A Hybrid Mode-Space/Real-Space Scheme for DFT+NEGF Device Simulations

F. Ducry*, M. H. Bani-Hashemian, and M. Luisier
Integrated Systems Laboratory (ETH Zurich), Email: fabian.ducry@iis.ee.ethz.ch

***Abstract*—Density functional theory based simulation techniques enable thorough investigation of the operational characteristics of nanoscale devices regardless of their configurational complexity. However, this flexibility comes with considerable computational cost. In this work, we present a hybrid mode-space/real-space scheme that utilizes a mode-space basis to represent periodic contacts while maintaining the real-space representation of the central device region. Reducing the size of the contact blocks via mode-space approximation speeds up the calculation of the open boundary conditions and reduces the overall size of the Hamiltonian and overlap matrices, which leads to significant improvements in the computational efficiency of simulations. Keeping the real-space representation of the device blocks preserves the versatility and accuracy of the *ab-initio* approach. The merits of the proposed method are demonstrated with the simulation of an amorphous device with metallic contacts.**

I. Introduction

Following the continuous decrease of the device dimensions, atomistic quantum mechanical simulation approaches have become essential to accurately predict the performance of nanoscale components. Complex material stacks, metallic contacts, and amorphous layers further call for an *ab-initio* treatment of their electronic properties. Coupling density functional theory (DFT) with the Non-equilibrium Green's Function (NEGF) formalism [1,2] meets this requirement. Consequently, DFT+NEGF has established itself as a reference to shed light on the behavior of nano-devices. However, it is typically limited to small atomic systems because of its heavy computational burden. This issue can be addressed by reducing the dimensionality of the real-space (RS) Hamiltonian matrix. Among the techniques proposed to do so, the so-called mode-space (MS) method [3,4] stands out. The MS transformation uses a reduced basis set to accurately reproduce the band structure of a periodic cell within a restricted energy interval only. As such, it does not lend itself to structures made of non-repeatable unit cells, which is true for most realistic components featuring interfaces and amorphous phases. This is the case, for example, of conductive bridging random access memories (CBRAMs), where an amorphous oxide is surrounded by two electrodes, as illustrated in Fig. 1(a) [5].

Here, we propose a hybrid MS/RS scheme that decouples the contact extensions, composed of periodic cells, from the central oxide region with random atomic placement. While the former parts are converted to MS, the latter remains expressed in RS. This approach offers two key advantages over pure RS simulations. By reducing the size of the contact blocks within the Hamiltonian matrix H, we significantly decrease the time to calculate the open boundary conditions (OBCs) and to solve the NEGF equations. At the same time, the memory required to store H drastically goes down.

Beside this innovation, we also successfully apply an MS transformation to metallic electrodes, fully capturing their band structure and accurately calculating the ballistic current flowing through the resulting device.

II. Approach

The MS transformation is used to precisely reproduce the RS band structure within the energy window of interest. This transformation usually introduces unphysical energy states that need to be removed. The contact RS-to-MS transformation matrix U can be obtained and refined using the method of Refs. [3,4] with a few modifications to process large metallic blocks. First, an initial guess for U is created based on the eigenvectors of the contact unit cell. Previous works disregarded interactions beyond nearest neighbors [4], which is acceptable for semiconductors. Metals however, usually feature delocalized electrons resulting in long range connections that cannot be ignored without loss of accuracy. To account for these, the following adaptations were developed. In a system periodic along the x-axis, the E-k relation is given by

$$H_{k_x}\Psi_{k_x} = E(k_x)S_{k_x}\Psi_{k_x}, \tag{1}$$

where H_{k_x} and S_{k_x} are the k_x-dependent Hamiltonian and overlap matrices, respectively, $E(k_x)$ the energy at k_x, and Ψ_{k_x} the corresponding wave function. For unit cells interacting with N_N neighbors along the $\pm x$-direction H_{k_x} is obtained according to

$$H_{k_x} = H_0 + \sum_{n=1}^{N_N}\left(H_n e^{ink_x} + H_n^\dagger e^{-ink_x}\right), \tag{2}$$

where H_n connects one periodic unit cell of width Δ_x at x_i to another one at $x_i + n\Delta_x$, as illustrated in Fig. 1(b). Note that S_{k_x} has the same structure as H_{k_x}. Based on H_{k_x} and S_{k_x} the initial guess is calculated according to [4].

