{\hbar}\frac{\partial \epsilon_\nu^{\text{kin}}}{\partial k_x}\frac{\partial f_{\nu,p}}{\partial x}}^{\text{free motion}} - \overbrace{\frac{1}{\hbar}\frac{\partial \epsilon_{\nu,p}}{\partial x}\frac{\partial f_{\nu,p}}{\partial k_x}}^{\text{force field}} = \overbrace{\mathcal{Q}_{\nu,p}[f]}^{\text{scatterings}} \right\}_{(\nu,p)}$$

Boltzmann Transport Equations

The electron-phonon interactions

Each of the seven electron-phonon scattering mechanisms has structure:

$$\mathcal{Q}_{\nu,p}[f](x,\boldsymbol{k}) = \sum_{\nu',p'} \int_{\mathbb{R}^2} \left[S_{(\nu',p',\boldsymbol{k}')\to(\nu,p,\boldsymbol{k})} f_{\nu',p'}(\boldsymbol{k}') - S_{(\nu,p,\boldsymbol{k})\to(\nu',p',\boldsymbol{k}')} f_{\nu,p}(\boldsymbol{k}) \right] \mathrm{d}\boldsymbol{k}',$$

where $S_{(\nu,p,\boldsymbol{k})\to(\nu',p',\boldsymbol{k}')} = C_{\nu\to\nu'} \frac{1}{W_{(\nu,p)\leftrightarrow(\nu',p')}} \delta\left(\epsilon^{\text{tot}}_{\nu',p'}(\boldsymbol{k}') - \epsilon^{\text{tot}}_{\nu,p}(\boldsymbol{k}) \pm \text{energy}\right)$

and $\frac{1}{W_{(\nu,p)\leftrightarrow(\nu',p')}} = \int_0^{L_z} |\psi_{\nu,p}|^2 |\psi_{\nu',p'}|^2 \,\mathrm{d}z$ is called *overlap integral*.

The unconfined dimension

Classical transport along x

$$\left\{ \frac{\partial f_{\nu,p}}{\partial t} + \overbrace{\frac{1}{\hbar}\frac{\partial \epsilon_\nu^{\text{kin}}}{\partial k_x}\frac{\partial f_{\nu,p}}{\partial x}}^{\text{free motion}} - \overbrace{\frac{1}{\hbar}\frac{\partial \epsilon_{\nu,p}}{\partial x}\frac{\partial f_{\nu,p}}{\partial k_x}}^{\text{force field}} = \overbrace{\mathcal{Q}_{\nu,p}[f]}^{\text{scatterings}} \right\}_{(\nu,p)}$$

Boltzmann Transport Equations

The electron-phonon interactions

Each of the seven electron-phonon scattering mechanisms has structure:

$$\mathcal{Q}_{\nu,p}[f](x,\boldsymbol{k}) = \sum_{\nu',p'} \int_{\mathbb{R}^2} \left[S_{(\nu',p',\boldsymbol{k}')\to(\nu,p,\boldsymbol{k})} f_{\nu',p'}(\boldsymbol{k}') - S_{(\nu,p,\boldsymbol{k})\to(\nu',p',\boldsymbol{k}')} f_{\nu,p}(\boldsymbol{k}) \right] \mathrm{d}\boldsymbol{k}',$$

where $S_{(\nu,p,\boldsymbol{k})\to(\nu',p',\boldsymbol{k}')} = C_{\nu\to\nu'} \frac{1}{W_{(\nu,p)\leftrightarrow(\nu',p')}} \delta\left(\epsilon^{\text{tot}}_{\nu',p'}(\boldsymbol{k}') - \epsilon^{\text{tot}}_{\nu,p}(\boldsymbol{k}) \pm \text{energy}\right)$

and $\frac{1}{W_{(\nu,p)\leftrightarrow(\nu',p')}} = \int_0^{L_z} |\psi_{\nu,p}|^2 |\psi_{\nu',p'}|^2 \,\mathrm{d}z$ is called *overlap integral*.

The model

BTE

The Boltzmann Transport Equation (one for each pair (ν, p)) reads

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

Schrödinger-Poisson block

$$-\frac{\hbar^2}{2}\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{1}{m_{z,\nu}}\frac{\mathrm{d}\psi_{\nu,p}}{\mathrm{d}z}\right] - q\left(V + V_c\right)\psi_{\nu,p} = \epsilon_{\nu,p}\psi_{\nu,p} \qquad \begin{array}{l}\text{input: } V\\ \text{output: } \{\epsilon_{\nu,p}, \psi_{\nu,p}\}_{p\geq 1}\end{array}$$

$$-\operatorname{div}_{x,z}\left[\varepsilon_R \nabla_{x,z} V\right] = -\frac{q}{\varepsilon_0}\left(2\sum_{\nu,p}\varrho_{\nu,p}[f]\,|\psi_{\nu,p}|^2 - N_D\right) \qquad \begin{array}{l}\text{input: } \psi_{\nu,p}\\ \text{output: } V\end{array}$$

These equations cannot be decoupled because we need the eigenfunctions to compute the potential, and we need the potential to compute the eigenfunctions.
Seen as a block: input: $\varrho_{\nu,p}[f] \longrightarrow$ output: $\{\epsilon_{\nu,p}, \psi_{\nu,p}\}_{(\nu,p)}$.

