The Non-Equilibrium Green's Function (NEGF) Formalism: An Elementary Introduction

Supriyo Datta
Email: datta@purdue.edu , Tel: 765-49-43511, Fax: 765-49-42706
School of Electrical and Computer Engineering, Purdue University,
West Lafayette, IN 47907

ABSTRACT

The non-equilibrium Green's function (NEGF) formalism provides a sound conceptual basis for the development of quantitative models for quantum transport (see for example, (1-4)). The purpose of this talk is to present a simple intuitive discussion of the NEGF equations illustrating the basic physics (5).

INTRODUCTION

Any device simulation program performs a self-consistent solution of a transport equation and a "Poisson" equation (Fig.1). The transport equation calculates the electron density, n(**r**) and the current, I for a given potential profile U(**r**), while the "Poisson" equation calculates the effective potential U(**r**) that an electron feels due to the presence of the other electrons. The two calculations are iterated till n(**r**) and U(**r**) converge to a self-consistent value. A quantum transport simulator also performs a similar iterative solution of a transport equation and a Poisson-like equation. Let us talk about these one by one.

TRANSPORT EQUATION

Rate equation for one discrete level: Consider first a really small device with just one energy level ε in the energy range of interest, connected to a source and a drain contact (Fig.2). What is the number of electrons, N in our device? The answer is clear if everything is in equilibrium with a common Fermi energy E_f, set by the work function of the source and drain contacts.

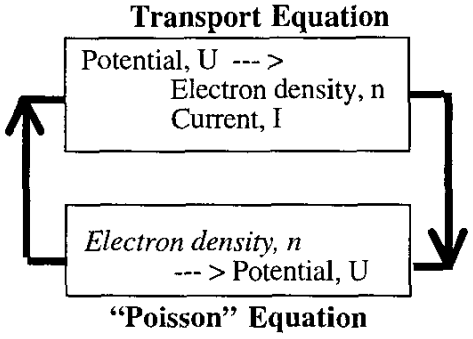

Fig.1. Any device simulation program performs an iterative self-consistent solution of a transport equation and a Poisson-like equation.

However, once we apply a drain bias V, the Fermi energies in the source and drain contacts, denoted by μ_1 and μ_2 will separate as follows (see Fig.2):

$$\mu_1 = E_f + (qV/2) \quad \text{and} \quad \mu_2 = E_f - (qV/2) \tag{1}$$

giving rise to two distinct Fermi functions for the two contacts. If the device were in equilibrium with the source, the number of electrons would equal f_1, but if the device were in equilibrium with the drain, the number of electrons would equal f_2, where

$$f_{1,2}(\varepsilon) = \frac{1}{\exp\left[\left(\varepsilon - \mu_{1,2}\right)/k_B T\right]} \tag{2}$$

The actual number of electrons N will clearly be intermediate between f_1 and f_2 and can be determined by writing simple rate equations for the currents $I_{1,2}$ crossing the source and drain interfaces (Fig.2):

$$I_1 = \frac{q\gamma_1}{\hbar}[f_1 - N] \quad \text{and} \quad I_2 = \frac{q\gamma_2}{\hbar}[N - f_2] \tag{3}$$

where the constants $(\gamma_1/\hbar)$ and $(\gamma_2/\hbar)$ represent the rates (per second) at which an electron inside the device will escape into the source and drain respectively.

Setting $I_1 = I_2 \equiv I$, we obtain the steady-state number of electrons N and the current I:

$$N = \frac{\gamma_1}{\gamma_1 + \gamma_2} f_1(\varepsilon) + \frac{\gamma_2}{\gamma_1 + \gamma_2} f_2(\varepsilon) \tag{4a}$$

$$I = \frac{q}{\hbar}\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}\left[f_1(\varepsilon) - f_2(\varepsilon)\right] \tag{4b}$$

This elementary derivation illustrates the essential physics of current flow through a small conductor attached to two reservoirs that strive to maintain it at two different levels of occupation f_1 and f_2. The actual occupation is intermediate between the two (Eq.(4a)) and one reservoir keeps pumping in electrons trying to increase the number while the other keeps emptying it trying to lower the number. The overall effect is a continuous flow of electrons from one reservoir to the other, leading to a net current in the external circuit (Eq.(4b)).