Once U is obtained, it needs to be iteratively refined to eliminate unphysical energy states. The elimination of these energies is achieved by iteratively adding additional normalized basis vectors in the form of $\Xi \cdot C$ to U. Ξ is a trial basis and C the vector that minimizes the following expression:

$$\mathcal{F}(C) = \frac{1}{n_z}\sum_q\sum_z\left[\frac{C^T A(q,z)C}{C^T B(q,z)C}(z-\epsilon_c)\right] + (C^T C - 1)^2, \tag{3}$$

978-1-7281-0940-4/19/$31.00 ©2019 IEEE

where q and z are a set of trial k_x and energies, respectively, A and B are matrices defined as in [4] and depending on U, H_q, S_q, and Ξ. The ability of C to effectively remove unphysical bands strongly depends on the choice of Ξ. Previous works used the commutator $[S_{k_{x1}}^{-1}H_{k_{x1}}, S_{k_{x2}}^{-1}H_{k_{x2}}]$, with $k_{x1} = 0$ and $k_{x2} = \pi$. We found a better performance of the removal procedure using the k_x with the largest number of unphysical energies, in our case $k_{x1,x2} = \pm\pi/5$ such that

$$\Xi = (1 - UU^{\dagger})[S_{k=\pi/5}^{-1}H_{k=\pi/5}, S_{k=-\pi/5}^{-1}H_{k=-\pi/5}]U. \quad (4)$$

Finding the global minimum of Eq. (3) is a challenging problem as the function has many local minima. Depending on the initial guess, an optimizer not always finds a vector C that removes an unphysical band from the MS band structure. To remedy this shortcoming, we applied multiple initial guesses and kept the result with the lowest value of $\mathcal{F}(C)$. The rows of $\Re\{U\}$ proved to be the best initial guesses for C.

Convergence of the optimization process was improved using the analytical first derivative:

$$\frac{\partial \mathcal{F}(C)}{\partial C} = \sum_q \sum_z \left[\frac{C^T A(q,z)}{C^T B(q,z) C} - \frac{C^T A(q,z) C}{(C^T B(q,z) C)^2} C^T B(q,z) \right] + 4(C^T C - 1)C, \quad (5)$$

which is a vector of the same length as C. Note that A and B are assumed to be symmetric without loss of generality and the scalar factors of Eq. (3) are absorbed into A.

The transformation of H_n (S_n) from RS to MS requires to project each contact block onto the created MS basis. The RS→MS coupling is achieved by partial, one-sided projection of the RS block onto MS. The MS blocks $\widetilde{H}_n$ and hybrid blocks $\mathcal{H}_n$ are then defined as

$$\widetilde{H}_n = U^{\dagger} H_n U, \ \mathcal{H}_{n,RM} = U^{\dagger} H_n \text{ or } \mathcal{H}_{n,MR} = H_n U. \quad (6)$$

The $\mathcal{H}_{n,RM/MR}$ matrices are the hybrid blocks at the RS-MS interface. Eq. (6) is only applied to the contacts, not to the central oxide and the metallic interface layers attached to it.

III. Results

As benchmark example, a CBRAM cell made of 3870 atoms, with two Cu contacts separated by amorphous silicon dioxide (a-SiO_2) is considered [6]. The H and S matrices were computed from DFT with CP2K [7] using double-zeta valence (DZVP) basis sets, and their contact extensions converted into MS. In RS each Cu unit cell is made of 240 atoms with 25 orbitals per atom summing up to 6000 orbitals per block. The RS→MS transformation reduced the original block size to 692 (11.5% of the original value), as listed in Fig. 2(a). The success of the transformation is illustrated with the RS and MS contact band structures in Fig. 2(b). The MS band structure is shown both before and after the refinement procedure, in the left and right half respectively. After refinement, the RS and MS results almost perfectly agree with a maximum error of 0.047 meV. Moreover, MS does not exhibit spurious states anymore within the energy interval of interest.

The sparsity pattern of the device Hamiltonian matrices before and after Eq. (6) are plotted in Fig. 3(a-b) for a subset of the CBRAM in Fig. 1(a) (left contact and Cu/a-SiO_2 interface). Five Cu contact blocks are converted to MS, while one Cu block at the interface and the a-SiO_2 remain in RS. It can be observed that the off-diagonal blocks coupling the MS to RS domains change shape and become thin rectangles.

To validate our approach, H and S were passed to our in-house quantum transport solver [8]. The energy-resolved transmission and low field IV-curve through the investigated cell, calculated with the RS and MS/RS matrices, are reported in Figs. 3(c) and 4(a): the MS/RS transmission overlaps almost perfectly with the RS one, and the IVs agree very well, confirming the strength and validity of the hybrid scheme.