The model

BTE

The Boltzmann Transport Equation (one for each pair (ν, p)) reads

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

Schrödinger-Poisson block

$$-\frac{\hbar^2}{2}\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{1}{m_{z,\nu}}\frac{\mathrm{d}\psi_{\nu,p}}{\mathrm{d}z}\right] - q\left(V + V_c\right)\psi_{\nu,p} = \epsilon_{\nu,p}\psi_{\nu,p} \qquad \begin{array}{l}\text{input: } V \\ \text{output: } \{\epsilon_{\nu,p}, \psi_{\nu,p}\}_{p\geq 1}\end{array}$$

$$-\operatorname{div}_{x,z}\left[\varepsilon_R \nabla_{x,z} V\right] = -\frac{q}{\varepsilon_0}\left(2\sum_{\nu,p}\varrho_{\nu,p}[f]\,|\psi_{\nu,p}|^2 - N_D\right) \qquad \begin{array}{l}\text{input: } \psi_{\nu,p} \\ \text{output: } V\end{array}$$

These equations cannot be decoupled because we need the eigenfunctions to compute the potential, and we need the potential to compute the eigenfunctions.
Seen as a block: input: $\varrho_{\nu,p}[f] \longrightarrow$ output: $\{\epsilon_{\nu,p}, \psi_{\nu,p}\}_{(\nu,p)}$.

Outline

Time integration

Adimensionalization

After complete adimensionalization, and in particular by a change of variables into ellipsoidal variables for k, we obtain the pdf in conservation-law form

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

Runge-Kutta

We advance in time by the third order Total Variation Diminishing Runge-Kutta scheme (no explicit time-dependency):

1. $\Phi^{n,1}_{\nu,p} = \Delta t H_{\nu,p}(\Phi^n)$
2. $\Phi^{n,2}_{\nu,p} = \frac{3}{4}\Phi^n_{\nu,p} + \frac{1}{4}\Phi^{n,1}_{\nu,p} + \frac{1}{4}\Delta t H_{\nu,p}(\Phi^{n,1})$
3. $\Phi^{n+1} = \frac{1}{3}\Phi^n_{\nu,p} + \frac{2}{3}\Phi^{n,2}_{\nu,p} + \frac{2}{3}H_{\nu,p}(\Phi^{n,2})$

Time integration

Adimensionalization

After complete adimensionalization, and in particular by a change of variables into ellipsoidal variables for k, we obtain the pdf in conservation-law form

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

Runge-Kutta

We advance in time by the third order Total Variation Diminishing Runge-Kutta scheme (no explicit time-dependency):

1. $\Phi^{n,1}_{\nu,p} = \Delta t H_{\nu,p}(\Phi^n)$
2. $\Phi^{n,2}_{\nu,p} = \frac{3}{4}\Phi^n_{\nu,p} + \frac{1}{4}\Phi^{n,1}_{\nu,p} + \frac{1}{4}\Delta t H_{\nu,p}(\Phi^{n,1})$
3. $\Phi^{n+1} = \frac{1}{3}\Phi^n_{\nu,p} + \frac{2}{3}\Phi^{n,2}_{\nu,p} + \frac{2}{3}H_{\nu,p}(\Phi^{n,2})$

Outline

Transport part

Partial derivatives

The three partial derivatives are approximated by means of WENO methods.

Inelastic phenomena

$$
\begin{aligned}
&\mathcal{Q}_{\nu,p}[\Phi]s_\nu(w) \\
=&C^{\mathcal{Q}}s_\nu(w)\sum_{\nu',p'}\frac{\gamma_{\nu'\to\nu}N_{\nu'\to\nu}}{W_{(\nu',p')\leftrightarrow(\nu,p)}}\mathbb{I}_{\{\Gamma_-\geq 0\}}\int_{\phi'=0}^{2\pi}\Phi_{\nu',p'}\left(\Gamma'_-,\phi'\right)\,\mathrm{d}\phi' \\
+&C^{\mathcal{Q}}s_\nu(w)\sum_{\nu',p'}\frac{\gamma_{\nu'\to\nu}(N_{\nu'\to\nu}+1)}{W_{(\nu',p')\leftrightarrow(\nu,p)}}\mathbb{I}_{\{\Gamma_+\geq 0\}}\int_{\phi'=0}^{2\pi}\Phi_{\nu',p'}\left(\Gamma_+,\phi'\right)\,\mathrm{d}\phi' \\
-&C^{\mathcal{Q}}2\pi\Phi_{\nu,p}(w,\phi)\sum_{\nu',p'}\frac{\gamma_{\nu\to\nu'}N_{\nu\to\nu'}}{W_{(\nu,p)\leftrightarrow(\nu',p')}}\mathbb{I}_{\{\Gamma_+\geq 0\}}s_{\nu'}\left(\Gamma_+\right) \\
-&C^{\mathcal{Q}}2\pi\Phi_{\nu,p}(w,\phi)\sum_{\nu',p'}\frac{\gamma_{\nu\to\nu'}(N_{\nu\to\nu'}+1)}{W_{(\nu,p)\leftrightarrow(\nu',p')}}\mathbb{I}_{\{\Gamma_-\geq 0\}}s_{\nu'}\left(\Gamma_-\right)
\end{aligned}
$$

Transport part

Partial derivatives

The three partial derivatives are approximated by means of WENO methods.

