Finite-element implementation for electron transport in nanostructures

P. Havu[a)]
Laboratory of Physics, Helsinki University of Technology-TKK, P.O. Box 1100, FIN-02015 TKK, Finland

V. Havu
Institute of Mathematics, Helsinki University of Technology-TKK, P.O. Box 1100, FIN-02015 TKK, Finland

M. J. Puska, M. H. Hakala, A. S. Foster, and R. M. Nieminen
Laboratory of Physics, Helsinki University of Technology-TKK, P.O. Box 1100, FIN-02015 TKK, Finland

We have modeled transport properties of nanostructures using Green's-function method within the framework of the density-functional theory. The scheme is computationally demanding, so numerical methods have to be chosen carefully. A typical solution to the numerical burden is to use a special basis-function set, which is tailored to the problem in question, for example, the atomic-orbital basis. In this paper we present our solution to the problem. We have used the finite-element method with a hierarchical high-order polynomial basis, the so-called *p* elements. This method allows the discretion error to be controlled in a systematic way. The *p* elements work so efficiently that they can be used to solve interesting nanosystems described by nonlocal pseudopotentials. We demonstrate the potential of the implementation with two different systems. As a test system a simple Na-atom chain between two leads is modeled and the results are compared with several previous calculations. Secondly, we consider a thin hafnium dioxide (HfO_2) layer on a silicon surface as a model for a gate structure of the next generation of microelectronics. © *2006 American Institute of Physics.* [DOI: 10.1063/1.2162900]

I. INTRODUCTION

Using small nanoscale lithographic structures, atomic aggregates and even single molecules, it is possible to fabricate new kinds of electronic devices.[1] The function and scale of these devices are based on quantum-mechanical phenomena and cannot be described within the classical regime. Of particular relevance are the electron-transport properties of these nanoscale devices, as these will determine their effectiveness in, for example, a new generation of transistors. As the experimental work on these devices grows, increasing emphasis is placed on developing a matching theoretical description.[2,3] Although some efforts have included a full description of an electronic circuit,[4,5] current research is mainly focused to study single electronic components.

Density-functional theory (DFT) is widely used in atomistic modeling of material properties, and recently, also properties and phenomena in nanostructures. The power of DFT is in its capacity to treat accurately systems with hundreds of atoms, yet retain a full quantum-mechanical treatment. Although the full justification of use of the DFT in electron-transport calculations is debated,[6,7] we adopt it as a practical scheme to describe the real systems and devices.

In the Kohn-Sham scheme of DFT the electron density is calculated using single-particle wave functions. The explicit use of the wave functions in constructing the density suffices well in two kinds of systems. Either the system has a repeating structure so that it can be modeled with periodic boundary conditions or the system is so small that it can be calculated as a whole. In nanoelectronics, however, a system consists usually of a small finite part, the nanostructure, which is connected to the surrounding infinite leads. If one enforces periodic boundary conditions, even a large repeating supercell or calculation volume can cause finite-size effects with spurious results for electron transport.

A commonly used solution to this problem, which we have also employed, is to combine DFT with Green's-function formalism.[1] Green's functions are first constructed for the semi-infinite leads by using the analytically known or easily calculated wave functions. Once Green's function for the combined nanostructure and leads is constructed, the wave functions are no longer needed explicitly. This makes it possible to use open boundary conditions between the nanostructure and lead. In this way we have an effectively infinite system without periodicity, making the finite-size effects small. It is also possible to calculate the electric current through the system for a finite bias voltage between the leads in a self-consistent manner with the electron density. The ensuing model for the current is analogous to the Landauer-Bütteker model.[2] We have used nonlocal pseudopotentials for modeling atoms and the ideal metal "jellium" model for the leads. The charge density in the leads can be varied according to the conducting properties of the leads we wish to model.

The use of Green's functions instead of the explicit use of wave functions is computationally demanding. This is why special care has to be taken in choosing the numerical methods. The first implementations used tight-binding methods,[8,9] but a more typical solution is to expand Green's functions in a special basis tailored for the system. Common
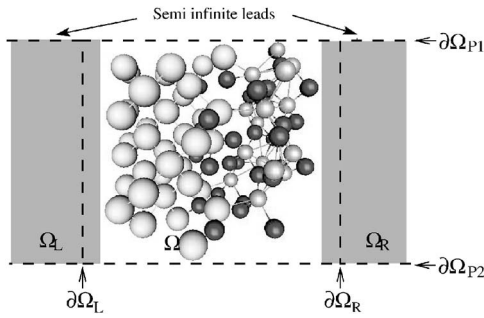
a)Electronic mail: pah@fyslab.hut.fi

FIG. 1. Schematic picture of the model. The HfO$_2$ interface is used as an example. The small and large gray spheres denote the Hf and Si atoms, respectively, and dark spheres the O atoms. The gray volumes are the jellium leads. The system consists of the volumes Ω_L, Ω, and Ω_R and the boundaries $\partial\Omega_L$, $\partial\Omega_R$, $\partial\Omega_{P1}$, $\partial\Omega_{P2}$, $\partial\Omega_{P3}$, and $\partial\Omega_{P4}$.

examples are localized atomic orbitals,[10,11] an $O(N)$ optimized basis,[12] a wavelet basis,[13] full-potential linearized augmented plane waves,[14] maximally localized Wannier functions,[15] a finite-difference method,[16] and a linear finite-element method.[17] Our solution is to use the finite-element method (FEM). It allows a systematic error control which is especially important in transport problems as there are many different properties which must be monitored. For example, the pole of Green's function can cause numerical problems. According to our experience electronic tunneling, in particular, is sensitive to numerical accuracy.