For improved simulation speed, the metal blocks at the interface can also be converted into MS at the cost of accuracy. The relative error of the IVs calculated from hybrid matrices with zero and one RS metal blocks is shown in Fig. 4(b). With a single Cu block in RS representation the current is within 1% of the original value. Transforming all electrode blocks to MS, however, causes an underestimation of the current by up to 20%. It is apparent that one RS block at the interface is required to capture all necessary interactions.

The computational efficiency of the proposed method was also tested on the Piz Daint supercomputer at CSCS by measuring the time to calculate the OBCs and NEGF equations and the total time needed per energy point. Results are presented in Fig. 5. The evaluation of the OBCs on CPUs could be accelerated by a factor of 55, irrespective of the number of RS blocks at the interface. Solving the NEGF system on GPUs is accelerated by a factor 40 (65) for one (zero) RS block, for an overall speed up of 20 (40) per energy point. The reduced memory consumption helped decrease the required number of nodes per energy point from 20 with only RS blocks to 4 with the hybrid approach. Thus, the total cost is lowered by a factor of at least 20 (speed up) x 5 (node reduction) = 100.

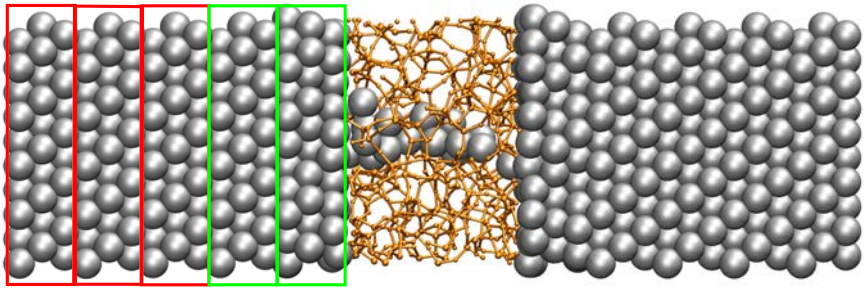
As a result of these improvements, at a fixed computational cost, the change in the resistance state of a CBRAM cell can be studied at more steps during the formation of filaments (see Fig. 6), as same MS transformation can be applied to the contact region of each case.

IV. Conclusion

We have developed a method that combines real-space and mode-space representations of non-homogeneous device structures, demonstrating both excellent physical accuracy and enhanced computational efficiency. By creating one transformation matrix U, we will be able to more rapidly simulate electron transport through dynamically evolving structures like CBRAM cells with different amorphous oxide layers and nano-filament configurations, but (almost) identical contacts.

Acknowledgment

This work was supported by the Werner Siemens Stiftung, by SNF under Grant No. PP00P2 159314, by ETH Research Grant ETH-35 15-2, and by a grant from the Swiss National Supercomputing Centre (CSCS) under Project s714.



(a) 3D CBRAM Structure

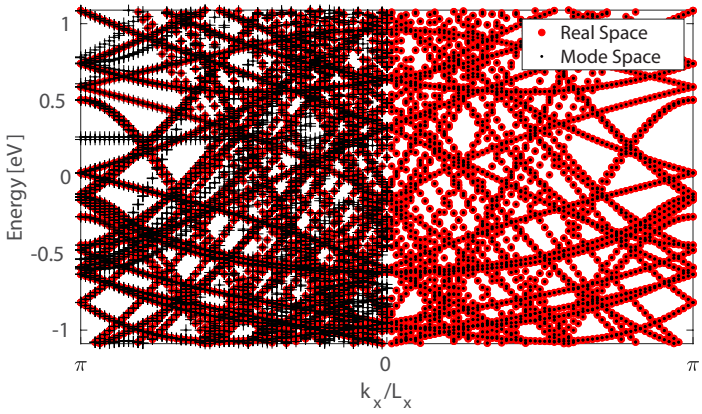
$$H = \begin{pmatrix} \ddots & & & & \\ & H_0 & H_1 & \cdots & H_{N_N} \\ & H_1^\dagger & H_0 & H_1 & \vdots \\ & \vdots & H_1^\dagger & H_0 & H_1 \\ & H_{N_N}^\dagger & \cdots & H_1^\dagger & H_0 \\ & & & & \ddots \end{pmatrix}$$

(b) Block N_B-diagonal Pattern of H

Fig. 1. (a) Atomistic Cu/a-SiO_2/Cu CBRAM cell structure used in this work to benchmark the proposed hybrid mode-space/real-space approach. It is composed of 3870 atoms. Grey spheres represent Cu atoms (25 orbitals per atom), the orange ones Si and O (both 13 orbitals per atom). The Cu contacts are in fact longer than shown here. The red rectangles mark the contact blocks, while the green ones refer to the interface blocks. (b) Typical block N_B-diagonal pattern of the real-space Hamiltonian matrix H corresponding to the structure in (a). Each block represents a unit cell. Due to long-range interactions, N_B=5 for N_N=2. The mode-space Hamiltonian has the same pattern with much smaller blocks.