Inelastic phenomena

$$
\begin{aligned}
&\mathcal{Q}_{\nu,p}[\Phi]s_\nu(w) \\
=&C^{\mathcal{Q}}s_\nu(w)\sum_{\nu',p'}\frac{\gamma_{\nu'\to\nu}N_{\nu'\to\nu}}{W_{(\nu',p')\leftrightarrow(\nu,p)}}\mathbb{I}_{\{\Gamma_-\geq 0\}}\int_{\phi'=0}^{2\pi}\Phi_{\nu',p'}\left(\Gamma'_-,\phi'\right)\mathrm{d}\phi' \\
+&C^{\mathcal{Q}}s_\nu(w)\sum_{\nu',p'}\frac{\gamma_{\nu'\to\nu}\left(N_{\nu'\to\nu}+1\right)}{W_{(\nu',p')\leftrightarrow(\nu,p)}}\mathbb{I}_{\{\Gamma_+\geq 0\}}\int_{\phi'=0}^{2\pi}\Phi_{\nu',p'}\left(\Gamma_+,\phi'\right)\mathrm{d}\phi' \\
-&C^{\mathcal{Q}}2\pi\Phi_{\nu,p}(w,\phi)\sum_{\nu',p'}\frac{\gamma_{\nu\to\nu'}N_{\nu\to\nu'}}{W_{(\nu,p)\leftrightarrow(\nu',p')}}\mathbb{I}_{\{\Gamma_+\geq 0\}}s_{\nu'}\left(\Gamma_+\right) \\
-&C^{\mathcal{Q}}2\pi\Phi_{\nu,p}(w,\phi)\sum_{\nu',p'}\frac{\gamma_{\nu\to\nu'}\left(N_{\nu\to\nu'}+1\right)}{W_{(\nu,p)\leftrightarrow(\nu',p')}}\mathbb{I}_{\{\Gamma_-\geq 0\}}s_{\nu'}\left(\Gamma_-\right)
\end{aligned}
$$

Outline

The Newton scheme

Iterative scheme

The Schrödinger-Poisson block

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

$$-\operatorname{div}\left[\varepsilon_R\nabla V\right]=-\frac{q}{\varepsilon_0}\left(2\sum_{\nu,p}\varrho_{\nu,p}\left|\psi_{\nu,p}[V]\right|^2-N_D\right)$$

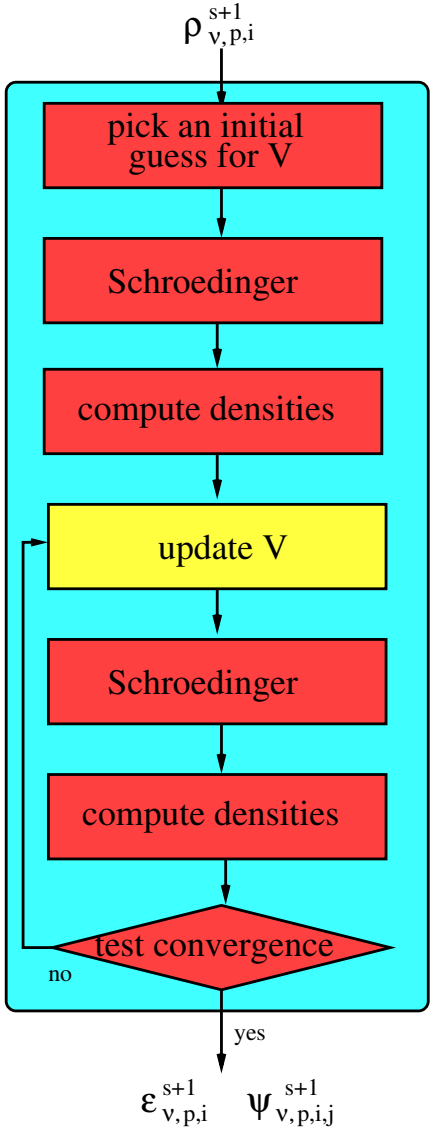
is solved thanks to a Newton-Raphson iterative methods. After calculations, the scheme boils down to refining (indexed on k)

$$-\operatorname{div}_{x,z}\left[\varepsilon_{\mathrm{R}}(x,z)\,\nabla V^{(k+1)}(x,z)\right]+\int\mathcal{A}^{(k)}(x,z,\zeta)\,V^{(k+1)}(x,\zeta)\,\mathrm{d}\zeta$$

$$=-2\sum_{\nu,p}\varrho_{\nu,p}\left|\psi^{(k)}_{\nu,p}[V]\right|^2+\int\mathcal{A}^{(k)}(x,z,\zeta)\,V^{(k)}(x,\zeta)\,\mathrm{d}\zeta$$

which represents a **linear system on $V^{(k+1)}$**. We iterate on k until convergence.