Besides systematic error control, the FEM also has other good properties which makes it a natural method for transport problems. It is a flexible method which allows one to take into account the geometry of the nanodevice exactly. Special boundary conditions are easy to derive without mixing the model with the numerical method and their implementation is straightforward. Moreover, the local nature of the basis produces sparse matrices for which efficient solving methods exist. Varying the size of the elements can be used to reduce the number of the basis functions and, consequently, the size of the system as compared with the finite-difference method. This is especially true for the high-order p method. Finally, there exists a lot of theoretical work together with tested and reliable tools, such as mesh generators and optimized linear solvers. These are used as standard building blocks for any FEM implementation granting easy access to state-of-the-art algorithms. Using the FEM, new theoretical or numerical ideas are easy to implement and test.

The structure of the paper is as follows: In Sec. II we describe the model itself in detail, including a discussion of the formalism of our implementation. In Sec. III we apply the model to two example systems: a Na-atom chain and HfO$_2$–Si interface between two leads. In Sec. IV we summarize the work. In this paper we use atomic units in all equations.

II. MODEL

A schematic picture of our model is shown in Fig. 1. Actually, the figure presents our second test case, the HfO$_2$–Si interface between two leads. We have an atomistic nanostructure between two semi-infinite leads. The system is divided into three parts, Ω being the calculation volume, and

Ω_L and Ω_R are left and right leads, respectively. The boundaries $\partial\Omega_{L/R}$ are open so that electrons can penetrate through them without any reflection or refraction. We use the DFT to model electron interactions. The basic quantity, the electron density, is calculated from single-particle Green's functions. Then we use the density to calculate the effective potential as

$$V_{\text{eff}} = V_{\text{ext}} + V_c + V_{\text{xc}} + V_{\text{bias}} + \hat{V}_{\text{nl}}, \tag{1}$$

where V_{ext} is the external potential caused by positive background charges, local parts of the pseudopotential operators, and the potential outside potential barriers. V_c is the Coulomb-Hartree interaction part and V_{xc} is the exchange-correlation part which we calculate using the local-density approximation based on the quantum Monte-Carlo data by Ceperley and Alder[18] and parametrized by Perdew and Zunger.[19] V_{bias} sets the boundary conditions if a bias voltage is applied. $\hat{V}_{\text{nl}}$ is the nonlocal part of the pseudopotential operators.

The Hartree potential is calculated from the modified Poisson equation,

$$\nabla^2 V_c^i - k_P^2 V_c^i = 4\pi(\rho_+ - \rho) - k_P^2 V_c^{i-1}, \tag{2}$$

where k_P is an adjustable parameter. k_P does not affect the final self-consistent result, but the stability and convergence of iterations are improved,[20] because the Coulomb potential due to charge redistribution between adjacent iterations is screened. The nonlocal pseudopotential is an operator given by

$$\hat{V}_{\text{nl}} v(\mathbf{r}) = \sum_{l,m} e_{l,m} \zeta_{l,m}(\mathbf{r}) \int_{\Omega} \zeta_{l,m}(\mathbf{r}') v(\mathbf{r}') d\mathbf{r}', \tag{3}$$

where $e_{l,m}$ and $\zeta_{l,m}(\mathbf{r})$ are defined using the Troullier-Martins pseudopotentials.[21,22] Equation (3) uses the projection of the function $v(\mathbf{r})$ (arbitrary function, which in practical calculations is a basis function) on the atomic-specific function $\zeta_{l,m}$ depending on the quantum numbers l and m corresponding to the angular momentum.

We have implemented the guaranteed-reduction Pulay[23] method for the mixing of the self-consistent iterations. It uses potentials from the five previous iterations for computing a new potential in such a way that the predicted norm of the potential residue is minimized. The simplest mixing scheme in which potentials are mixed with a linear feedback coefficient does not work well in open systems. The calculations are rather unstable so that quite a small feedback coefficient has to be used. This is because the net charge in the calculation volume Ω varies during the calculations.

A. Green's-function model

The details of Green's-function method for electron transport in nanostructures are explained, for example, in Ref. 24. Here we give only a short introduction to the equations to be solved. The retarded Green's function G^r is solved from the equation

$$(\omega - \hat{H}(\mathbf{r}))G^r(\mathbf{r},\mathbf{r}';\omega) = \delta(\mathbf{r}-\mathbf{r}'), \tag{4}$$

where ω is the electron energy and $\hat{H}$ is the DFT Hamiltonian of the system,

$$\hat{H}(\mathbf{r}) = -\frac{1}{2}\nabla^2 + V_{\text{eff}}(\mathbf{r}). \tag{5}$$

When we know G^r we can calculate the so-called lesser Green's function $G^<$. In the equilibrium when no bias voltage is applied over the nanostructure, it is obtained from

$$G^<(\mathbf{r},\mathbf{r}';\omega) = 2f_{L/R}(\omega)G^r(\mathbf{r},\mathbf{r}';\omega), \tag{6}$$

where $f_{L/R}$ are the Fermi functions of the leads. In the equilibrium, $f_L = f_R$. For a finite bias voltage $f_L \neq f_R$ and a more complicated equation for $G^<$ has to be used. To obtain it we write Eq. (4) in the form

$$(\omega - \hat{H}_0 - \Sigma_L^r(\omega) - \Sigma_R^r(\omega))G^r(\mathbf{r},\mathbf{r}';\omega) = \delta(\mathbf{r}-\mathbf{r}'), \tag{7}$$

where $\hat{H}_0$ is the Hamiltonian of the isolated volume Ω and $\Sigma_{L/R}^r$ are the so-called self-energies of the leads. We also define

$$i\Gamma_{L/R} = \Sigma_{L/R}^r - \Sigma_{L/R}^a = 2i\,\text{Im}(\Sigma_{L/R}^r), \tag{8}$$

and can solve $G^<$ for a finite bias voltage as

$$\begin{aligned}
G^<(\mathbf{r},\mathbf{r}';\omega) = &-if_R(\omega)\int_{\partial\Omega_R}\int_{\partial\Omega_R} G^r(\mathbf{r},\mathbf{r}_R;\omega)\Gamma_R(\mathbf{r}_R,\mathbf{r}'_R;\omega) \\
&\times G^a(\mathbf{r}'_R,\mathbf{r}';\omega)d\mathbf{r}_R d\mathbf{r}'_R \\
&-if_L(\omega)\int_{\partial\Omega_L}\int_{\partial\Omega_L} G^r(\mathbf{r},\mathbf{r}_L;\omega)\Gamma_L(\mathbf{r}_L,\mathbf{r}'_L;\omega) \\
&\times G^a(\mathbf{r}'_L,\mathbf{r}';\omega)d\mathbf{r}_L d\mathbf{r}'_L.
\end{aligned} \tag{9}$$