0-7803-7462-2/02/$17.00 ©2002 IEEE

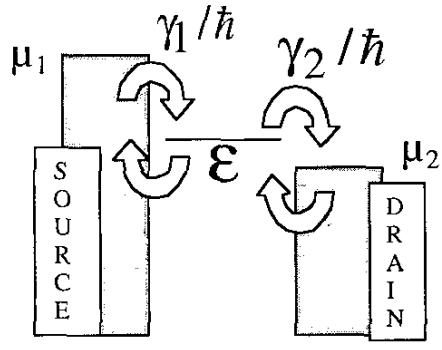

Fig.2. Derivation of the transport equation for one discrete level.

Broadening: The coupling to the source and drain contacts broadens the discrete level into a distribution

$$D(E) = \frac{\gamma/2\pi}{(E-\varepsilon-\Delta)^2+(\gamma/2)^2}$$

having a linewidth of γ along with a possible shift in the level from ε to $\varepsilon+\Delta$, where $\gamma=\gamma_1+\gamma_2$, $\Delta=\Delta_1+\Delta_2$. We can account for this broadening by modifying Eqs.(4a,b) to include an integral over all energies, weighted by the distribution $D(E)$:

$$N = \int_{-\infty}^{+\infty} dE\, D(E)\left[\frac{\gamma_1}{\gamma_1+\gamma_2} f_1(E) + \frac{\gamma_2}{\gamma_1+\gamma_2} f_2(E)\right] \quad (5a)$$

$$I = \frac{q}{\hbar}\int_{-\infty}^{+\infty} dE\, D(E)\,\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}\left[f_1(E)-f_2(E)\right] \quad (5b)$$

Using straightforward algebra, Eqs.(5a,b) can be rewritten as

$$N = \int_{-\infty}^{+\infty}\frac{dE}{2\pi}\left[A_1(E)\, f_1(E) + A_2(E)\, f_2(E)\right] \quad (6a)$$

$$I = \frac{q}{h}\int_{-\infty}^{+\infty} dE\, \overline{T}(E)\left[f_1(E)-f_2(E)\right] \quad (6b)$$

where $A_1 = G\,\gamma_1\,G^+$, $A_2 = G\,\gamma_2\,G^+$, $\overline{T} = \gamma_1\,G\,\gamma_2\,G^+$ (7)

$$G = \left[E-\varepsilon-\sigma_1-\sigma_2\right]^{-1},\ \sigma_{1,2} \equiv \Delta_{1,2} - i\,\gamma_{1,2}/2 \quad (8)$$

The "real" thing : So far we have assumed that the device has just one energy level with energy ε. What does this have to do with real devices (like the one shown in Fig.3) which surely have multiple energy levels in the energy range of interest? The answer is that any device, in general , is described by a Hamiltonian ***matrix*** [H] whose eigenvalues tell us the allowed energy levels. For example if we describe the device using an effective mass Hamiltonian $H = -\left(\hbar^2/2m^*\right)\nabla^2 + U(r)$, we could represent it with a (NxN) matrix by choosing a discrete lattice with N points and using the method of finite differences (see for example, (6)). This corresponds to using a (discretized) real space basis. More generally we could use valence atomic orbitals like sp^3s^* as a basis and write down a semi-empirical Hamiltonian, or go further and include the core atomic orbitals as well, with an ab intio approach.

Similarly, once we have chosen a basis or a representation, we can define self-energy ***matrices*** $[\Sigma_{1,2}]$ that describe the broadening and shift of the energy levels due to the coupling to the source and drain. The appropriate NEGF equations are obtained from Eqs.(6a,b) simply by replacing the scalar quantities like ε and $\sigma_{1,2}$ with the corresponding matrices [H] and $[\Sigma_{1,2}]$ (that is why we rewrote Eqs.(5a,b) in this form!). This yields

$$G = \left[EI-H-\Sigma_1-\Sigma_2\right]^{-1},\ \Gamma_{1,2} = i\left[\Sigma_{1,2}-\Sigma_{1,2}^+\right] \quad (9)$$

$$A_1(E) = G\,\Gamma_1\,G^+,\quad A_2(E) = G\,\Gamma_2\,G^+ \quad (10)$$

where I is an identity matrix of the same size as the rest. The number of electrons N (Eq.(6a)) is replaced by the density matrix given by an analogous quantity:

$$[\rho] = \int_{-\infty}^{+\infty}\frac{dE}{2\pi}\left\{\left[A_1(E)\right] f_1(E) + \left[A_2(E)\right] f_2(E)\right\} \quad (11)$$

The current is still given by Eq.(6b) if we define the transmission as the trace of the analogous matrix quantity

$$\overline{T}(E) = \text{Trace}\left[\Gamma_1\, G\,\Gamma_2\, G^+\right] \quad (12)$$

Both Eqs.(10) and (6b) should be multiplied by 2 for spin degeneracy, unless spin is explicitly accounted for in the [H] matrix itself.