Framework


$\rho^{s+1}_{\nu,p,i}$
pick an initial guess for V
Schroedinger
compute densities
update V
Schroedinger
compute densities
test convergence
no
yes
$\varepsilon^{s+1}_{\nu,p,i}$
$\psi^{s+1}_{\nu,p,i,j}$

Outline

Summary

The dimensions

(i). The **valley** (the silicon band structure) $\nu \in \{0, 1, 2\}$.

(ii). The **subband** (the energy level's index): $p \in \{0, \ldots, N_{\text{sbn}} - 1\}$.

(iii). The **longitudinal dimension** (unconfined): $x_i = i \times \underbrace{\frac{1}{N_x - 1}}_{\Delta x}$.

(iv). The **transversal dimension** (confined): $z_j = j \times \underbrace{\frac{1}{N_z - 1}}_{\Delta z}$.

(v). The **energy**: $w_\ell = (\ell + 0.5) \times \underbrace{\frac{w_{\max}}{N_w - 1}}_{\Delta w}$.

(vi). The **angle**: $\phi_m = m \times \underbrace{\frac{2\pi}{N_\phi}}_{\Delta \phi}$.

(vii). As for the **time** step, it is adapted following a CFL condition.

Summary: the transport part

The magnitudes

	dimensions N	1	2	3	4	5	RAM (KB)
$\Phi_{\nu,p,i,\ell,m}$	5	m	ℓ	p	ν	i	≈ 130000
$D_{\nu,p,i,\ell,m}$	5	m	ℓ	p	ν	i	≈ 130000
$Q_{\nu,p,i,\ell,m}$	5	m	ℓ	p	ν	i	≈ 130000
$\mathcal{H}_{\nu,p,i,\ell,m}$	5	m	ℓ	p	ν	i	≈ 130000
$\mathcal{W}_{\nu,p,\nu',p',i}$	5	p'	ν'	p	ν	i	≈ 165
$\tilde{\Phi}_{\nu,p,i,\ell}$	4	ℓ	p	ν	i		≈ 2740
	dimensions N	1	2	3	4	5	RAM and DRAM (KB)
$\epsilon_{\nu,p,i}$	3	p	ν	i			≈ 28
$\varrho_{\nu,p,i}$	3	p	ν	i			≈ 28
$\psi_{\nu,p,i,j}$	4	j	p	ν	i		≈ 1800
	dimensions N	1	2	3	4	5	DRAM (KB)
$V_{i,j}$	2	j	i				≈ 33
$N_{\nu,p,i,j}$	4	p	ν	j	i		≈ 1800
$\mathcal{A}_{i,j,j'}$	3	j'	j	i			≈ 2140

Summary: the transport part

WENO

$$\frac{\partial}{\partial x}\left[a^1 \cdot \Phi^s\right]_{\nu,p,i,\ell,m} \approx \frac{\left(\widehat{a^1_{\nu,\ell,m} \cdot \Phi^s_{\nu,p,\cdot,\ell,m}}\right)_{i+\frac{1}{2}} - \left(\widehat{a^1_{\nu,\ell,m} \cdot \Phi^s_{\nu,p,\cdot,\ell,m}}\right)_{i-\frac{1}{2}}}{\Delta x},$$

$$\frac{\partial}{\partial w}\left[a^{2,s} \cdot \Phi^s\right]_{\nu,p,i,\ell,m} \approx \frac{\left(\widehat{a^{2,s}_{\nu,p,i,\cdot,m} \cdot \Phi^s_{\nu,p,i,\cdot,m}}\right)_{\ell+\frac{1}{2}} - \left(\widehat{a^{2,s}_{\nu,p,i,\cdot,m} \cdot \Phi^s_{\nu,p,i,\cdot,m}}\right)_{\ell-\frac{1}{2}}}{\Delta w},$$

$$\frac{\partial}{\partial \phi}\left[a^{3,s} \cdot \Phi^s\right]_{\nu,p,i,\ell,m} \approx \frac{\left(\widehat{a^{3,s}_{\nu,p,i,\ell,\cdot} \cdot \Phi^s_{\nu,p,i,\ell,\cdot}}\right)_{m+\frac{1}{2}} - \left(\widehat{a^{3,s}_{\nu,p,i,\ell,\cdot} \cdot \Phi^s_{\nu,p,i,\ell,\cdot}}\right)_{m-\frac{1}{2}}}{\Delta \phi}.$$