The first and second terms correspond to electrons originating from the right and left leads, respectively. The electron density is calculated from

$$\rho(\mathbf{r}) = \frac{-1}{2\pi}\int_{-\infty}^{\infty}\text{Im}(G^<(\mathbf{r},\mathbf{r};\omega))d\omega \tag{10}$$

and the tunneling probability from

$$\begin{aligned}
T(\omega) = &\int_{\partial\Omega_L}\int_{\partial\Omega_L}\int_{\partial\Omega_R}\int_{\partial\Omega_R}\Gamma_L(\mathbf{r}_L,\mathbf{r}'_L;\omega)G^r(\mathbf{r}'_L,\mathbf{r}_R;\omega) \\
&\times \Gamma_R(\mathbf{r}_R,\mathbf{r}'_R;\omega)G^a(\mathbf{r}'_R,\mathbf{r}_L;\omega)d\mathbf{r}_L d\mathbf{r}'_L d\mathbf{r}_R d\mathbf{r}'_R, \tag{11}
\end{aligned}$$

from the values of the functions at the boundaries $\partial\Omega_{L/R}$. Finally, the current is determined as

$$I = \frac{1}{\pi}\int_{-\infty}^{\infty}T(\omega)(f_L(\omega)-f_R(\omega))d\omega. \tag{12}$$

We use the FEM in the numerical implementation. Therefore we first cast Eq. (4) in the variational form with open boundary conditions (for the derivation, see Ref. 25) as

$$\begin{aligned}
\int_{\Omega}&\left\{-\nabla\upsilon(\mathbf{r})\cdot\frac{1}{2}\nabla G^r(\mathbf{r},\mathbf{r}';\omega)\right. \\
&\left.+\upsilon(\mathbf{r})[\omega - V_{\text{eff}}(\mathbf{r})]G^r(\mathbf{r},\mathbf{r}';\omega)\right\}d\mathbf{r} \\
&-\langle\hat{\Sigma}_L G^r,\upsilon\rangle - \langle\hat{\Sigma}_R G^r,\upsilon\rangle = \upsilon(\mathbf{r}'),
\end{aligned} \tag{13}$$

where $\upsilon(\mathbf{r})$ is a smooth arbitrary function and the self-energy operators

$$\begin{aligned}
\langle\hat{\Sigma}_L G^r,\upsilon\rangle = &\int_{\partial\Omega_L}\int_{\partial\Omega_L}\frac{1}{4}G^r(\mathbf{r}'_L,\mathbf{r}';\omega) \\
&\times\frac{\partial^2 g_e(\mathbf{r}'_L,\mathbf{r}_L;\omega)}{\partial\mathbf{n}_L\,\partial\mathbf{r}'_L}\upsilon(\mathbf{r}_L)d\mathbf{r}'_L d\mathbf{r}_L. \tag{14}
\end{aligned}$$

Above, g_e is Green's function of the semi-infinite lead in the domain $\Omega_{L/R}$ with the zero-value condition on the boundary $\partial\Omega_{L/R}$. In our implementation the leads are described by a uniform positive background charge and therefore g_e can be calculated partly analytically. Thus our model means that the leads are of some kind of ideal generic metals. The important interface between the nanostructure, e.g., a molecule, and the actual metallic lead can be described accurately by including some lead metal atoms in the computational domain Ω. It is also possible to use fully atomistic leads by calculating numerically g_e for them.

Note that Eqs. (13) and (14) are analogous to those derivations of the open boundary conditions in which truncated matrices[1] are used. In the continuum limit these two forms give the same results. However, the weak form (13) is more natural in the FEM formulation and more suitable for theoretical purposes when analyzing nonlinear partial differential equations. It is also straightforward to use, and the error control is systematic. Note that this formulation can be used with any continuous basis set, not only with the FEM. In the context of basis-set methods, the weak formulation case is known as the Galerkin method. In practice Green's functions are approximated with respect to this basis so that

$$G^r(\mathbf{r},\mathbf{r}';\omega) \approx \sum_{i,j=1}^{N} g_{ij}(\omega)\phi_i(\mathbf{r})\phi_j(\mathbf{r}'). \tag{15}$$

The coefficients $g_{ij}(\omega)$ can be solved from (13) by choosing $\upsilon = \phi_k$ and evaluating the equations.

B. Finite-element p basis

In the FEM we partition the calculation volume to (in our case, tetrahedral) subdomains called elements, and the basis functions ϕ_i are constructed using globally continuous (but not necessarily continuously differentiable) piecewise polynomials with respect to the finite-element mesh. This gives both unique flexibility of the approximating functions as well as completeness of the basis with respect to almost any norm. Each basis function has a support that is concentrated to only a few neighboring elements. This makes the basis local and results in sparse system matrices.

There are several options on how to choose the finite-element basis, and one has to be careful in achieving accept-

able accuracy. The simplest basis is the linear one. It is easy to implement and works well, especially in systems with rapidly varying functions. A typical improvement to this basis is to use node-based higher-order elements. These elements converge faster to a smooth solution than the linear ones. However, practically only relatively low orders, two and three, can be used because of numerical stability problems.

In this work we have used the so-called hierarchical p elements. They also span higher-order polynomials, but the choice of the local basis ensures that stability problems do not appear. This is because the basis functions are chosen so that their derivatives are close to orthogonal in the L_2 norm. The hierarchical nature also makes it easy to change the order of the basis from element to element within the same mesh.