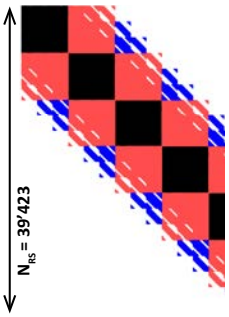
RS block size:	6000
Initial guess size:	538
Refinement iterations:	154
MS block size:	692
Size reduction:	88.5%
Energy window:	2 eV
Number of k-points:	128
Max. error band structure:	0.047 meV

(a) Mode-Space Transformation

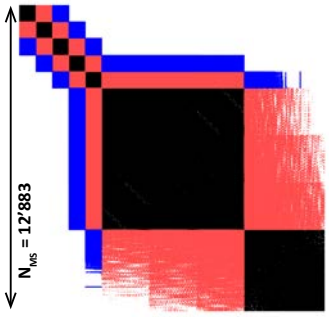


(b) Cu Contact Band Structure

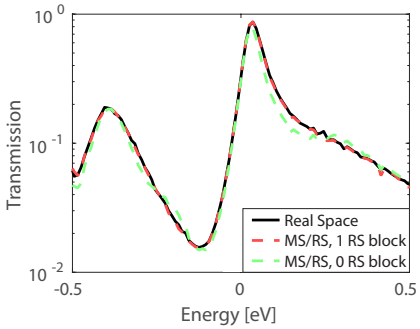
Fig. 2. (a) Table summarizing the real-space to mode-space transformation parameters. For an energy window of 2 eV around the Fermi level, 128 k-points are necessary to sample the Cu contact band structure and obtain accurate results, leading to a reduction of the block size by 88.5%. The largest error in the MS band structure is well below 10^{-4} eV. (b) Band structure of the Cu contact. Large red dots refer to the real-space. On the left branch the black crosses correspond to the MS band structure before refinement. The right-hand-side shows with small black dots the final mode-space results. The unphysical bands have been removed and all spurious states have been successfully eliminated.



(a) Real-Space Hami

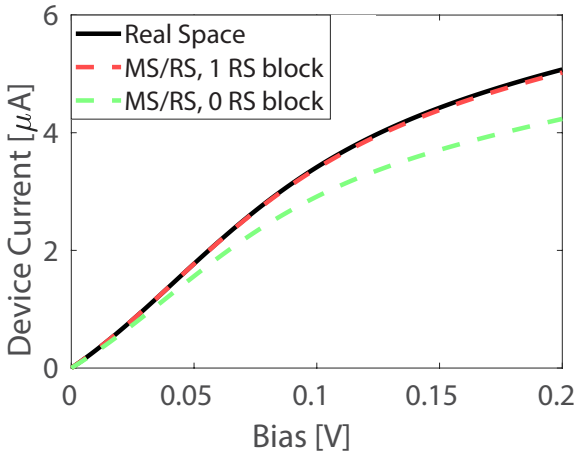


(b) Hybrid Hamiltonian

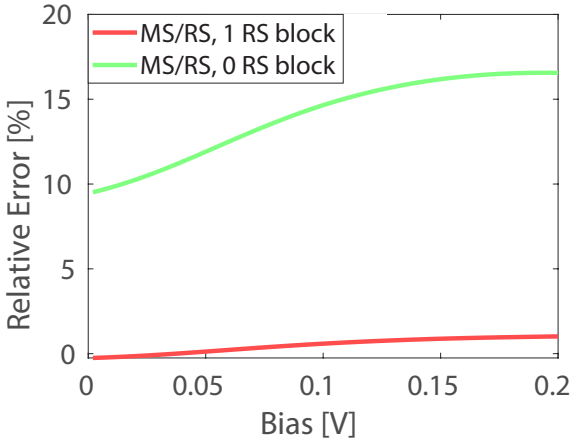


(c) Transmission profile through the device

Fig. 3. (a) Sparsity pattern of the real-space, penta-diagonal Hamiltonian matrix (left contact and Cu/a-SiO_2 interface). Black blocks correspond to the different Cu contact and a-SiO_2 (last block, bottom-right) unit cells, red (blue) blocks to their (second) nearest-neighbor connections. (b) Same as (a) after transformation into mode-space. Note the rectangular shape of the RS→MS coupling blocks and the significant size reduction from $N_{RS} = 39'423$ to $N_{MS} = 12'883$. (c) Transmission function through the structure in Fig. 1(a) as a function of energy around the Fermi energy (0 eV). Real-space (solid black line) and mode-space (dashed lines) are shown. The red (blue) dashed line corresponds to hybrid Hamiltonian with one (zero) RS metal block. While the red and black curve perfectly agree, the blue one is slightly off.