Summary: the transport part

Scatterings

$$Q^{\mu,\text{gain}}_{\nu,p,i,\ell} = \mathcal{C}^{\mu}\, 2 \sum_{p'=0}^{N_{\text{sbn}}-1} \sum_{\nu'=0}^{2} \mathcal{W}_{\nu,p,\nu',p',i}$$
$$\left\{ \mathbb{1}\left(\Gamma^{\mu,+}_{\nu,p,\nu',p',i,\ell} \geq 0\right) \cdot \left(N^{\mu}_{\nu',\nu} + 1\right) \cdot s_{\nu'}(w_\ell) \cdot \text{LI}\left[\tilde{\Phi}_{\nu',p',i,\cdot}\right]\left(\Gamma^{\mu,+}_{\nu,p,\nu',p',i,\ell}\right) \right.$$
$$\left. + \mathbb{1}\left(\Gamma^{\mu,-}_{\nu,p,\nu',p',i,\ell} \geq 0\right) \cdot N^{\mu}_{\nu',\nu} \cdot s_{\nu}(w_\ell) \cdot \text{LI}\left[\tilde{\Phi}_{\nu',p',i,\cdot}\right]\left(\Gamma^{\mu,-}_{\nu,p,\nu',p',i,\ell}\right) \right\}$$

$$Q^{\mu,\text{loss}}_{\nu,p,i,\ell,m} = -\mathcal{C}^{\mu}\, 4\pi\, \Phi^{s}_{\nu,p,i,\ell,m} \sum_{p'=0}^{N_{\text{sbn}}-1} \sum_{\nu'=0}^{2} \mathcal{W}_{\nu,p,\nu',p',i}$$
$$\left\{ \mathbb{1}\left(\Gamma^{\mu,+}_{\nu,p,\nu',p',i,\ell} \geq 0\right) N^{\mu}_{\nu,\nu'} \cdot s_{\nu'}\left(\Gamma^{\mu,+}_{\nu,p,\nu',p',i,\ell}\right) \right.$$
$$\left. + \mathbb{1}\left(\Gamma^{\mu,-}_{\nu,p,\nu',p',i,\ell} \geq 0\right) + \left(N^{\mu}_{\nu,\nu'} + 1\right) \cdot s_{\nu'}\left(\Gamma^{\mu,-}_{\nu,p,\nu',p',i,\ell}\right) \right\}.$$

Summary: the transport part

Scatterings

The overlap integral is: $\mathcal{W}_{\nu,p,\nu',p',i} = \Delta z \sum_{j=1}^{N_z-2} \left|\psi_{\nu,p,i,j}\right|^2 \left|\psi_{\nu',p',i,j}\right|^2$.

The ϕ-integrated distribution function is $\tilde{\Phi}_{\nu,p,i,\ell} := \Delta\phi \sum_{m=0}^{N_\phi-1} \Phi^s_{\nu,p,i,\ell,m}$.

The linear interpolation is:

$$\mathrm{LI}\left[\tilde{\Phi}_{\nu,p,i,\cdot}\right](\Gamma) := \frac{\tilde{\Phi}_{\nu,p,i,\ell_\mathrm{u}} - \tilde{\Phi}_{\nu,p,i,\ell_\mathrm{d}}}{\Delta w} \cdot \Gamma + \frac{w_{\ell_\mathrm{u}} \cdot \tilde{\Phi}_{\nu,p,i,\ell_\mathrm{d}} - w_{\ell_\mathrm{d}} \cdot \tilde{\Phi}_{\nu,p,i,\ell_\mathrm{u}}}{\Delta w} \times \mathbb{1}\left(\Gamma \geq 0 \quad \wedge \quad \ell_\mathrm{d} \leq N_w - 2\right)$$

Surface densities

The surface densities are $\varrho_{\nu,p,i} = \Delta w \sum_{\ell=0}^{N_w-1} \tilde{\Phi}_{\nu,p,i,\ell}$.

Summary: the confinement

Schrödinger

We have to compute selected eigenvalues and eigenvectors of matrices whose elements in the diagonal are $\left(\frac{1/4}{m_{z,\nu,i,j-1}} + \frac{1/2}{m_{z,\nu,i,j}} + \frac{1/4}{m_{z,\nu,i,j+1}} \over \Delta z^2 - V_{i,j} \right)_{j=1,\ldots,N_z-2}$

and in the sub- and super-diagonals are $\left(-\frac{1/4}{m_{z,\nu,i,j}} + \frac{1/4}{m_{z,\nu,i,j+1}} \over \Delta z^2 \right)_{j=1,N_z-3}$.

We have 195 independent problems of diagonalization of 63×63 matrices.