The actual FEM implementation consists of a reference element and reference basis that are mapped separately to each of the elements of the mesh. Our reference element is a tetrahedron with nodes at the coordinates 1: $(-1,0,0)$, 2: $(1,0,0)$, 3: $(0,\sqrt{3},0)$, and 4: $\left(0,\frac{1}{\sqrt{3}},2\sqrt{\frac{2}{3}}\right)$. One can easily show that there exists an affine map taking the reference element to any of the tetrahedron. The order of our basis is p, meaning that in each element polynomials of the order of p are employed. The basis is constructed hierarchically. First, there are four linear node basis functions inside the elements sharing a common node. In the reference element they are

$$L_1 = \frac{1}{2}\left(1 - \xi - \frac{\mu}{\sqrt{3}} - \frac{\zeta}{\sqrt{6}}\right),$$

$$L_2 = \frac{1}{2}\left(1 + \xi - \frac{\mu}{\sqrt{3}} - \frac{\zeta}{\sqrt{6}}\right),$$

$$L_3 = \frac{1}{\sqrt{3}}\left(\mu - \frac{\zeta}{\sqrt{8}}\right),$$

$$L_4 = \sqrt{\frac{3}{8}}\zeta, \tag{16}$$

where ξ, μ, and ζ are the Cartesian coordinates of the reference element. Second, for $p > 1$ we have $6(p-1)$ edge functions. For example, for the edge between nodes 1 and 2,

$$N_{i-1}^{(1,2)} = L_1 L_2 \varphi_i(L_2 - L_1) \quad i = 2,\ldots,p. \tag{17}$$

Here one usually sets

$$\varphi_i(\xi) = \frac{4\phi_i(\xi)}{1 - \xi^2},$$

$$\phi_i(\xi) = \sqrt{\frac{1}{2(2i-1)}}(P_i(\xi) - P_{i-2}(\xi)). \tag{18}$$

Above P_i is the Legendre polynomial of the order of i. Third, we have $2(p-1)(p-2)$ face functions. For example, for the face between nodes 1, 2, and 3 they are

$$N_{i,j}^{(1,2,3)} = L_1 L_2 L_3 P_i(L_2 - L_1) P_j(2L_3 - 1),$$

$$i,j = 0,\ldots,p-3, i+j = 0,\ldots,p-3. \tag{19}$$

Fourth, we have $\frac{1}{6}(p-1)(p-2)(p-3)$ bubble functions, which are supported only in a single element. These are

$$N_{i,j,k} = L_1 L_2 L_3 L_4 P_i(L_2 - L_1)$$
$$\times P_j(2L_3 - 1)P_k(2L_4 - 1),$$

$$i,j,k = 0,\ldots,p-4, i+j+k = 0,\ldots,p-4. \tag{20}$$

When p elements are used one must take care of the continuity of the basis. This is because, for example, the local basis function $N_3^{(1,2)}$ has an orientation on the boundary. The basis includes the function $\varphi(L_2-L_1)$, not $\varphi(L_1-L_2)$, which would be another possibility. This means that all the edges in the mesh have to have information about the direction. Otherwise there is very likely a continuity problem on some boundaries. In practice, for tetrahedral elements the orientation problem can be handled for arbitrary finite-element meshes using only two reference elements.[26]

The benefits of selecting the basis described above are rather clear. The polynomial basis is very easy to realize and has good approximating properties. For smooth solutions the p basis is known to give exponential convergence rates with respect to the number of basis functions used. In the DFT methods the theory is typically developed to the direction that the solutions are as smooth as possible. For example, pseudopotential operators are designed so that they produce as smooth an electron potential as possible. This is because the plane-wave basis set needs smooth solutions in order to work efficiently. On the other hand, in the case of nonsmooth solutions one can benefit from the piecewise nature of the FEM basis allowing one to approximate even singular solutions to some extent. Moreover, the finite-element mesh can be refined in regions where the solution changes rapidly. When modeling molecules there is also a lot of empty space in the calculation domain. It is then practical to use large elements in the empty space and smaller ones near atoms.

C. Linear algebra methods

The use of Green's-function method is computationally demanding in comparison to explicit wave-function methods. Since the main computational burden of our method is to find a subset of the coefficients of Green's function in question, a special consideration must be taken when choosing the methods of linear algebra to be used. The eigenvalue problems in explicit wave-function methods are typically solved by iterative methods. In our case it is better to use direct solvers, because a set of linear equations needs to be solved. We have opted for the frontal method widely used in the solution of sparse linear systems[27,28] and extremely suitable for finite-element matrices. The actual implementation is ME47 of the Harwell Subroutine Library[29] (HSL) (see Refs. 30–32 for other similar approaches). In the frontal method, one first finds a permutation of the sparse matrix aiming to minimize the fill-in resulting from the factorization process. Next, a LU-decomposition (or Cholesky decompo-

sition for symmetric problems) of the matrix A is found, and finally, two systems with triangular coefficient matrices, $Lz = b$ and $Ux = z$ (where $U = L^T$ for symmetric problems), are solved. To find all the required coefficients of the solution we must vary the right-hand side b of the system.

For a three-dimensional problem the size of the linear system can grow so large that the CPU time and memory requirements of different systems have to be addressed. The main question is how large systems can be calculated using these methods so that the calculation time for a single self-consistent iteration is not too large. Currently a system of several tens of thousands of unknowns can be solved in a commodity CPU cluster environment.

In detail, Green's-function method includes a computation of the elements for the inverse of a sparse matrix, so that the calculation time requirements increase relatively fast with the system size. A classical complexity result for the solution (and inversion) of a general $N \times N$ system with a direct method is $O(N^3)$. However, for sparse systems and modern frontal methods this bound is too pessimistic.[33] The CPU time requirement depends on the fill-in of the inversion problem. For very simple cases one can show that the key statistics of the problem, the number of non-zeros (nnz's) present in the factors L and U, satisfies $\text{nnz}(L) \sim \text{nnz}(U) = O(N \log(N))$.[33] Then the solution of each of the systems requires $O(\text{nnz}(L) + \text{nnz}(R))$ floating-point operations, and in the worst case we must solve these with N different right-hand sides, effectively giving us the inverse of the matrix A, so that the total cost is $O(N(\text{nnz}(L) + \text{nnz}(R)))$. However, in modern computer systems the complexity is not the only relevant measure since the performance may be highly nonlinear [see, e.g., Refs. 34 and 35 for an example on basic linear algebra subroutine (BLAS) tuning].