(a) Low-field IV characteristics of the device



(b) Relative error with respect to the RS IV curve

Fig. 4. (a) Low-field IV curves for the same simulations as in Fig. 3(c). The RS and MS curves agree given that at least one interface block remains in RS. (b) Relative error of the MS IVs shown in (a) with respect to RS. With one RS metal block the relative error of the current remains below 1%. With all metal blocks transformed to MS, however, the current is underestimated by almost 20% at a bias of 0.2 V.

	RS	MS/RS Hybrid 1 RS block	Gain vs RS	MS/RS Hybrid 0 RS blocks	Gain vs RS
Matrix size:	78'846	25'766	3.1	15'150	5.2
Non-zero elements:	673.9e6	132.0e6	5.1	47.5e6	14.2
Time OBC [s]:	308.5	5.3	57.8	5.8	53.2
Time linear system [s]:	273.9	6.4	42.5	4.1	66.8
Total time [s]:	336.0	16.5	20.4	8.2	41.0
Nodes:	20	4	5	2	10
Cost [time*nodes]:	6720	66	101.8	16.2	414.8

Fig. 5. Computational benchmark: Table summarizing the numerical problem size and computing times obtained for both the real-space and mode-space Hamiltonians on the Piz Daint supercomputer at CSCS (http://www.cscs.ch). Each node of this machine features 12 Intel Haswell cores (64 GB RAM) and 1 Pascal 100 GPU (16 GB). All benchmarks were run on the same number of cores (20) and GPUs (4), but different number of nodes, 20 for RS and 4 (2) for MS/RS with one (zero) RS blocks. Significant speed up factors of 20 (40) and reductions of memory footprint (using 4 (2) instead of 20 nodes) were achieved by applying the proposed hybrid MS/RS scheme.

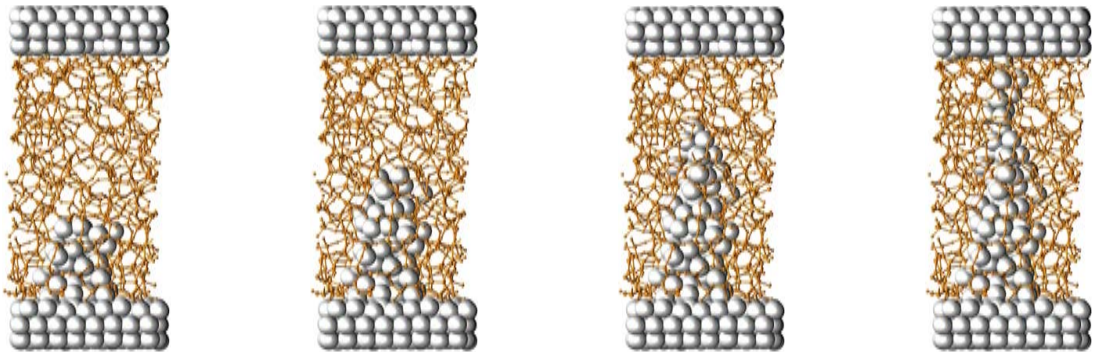


Fig. 6. Different CBRAM cells with the same cross section dimensions, but different nano-filament states. In the left-most structure, the central oxide is almost free of Cu atoms, while the top and bottom contacts are short-circuited by a filament in the right-most one. As the electrodes remain identical for all configurations, the same mode-space transformation can be applied to all of them to reduce their block size.

References: [1] M. Brandbyge et al., *Phys. Rev. B* 65, 165401 (2002). [2] M. V. Fernández-Serra et al., *Nano Lett.* 6, 2674 (2006). [3] G. Mil'nikov et al., *Phys. Rev. B* 85, 035317 (2012). [4] M. Shin et al., *J. Appl. Phys* 119, 154505 (2016). [5] I. Valov et al., *Nanotechnology* 22, 254003 (2011). [6] F. Ducry et al., in *Proceedings of the 2017 IEEE IEDM*, 4.2.1 (2017). [7] J. Hutter et al., *Comp. Molec. Sci.* 4, 15 (2014). [8] M. Luisier et al., *Phys. Rev. B* 74, 205323 (2006).