Summary: the confinement

The linear system

$$
\begin{aligned}
&\left(\operatorname{div}\left[\varepsilon_R \nabla V^{(k+1)}\right]\right)_{i,j} \\
&= \left(\frac{\frac{1}{2}(\varepsilon_R)_{i-1,j} + \frac{1}{2}(\varepsilon_R)_{i,j}}{\Delta x^2}\right) V_{i-1,j}^{(k+1)} + \left(\frac{\frac{1}{2}(\varepsilon_R)_{i,j-1} + \frac{1}{2}(\varepsilon_R)_{i,j}}{\Delta z^2}\right) V_{i,j-1}^{(k+1)} \\
&- \left(\frac{\frac{1}{2}(\varepsilon_R)_{i-1,j} + (\varepsilon_R)_{i,j} + \frac{1}{2}(\varepsilon_R)_{i+1,j}}{\Delta x^2} + \frac{\frac{1}{2}(\varepsilon_R)_{i,j-1} + (\varepsilon_R)_{i,j} + \frac{1}{2}(\varepsilon_R)_{i,j+1}}{\Delta z^2}\right) V_{i,j}^{(k+1)} \\
&+ \left(\frac{\frac{1}{2}(\varepsilon_R)_{i,j} + \frac{1}{2}(\varepsilon_R)_{i,j+1}}{\Delta z^2}\right) V_{i,j+1}^{(k+1)} + \left(\frac{\frac{1}{2}(\varepsilon_R)_{i,j} + \frac{1}{2}(\varepsilon_R)_{i+1,j}}{\Delta x^2}\right) V_{i+1,j}^{(k+1)} \\
&+ \frac{\Delta z}{2} \cdot \left[\sum_{j'=0}^{N_z-2} \mathcal{A}_{i,j,j'}^{(k)} V_{i,j'}^{(k+1)} + \sum_{j'=1}^{N_z-1} \mathcal{A}_{i,j,j'}^{(k)} V_{i,j'}^{(k+1)}\right] = \text{right hand side}
\end{aligned}
$$

where $\mathcal{A}_{i,j,j'}^{(k)} = 2\sum_{\nu,p}\sum_{p'\neq p} \frac{\varrho_{\nu,p,i}^{s+1} - \varrho_{\nu,p',i}^{s+1}}{\epsilon_{\nu,p',i}^{(k)} - \epsilon_{\nu,p,i}^{(k)}} \times \psi_{\nu,p,i,j'}^{(k)} \psi_{\nu,p',i,j'}^{(k)} \psi_{\nu,p',i,j}^{(k)} \psi_{\nu,p,i,j}^{(k)}$

The matrix is square of order 4225, has 129 diagonals and is sparse (98%).

Outline

CUDA and openMP

The language

The solver is implemented in C++.

The iterative part *ITER*

The iterative scheme is solved on the CPU, and is parallelized using openMP.

The transport part *BTE*

The transport part is fully solved on the GPU, by means of CUDA extensions to C++.

CUDA and openMP

The language

The solver is implemented in C++.

The iterative part *ITER*

The iterative scheme is solved on the CPU, and is parallelized using openMP.

The transport part *BTE*

The transport part is fully solved on the GPU, by means of CUDA extensions to C++.

CUDA and openMP

The language

The solver is implemented in C++.

The iterative part *ITER*

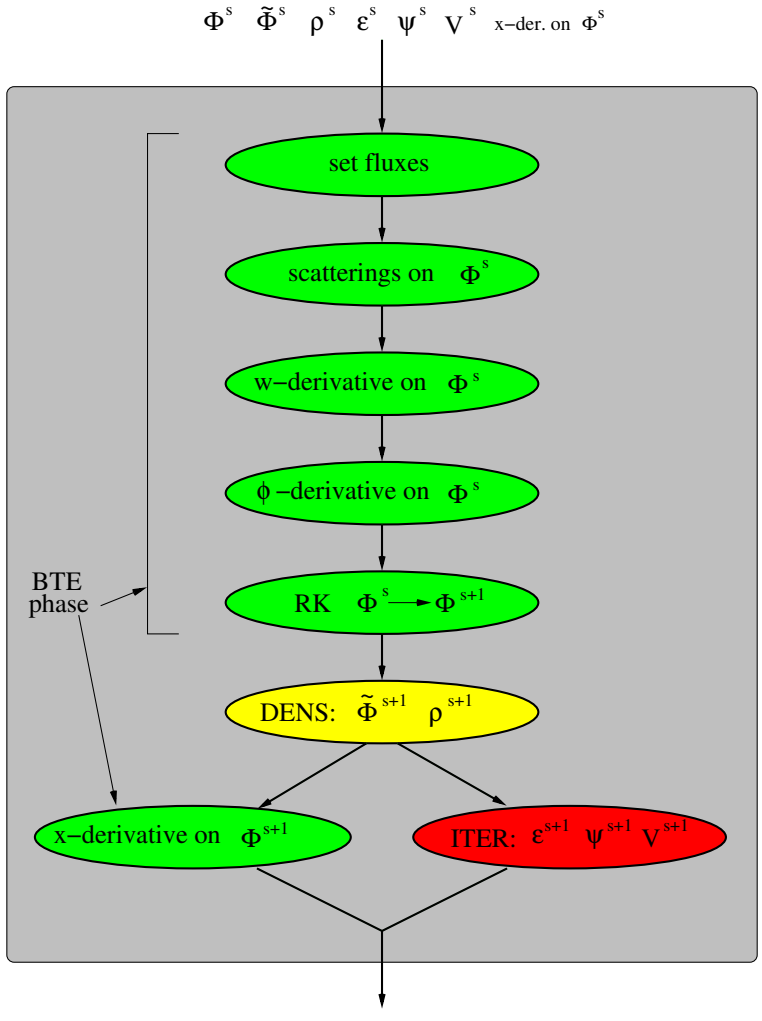
The iterative scheme is solved on the CPU, and is parallelized using openMP.