Another topic related directly to the performance of modern computer systems is the relation between processor power and memory bandwidth. This is especially true for the computation of Green's function, where the actual bottleneck is the lack of available memory bandwidth in commodity-based cluster systems used in calculations, not the floating-point performance of the processor itself.

It is likely that a better performance can be achieved by upgrading several parts of the algorithms. First, the current parallel solver is implemented using the message passing interface (MPI).[36] However, in symmetric multiprocessor (SMP) systems it is likely that well-designed OPENMP (Ref. 37) (or similar) parallelism would reduce the need for data transfer and thus increase performance. It would also decrease the memory requirements of the problems. Second, at the moment the solution of Green's function is computed by varying one vector on the right-hand side at a time. A better performance could be obtained if the equations could be solved for multiple right-hand sides at a time, allowing the use of BLAS3 routines. Finally, it is likely that computations would benefit from a computer system having a larger memory bandwidth than our present commodity-based one.
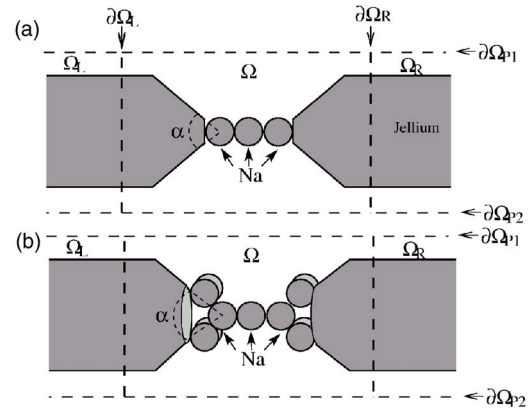

FIG. 2. Two models of the Na-atom chain. (a) Na atoms are directly connected to the cone-shaped leads. (b) there are four Na atoms shown as squares at the tips of the leads.

III. EXAMPLE SYSTEMS

A. Atomic wire

Using the atomic force microscope or the mechanically controlled break junction technique, a chain of atoms can be made from certain metals.[38] It has been observed that the conductances of atom chains vary as a function of the number of atoms in the chain.[39] The conductances of these systems have been studied also theoretically in several works. In order to benchmark our results against other calculations, we use Na-atom chains as test systems. They have been simulated in several previous studies[40–43] using different models. According to these calculations the conductances of the wires show even-odd oscillation as a function of the number of atoms in the wire.

In our setup, the Na-atom chain is located between two leads, with the lead shape defined by a 70° cone angle (see Fig. 2). We consider two different connections of the atom chain to the electrodes. In Fig. 2(a) we just have three Na atoms between the jellium leads. This resembles closely the system used in Ref. 41. In Fig. 2(b), there are four Na atoms at the tips of the leads in a square form. This makes the connection between the atom chain and the leads more realistic. This kind of structure is modeled also in Ref. 42.

The conductances as a function of the number of chain atoms for systems A and B are shown in Fig. 3. In the Na-atom chain, electrons have only one conducting mode so that the conductance can be one conductance quantum $2e^2/h$ at maximum. Both systems A and B exhibit conductance oscillations as a function of the number of atoms. These oscillations arise from resonance states in the atom chain. Depending on the position of the resonances relative to the Fermi level, the conductance has either a maximum or minimum value, so that the maxima and minima correspond to approximately half and fully occupied resonance states, respectively. The oscillation is within the range of $(0.9–1.0) \times 2e^2/h$ for system A and $(0.6–1.0) \times 2e^2/h$ for system B. The difference between the oscillation amplitudes is due to the different strengths of the connection of the chains to the leads. System B has a weaker coupling to the leads than system A. Weak connections make the resonances also sharper, as is seen in the tunneling probability in Fig. 4. In
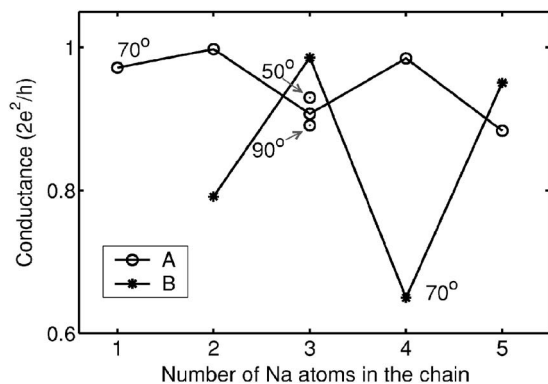
FIG. 3. Conductance through the Na-atom chain as a function of the number of Na atoms in the chain. The results for systems A and B (see Fig. 2) are denoted by circles and stars, respectively. For system A with three Na atoms, results corresponding to 50° and 90° cone angles α are also given.

contrast to Ref. 44, we do not see a strong lead-shape dependence in the conductance. The widening of the cone angle lowers the conductance as the edges of the wire become sharper.

The electron tunneling probabilities through chains of three- and four-atom systems A and B are shown in Fig. 4. The probability function $T(\omega)$ is defined in Eq. (11). The conductance of the system in the zero-bias limit can be read at the Fermi level. Here, as well as in Fig. 3, we see that the conductance oscillations for systems A and B are in a different phase. This is because in system B the atom chain is effectively shorter than in A, as the first and the last chain atoms are partly inside the square of the four Na atoms.

When we compare the conductance oscillations of system A (see Fig. 3) to those in Ref. 41 obtained by using semi-infinite jellium leads with planar surfaces ($\alpha=180°$), we see that the even-odd oscillations in the conductance are in the same phase. In the case of system B we can directly compare the tunneling probability of Fig. 4 with those in Ref. 42, where the atom chain is connected also through a square of four Na atoms to jellium. The phase and amplitude
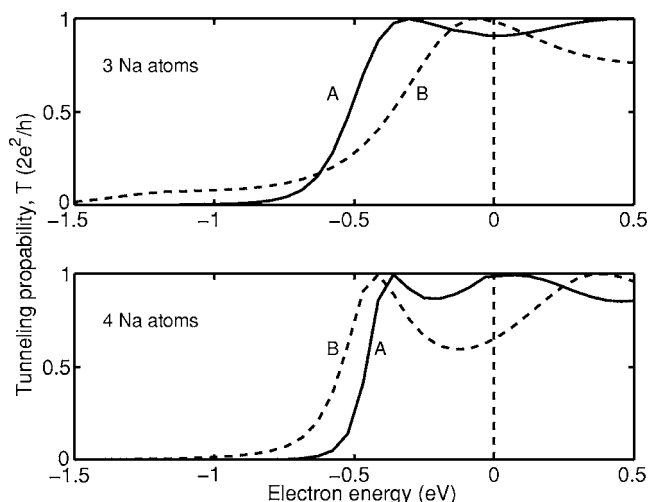

FIG. 4. Tunneling probability from Eq. (11) for three- and four-atom chains between two semi-infinite jellium leads. The solid and dashed lines correspond to systems A and B shown in Fig. 2, respectively. The Fermi level is marked by dashed vertical lines. The cone angle $\alpha=70°$.

of the conductance oscillations of these results are in good agreement with our values in Fig. 3. Now that we have satisfied ourselves that the method provides a good model for electron transport, we can consider a more interesting and demanding example.

B. Thin insulating layer

The general increase in the performance of microelectronic devices in the past few decades has been made possible by continuous transistor scaling based on a reduction in the thickness of the gate dielectric in typical metal-oxide-semiconductor field-effect transistors (MOSFETs). At present the process has reached a bottleneck, as further reduction leads to a large increase in leakage current due to direct tunneling across the thin silicon dioxide (SiO_2) layer. Several possible approaches to resolve this are being considered,[45] but retaining conventional MOSFET design remains an economically attractive choice, and a leading option is just to replace SiO_2 with another oxide of higher dielectric constant (high k). A high-k oxide would provide a higher effective capacitance to a comparable SiO_2 layer, hence allowing thicker layers to be used to reduce losses due to tunneling. The specific choice of oxide is determined by a set of requirements[46] based on both the intrinsic properties of the grown oxide and its integration into the fabrication process, and at present hafnium oxide (HfO_2) remains a leading candidate.

In order to study the transport properties of thin HfO_2 films, we have simulated the growth of the oxide on a silicon surface via first-principles molecular dynamics.[47] Here we consider three model interfaces: (i) a nonstoichiometric oxide interface (C), which is basically metallic due to Hf–Hf and Hf–Si bonds across the interface; (ii) a stoichiometric oxide interface (D), which has a localized state in the band gap due to a few Hf–Hf bonds; and (iii) a more idealistic interface (E), which remains insulating if no defects are present. The last model is based on the interface used in Ref. 48, but slightly reduced in size to make it computationally manageable.[49] These models were calculated with periodic boundary conditions with k points on the boundaries $\partial\Omega_{P1/P2/P3/P4}$. The effective potentials have been calculated for systems C and D using the gamma point, and for system E, four k points. All the tunneling probabilities $T(\omega)$ are calculated using four k points, which were enough to converge the probabilities to a good accuracy.

As shown in Fig. 1, the interface models are positioned between two leads. The charge density in the leads is chosen so that in the right lead $r_s=2$ [electron density $n_e=3/(4\pi r_s^3)$], representing a metal, and in the left one $r_s=3.1$, representing doped silicon as in a standard MOSFET design.

The tunneling profiles of the systems are shown in Fig. 5. Here it is seen that systems C and D show clearly metallic behavior, with a large tunneling probability at the Fermi energy. Although, in principle, the stoichiometric interface (D) has a much lower density of metallic bonds, it is clear that both interfaces C and D around two channels dominate the
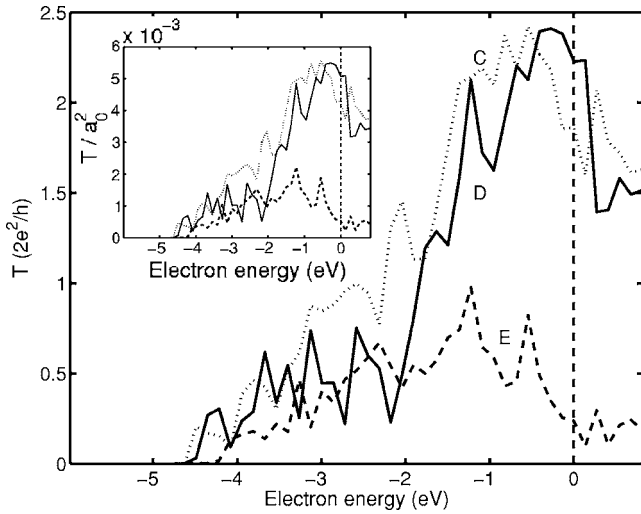
FIG. 5. Tunneling probability $T(\omega)$ through thin HfO_2 layers. Results for system C (dotted line): nonstoichiometric interface, system D (solid line): stoichiometric interface, and system E: ideal interface are shown. The inset shows $T(\omega)$ normalized with the conducting area, enabling the comparison of actual insulating properties of different systems.
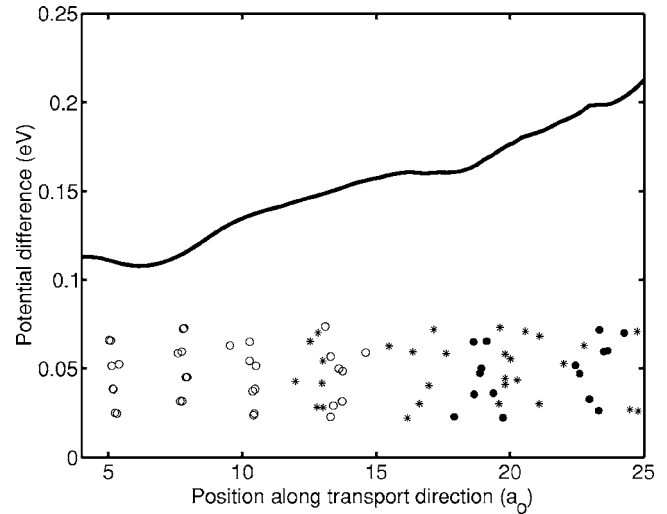