The transport part *BTE*

The transport part is fully solved on the GPU, by means of CUDA extensions to C++.

Overall design of the solver

Here we depict one Runge-Kutta step. Note the **overlap** of some computations.



Some remarks

Memory usage

The only magnitudes that are allocated on both the RAM (CPU) and the DRAM (GPU) are:

- $\varrho_{\nu,p}(x)$, that the GPU transfers to the CPU (about 28 KB).
- $\epsilon_{\nu,p}(x)$, that the CPU transfers to the GPU (about 28 KB).
- $\psi_{\nu,p}(x,z)$, that the CPU transfers to the GPU (about 1800 KB).

All the other magnitudes are either only on the RAM or only on the DRAM.

Linear system

For the linear system, the Library of Iterative Solvers (LIS) has been used. An iterative method has been chosen, the BICGSTAB, preconditioned by ILUT.

Eigenproblems

The `dsgetr` LAPACK routine is exploited for the computation of eigenstates and eigenvalues (bounded to 6, in our configuration).

Some remarks

Memory usage

The only magnitudes that are allocated on both the RAM (CPU) and the DRAM (GPU) are:

- $\varrho_{\nu,p}(x)$, that the GPU transfers to the CPU (about 28 KB).
- $\epsilon_{\nu,p}(x)$, that the CPU transfers to the GPU (about 28 KB).
- $\psi_{\nu,p}(x,z)$, that the CPU transfers to the GPU (about 1800 KB).

All the other magnitudes are either only on the RAM or only on the DRAM.

Linear system

For the linear system, the Library of Iterative Solvers (LIS) has been used. An iterative method has been chosen, the BICGSTAB, preconditioned by ILUT.

Eigenproblems

The `dsgetr` LAPACK routine is exploited for the computation of eigenstates and eigenvalues (bounded to 6, in our configuration).

Some remarks

Memory usage

The only magnitudes that are allocated on both the RAM (CPU) and the DRAM (GPU) are:

- $\varrho_{\nu,p}(x)$, that the GPU transfers to the CPU (about 28 KB).
- $\epsilon_{\nu,p}(x)$, that the CPU transfers to the GPU (about 28 KB).
- $\psi_{\nu,p}(x,z)$, that the CPU transfers to the GPU (about 1800 KB).

All the other magnitudes are either only on the RAM or only on the DRAM.

Linear system

For the linear system, the Library of Iterative Solvers (LIS) has been used. An iterative method has been chosen, the BICGSTAB, preconditioned by ILUT.

Eigenproblems

The `dsgetr` LAPACK routine is exploited for the computation of eigenstates and eigenvalues (bounded to 6, in our configuration).

Some remarks

Overlap

The x-derivative $\frac{\partial}{\partial x}\left[a_\nu^1 \Phi_{\nu,p}\right]$ and the *ITER* block can be performed simultaneously, because the flux coefficient a_ν^1 is constant in time, hence it does not depend on the eigenstates.

Shared memory

Use of shared memory for the sake of coalescent reading when the data distribution in the DRAM is not favourable. This is exploited for the computations of the w-derivative, the ϕ-derivative, the ϕ-integration of the pdf and the integration of the loss part of the scattering operator.

Bank conflicts

Reading from the shared memory must be performed carefully in order to avoid bank conflicts.

Some remarks

Overlap

The x-derivative $\frac{\partial}{\partial x}\left[a_\nu^1 \Phi_{\nu,p}\right]$ and the *ITER* block can be performed simultaneously, because the flux coefficient a_ν^1 is constant in time, hence it does not depend on the eigenstates.

Shared memory

Use of shared memory for the sake of coalescent reading when the data distribution in the DRAM is not favourable. This is exploited for the computations of the w-derivative, the ϕ-derivative, the ϕ-integration of the pdf and the integration of the loss part of the scattering operator.

Bank conflicts

Reading from the shared memory must be performed carefully in order to avoid bank conflicts.

Some remarks

Overlap

The x-derivative $\frac{\partial}{\partial x}\left[a_\nu^1 \Phi_{\nu,p}\right]$ and the *ITER* block can be performed simultaneously, because the flux coefficient a_ν^1 is constant in time, hence it does not depend on the eigenstates.

Shared memory

Use of shared memory for the sake of coalescent reading when the data distribution in the DRAM is not favourable. This is exploited for the computations of the w-derivative, the ϕ-derivative, the ϕ-integration of the pdf and the integration of the loss part of the scattering operator.