FIG. 6. Change of the average effective potential in interface D when a 0.25 V bias voltage is applied over the HfO_2 layer. In every position along the transport direction, the effective potential is averaged over the perpendicular coordinates. Atom positions indicated by open circles are silicon, filled circles and stars are hafnium and oxygen atoms, respectively.

transport. The localized defect state in the band gap of system D plays an equivalent role in transport to the metallic bonds in interface C.

As expected, the tunneling probability for the more ideal interface E is an order of magnitude smaller at the Fermi energy than those for interfaces C and D. Yet we also see that it remains significant; this is largely due to the structure of the interface itself.[48] Although bulk HfO_2 is a wide-band-gap insulator, at the interface it exists as almost tetragonal $HfSiO_4$, and the effective band gap is actually smaller than that of bulk silicon below the interface. This means that there is a negative conduction band offset between silicon and HfO_2, and no real barrier for leakage. Although some of this is caused by the underestimation of the band gap in the DFT, this also reduces the silicon band gap (although the effect is not systematic).

The poor performance of interface D can also be seen in its capacity for dropping the potential. Figure 6 shows the potential change for the 0.25 V applied bias voltage. The potential drop across HfO_2 is less than 0.05 eV, demonstrating that the oxide hardly perturbs the electron flow from the right lead. The potential drops fastest at the right-hand side of the HfO_2 layer where pure HfO_2 exists, and much more slowly in the thin layer of SiO_2 formed due to diffusion of oxygen. Electrons can rather freely flow through the oxide layer from the right lead to the left lead. Because the electron density is assumed to be unperturbed in the leads beyond the boundaries $\partial\Omega_L/R$, there will be a surplus (deficiency) of electrons in the left (right) lead inside the computational volume. These net charges account for the bias voltage drop outside the oxide (and Si) layers.

In the rigid-band approximation (used, for example, in Ref. 48) it is assumed that the shape of the tunneling probability stays constant and is only shifted in energy so that $T(\omega, V_{bias}) = T(\omega + \eta V_{bias})$, where η is the ratio of potential drop at the other end of the nanostructure to the total drop over the nanostructure. In Fig. 7 we have studied how well

this approximation works for interface D. The curves are plotted so that the zero-bias Fermi level is in the middle of the left and right Fermi levels of the biased interface. This corresponds to the symmetric case with $\eta = 0.5$. We see that the tunneling probability curves roughly coincide. This indicates that the potential drops symmetrically over the nanostructure, and the rigid-band approximation gives a rather reasonable result.

The above results show, in agreement with previous calculations,[48] that tunneling through a more ideal, insulating interface is still significant due to a negative band offset with silicon. Since the only HfO_2 interfaces providing significant band offsets to silicon were built very idealistically (i.e., assuming no significant atom migration nor interfacial SiO_2
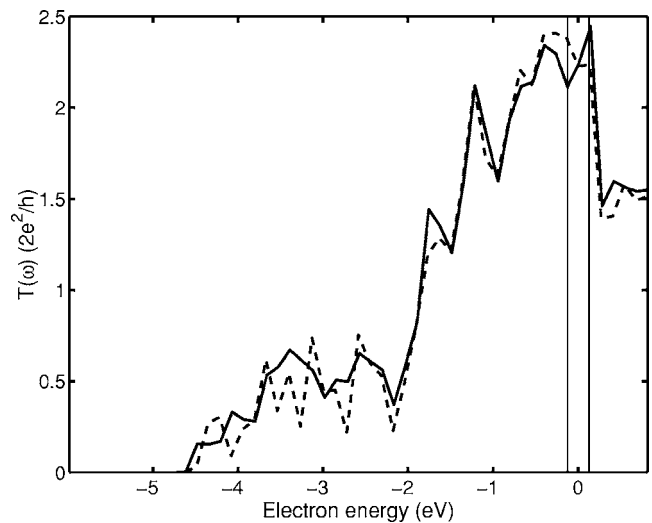

FIG. 7. Tunneling probability through a thin HfO_2 layer. Results for interface D with 0.25 V bias (solid line) and zero bias (broken line) voltages are shown. The vertical lines show the positions of the Fermi levels in the left and right leads for the biased system. Between them is located the so-called conductance window [see Eq. (12)]. The Fermi level of the nonbiased system is in the middle of these lines.

growth),[50,51] this indicates that fabricating a *good* interface directly between silicon and HfO$_2$ is very difficult. A more viable alternative may be to sacrifice somewhat the dielectric constant and grow HfO$_2$ onto a preexisting SiO$_2$ layer. These possibilities will be explored in more detail in a further work.[47]

IV. CONCLUSIONS

In this paper we present a finite-element implementation of the nonequilibrium Green's-function method which is combined to the density-functional theory. Although Green's-function method is computationally demanding, we demonstrate that by using hierarchical p elements, large, physically relevant systems become tractable. More importantly, our method offers a much more rigorous control of accuracy than is usually possible in transport calculations.

We demonstrate the functionality of our implementation with two kinds of systems, the sodium-atom chain wire and the silicon-HfO$_2$ interface. For the atom chain, we show that the method reproduces the previous results of other Green's-function transport methods. This gives us confidence to apply it to the more complex system: a thin layer of hafnium oxide on a silicon substrate. Here we show that the transport properties are an even more sensitive indicator of the role of defects than the electronic structure. Comparison of stoichiometric and nonstoichiometric HfO$_2$ oxide layers demonstrates that even one or two defects in a stoichiometric interface can result in tunneling comparable to that of a fully metallic nonstoichiometric interface.

[1] S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, Cambridge, 1995).
[2] A. Nitzan, Annu. Rev. Phys. Chem. **52**, 681 (2001).
[3] A. Nitzan and M. A. Ratner, Science **30**, 1384 (2003).
[4] Y. Luo, C. P. Collier, J. O. Jeppesen *et al.*, ChemPhysChem **3**, 519 (2002).
[5] Y.-H. Kim, S. S. Jang, Y. H. Jang, and W. A. Goddard, Phys. Rev. Lett. **94**, 156801 (2005).
[6] K. Burke, R. Car, and R. Gebauer, Phys. Rev. Lett. **94**, 146803 (2005).
[7] K. S. Thygesen and K. W. Jacobsen, Phys. Rev. B **72**, 033401 (2005).
[8] P. Sautet and C. Joachim, Phys. Rev. B **38**, 12238 (1988).
[9] L. Chico, L. X. Benedict, S. G. Louie, and M. L. Cohen, Phys. Rev. B **54**, 2600 (1996).
[10] J. Taylor, H. Guo, and J. Wang, Phys. Rev. B **63**, 245407 (2001).
[11] M. Brandbyge, J. L. Mozos, P. Ordejón, J. L. Taylor, and K. Stokbro, Phys. Rev. B **65**, 165401 (2002).
[12] M. B. Nardelli, J.-L. Fattebert, and J. Bernholc, Phys. Rev. B **64**, 245423 (2001).
[13] K. S. Thygesen, M. V. Bollinger, and K. W. Jacobsen, Phys. Rev. B **67**, 115404 (2003).
[14] D. Wortmann, H. Ishida, and S. Blugel, Phys. Rev. B **66**, 075113 (2002).
[15] A. Calzolari, N. Marzari, I. Souza, and M. B. Nardelli, Phys. Rev. B **69**, 035108 (2004).
[16] P. A. Khomyakov and G. Brocks, Phys. Rev. B **70**, 195402 (2004).
[17] E. Polizzi and A. N. Ben, J. Comput. Phys. **202**, 150 (2005).
[18] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).
[19] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[20] J. Arponen, P. Hautojärvi, R. Nieminen, and E. Pajanne, J. Phys. F: Met. Phys. **3**, 2092 (1973).
[21] M. Fuchs and M. Scheffler, Comput. Phys. Commun. **119**, 67 (1999).
[22] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).
[23] D. R. Bowler and M. J. Gillan, Chem. Phys. Lett. **325**, 473 (2000).
[24] Y. Xue, S. Datta, and M. A. Ratner, Chem. Phys. **281**, 151 (2002).
[25] P. Havu, V. Havu, M. J. Puska, and R. M. Nieminen, Phys. Rev. B **69**, 115325 (2004).
[26] M. Ainsworth and J. Coyle, Int. J. Numer. Methods Eng. **58**, 2103 (2003).
[27] J. K. Reid and I. S. Duff, ACM Trans. Math. Softw. **9**, 302 (1983).
[28] I. S. Duff, Comput. Phys. Commun. **97**, 45 (1996).
[29] The Harwell Subroutine Library, see http://www.cse.clrc.ac.uk/nag/hsl/
[30] A. Gupta, ACM Trans. Math. Softw. **28**, 321 (2002).
[31] T. A. Davis, ACM Trans. Math. Softw. **30**, 353 (2004).
[32] J. W. Demmel, S. C. Eisenstat, J. R. Gilbert, X. S. Lia, and J. W. H. Li, SIAM J. Matrix Anal. Appl. **20**, 720 (1999).
[33] A. George and J. W.-H. Liu, *Computer Solution of Large Sparse Positive Definite Systems* (Prentice-Hall, Englewood Cliffs, NJ, 1981).
[34] *Automatically Tuned Linear Algebra Software (ATLAS)*, http://math-atlas.sourceforge.net/
[35] K. Goto and R. van de Geijn, Department of Computer Sciences, The University of Texas at Austin, Report No. TR-2002-55, 2002 (unpublished).
[36] *MPI*, http://www.mpi-forum.org/
[37] *OPEN MP*, http://www.openmp.org/
[38] N. Agrait, A. L. Yeyati, and J. M. van Ruitenbeek, Phys. Rep. **377**, 81 (2003).
[39] R. H. M. Smit, C. Untiedt, G. Rubio-Bollinger, R. C. Segers, and J. M. van Ruitenbeek, Phys. Rev. Lett. **91**, 076805 (2003).
[40] H.-S. Sim, H.-W. Lee, and K. J. Chang, Phys. Rev. Lett. **87**, 096803 (2001).
[41] N. D. Lang, Phys. Rev. Lett. **79**, 1357 (1997).
[42] S. Tsukamoto and K. Hirose, Phys. Rev. B **66**, 161402(R) (2002).
[43] Y.-J. Lee, M. Brandbyge, M. J. Puska, J. Taylor, K. Stokbro, and R. M. Nieminen, Phys. Rev. B **69**, 125409 (2004).
[44] P. Havu, T. Torsti, M. J. Puska, and R. M. Nieminen, Phys. Rev. B **66**, 075401 (2002).
[45] A. I. Kingon, J. P. Maria, and S. K. Streiffer, Nature (London) **406**, 1032 (2000).
[46] H. R. Huff, A. Hou, C. Lim *et al.*, Microelectron. Eng. **69**, 152 (2003).
[47] M. H. Hakala and A. S. Foster, J. Appl. Phys. (to be published).
[48] L. R. C. Fonseca, A. A. Demkov, and A. Knizhnik, Phys. Status Solidi B **239**, 48 (2003).
[49] J. L. Gavartin, L. Fonseca, G. Bersuker, and A. L. Shluger, Microelectron. Eng. **84**, 412 (2005).
[50] V. Fiorentini and G. Gulleri, Phys. Rev. Lett. **89**, 266101 (2002).
[51] P. W. Peacock and J. Robertson, Phys. Rev. Lett. **92**, 057601 (2004).