Bank conflicts

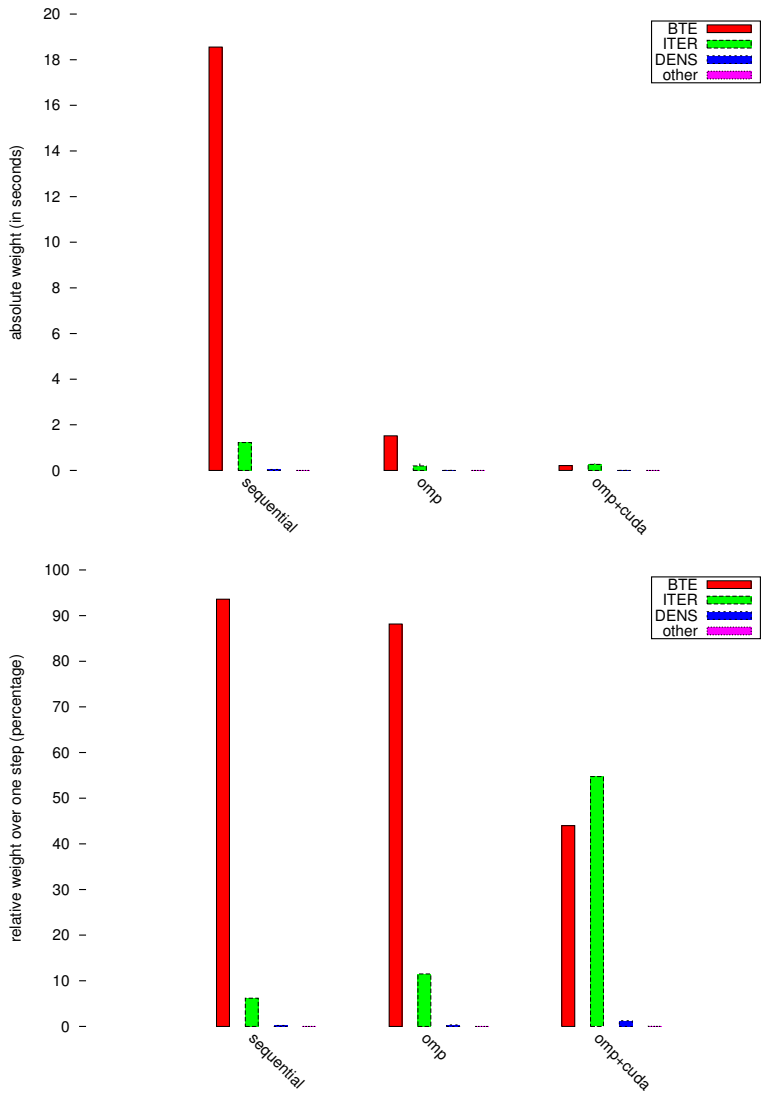
Reading from the shared memory must be performed carefully in order to avoid bank conflicts.

Outline

Speedups

Version	step	BTE	DENS	ITER	FD-WENO-5	scatt.	lin. sys.	Schröd.	NR kernel
sequential	19.82	18.56	0.044	1.22	3.69	13.70	0.59	0.23	0.37
OMP 2-core	10.05	9.35	0.024	0.67	1.87	6.89	0.34	0.12	0.19
OMP 4-core	5.60	5.18	0.0146	0.40	1.03	3.82	0.207	0.066	0.106
OMP 6-core	4.05	3.72	0.011	0.317	0.74	2.74	0.167	0.048	0.076
OMP 8-core	3.32	3.04	0.0103	0.27	0.61	2.24	0.145	0.04	0.063
OMP 10-core	2.83	2.58	0.0086	0.24	0.618	1.8	0.136	0.032	0.051
OMP 12-core	2.39	2.16	0.0075	0.21	0.53	1.5	0.119	0.026	0.042
OMP 14-core	2.07	1.85	0.0068	0.208	0.46	1.28	0.121	0.022	0.036
OMP 16-core	1.73	1.53	0.0062	0.199	0.31	1.12	0.119	0.02	0.032
OMP 16-core/GPU	0.47	0.21	0.00618	0.26	0.087	0.105	0.15	0.03	0.05

Main computational phases



GFlops/s

kernel	avg. time	GFlops/s
phonons, loss	32.8 ms	539
ϕ-derivative	10.7 ms	598
w-derivative	10.6 ms	284
x-derivative	6.66 ms	389
a^3 computation	2.92 ms	226
ϕ-integrated pdf	1.81 ms	9
phonons, gain	2.47 ms	275
RK, 2nd stage	3.59 ms	28
RK, 3rd stage	3.59 ms	28
RK, 1st stage	2.90 ms	11
overlap integral	.297 ms	16
surface densities	.160 ms	2

Gràcies!

Los autores agradecen los proyectos **MTM2014-52056-P** y **MTM2017-85067-P** financiados por el Ministerio de Economía y competitividad, y el Fondo Europeo de Desarrollo Regional (ERDF/FEDER).