I do not wish to imply that what we have done here represents a "derivation" of the NEGF equations. What we have derived carefully is the one-level scalar version (Eqs.(5-8)). The multi-level matrix version (Eqs.(6b), (9)-(12)) follows from it only if all the matrices are diagonal. But in general one cannot diagonalize both the Hamiltonian [H] and the self-energy matrices $[\Sigma_{1,2}]$ simultaneously and a more careful treatment is called for (7). The real value of this discussion is that gives us an intuitive feeling for the meanings of the quantities appearing in the NEGF equations.

The NEGF equations presented here do not reflect the effect of incoherent scattering processes (such as electron-phonon interaction) inside the device which become increasingly important as the device gets longer. In an approximate qualitative sense, scattering processes are like additional floating contacts that extract electrons from the device. However, unlike the source and drain contacts, they reinject the electrons so as not to draw any net current (8).

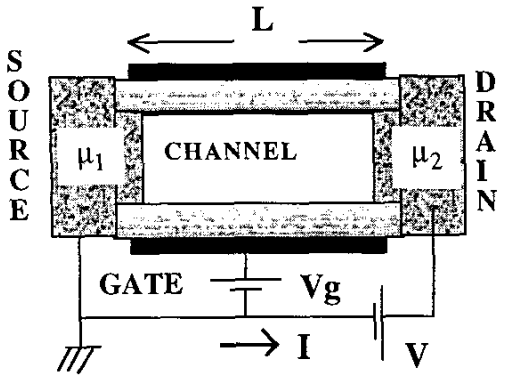

Fig.3. (a) Sketch of a generic nanoscale FET. The channel could be silicon or some non-silicon alternative like carbon nanotubes. (b) The Hamiltonian matrix [H] describes the energy levels of the channel, while the self-energy matrices $[\Sigma_1]$ and $[\Sigma_2]$ describe the effect of coupling it to the source and drain contacts respectively.

Scattering processes can be included in the NEGF formalism by defining additional self-energy matrices like the ones we have defined for the source and drain contacts but the details are more complicated – at least operationally, if not conceptually (see Section 5 of Ref.(6)). Indeed the NEGF equations were first developed to describe quantum transport in bulk solids with the scattering processes described by self-energy matrices; the use of self-energy matrices to describe the source and drain contacts came later.

Density matrix: The diagonal elements of the density matrix $[\rho]$ are interpreted as the number of electrons occupying the corresponding basis orbital. If we are using a real space lattice as our basis, then the diagonal elements tell us the electron density in real space, n(**r**): $n(r) = \rho(r,r)$. One can go from one representation to another through an appropriate unitary transformation. The total number of electrons is obtained by adding up all the diagonal elements (Trace$[\rho]$) which is the same in any representation. The off-diagonal elements of $[\rho]$ do not have a simple physical interpretation, but it is clear that they are needed to make sure that $[\rho]$ transform correctly from one representation to another. For example, if we only knew the electron density n(r) in real space we would not be able to calculate n(k), but if we knew the full $\rho(r,r')$ we could perform a unitary transformation to $\rho(k,k')$, obtain n(k) from the diagonal elements and even calculate the current from it (instead of using Eq.(6)).

"POISSON" EQUATION

Finally let us talk about the second half of the self-consistent procedure depicted in Fig.1. We write the overall Hamiltonian as $[H] = [H_0] + [U_{scf}]$, where the self-consistent field $[U_{scf}]$ is a functional of the density matrix $[\rho]$ that is chosen so as to account for electron-electron interactions. The lowest order approximation to the self-consistent potential is given by the Hartree potential U_H obtained from the Poisson equation used in standard device simulation programs:

$$\vec{\nabla}.\left[\varepsilon\vec{\nabla}U_H(r)\right] = -q^2\left[n(r) - n_0(r)\right] \qquad (13)$$

If we are using a discrete real space basis, then the corresponding $[U_{scf}]$ is diagonal with the diagonal elements equal to U_H: $U_{scf}(r,r) = U_H(r)$. In general for a given set of basis functions, the matrix elements for $[U_{scf}]$ have to be determined following the usual quantum mechanical prescription. In Eq.(13) $n_0(r)$ represents the reference electron density corresponding to the unperturbed Hamiltonian $[H_0]$. Also, the dielectric constant ε should only include those effects that are not included in the Hamiltonian [H]. For example, if [H] includes the core electrons, then their contribution to ε should be excluded.

It is generally recognized that this Hartree approximation overestimates the electron-electron repulsion. The interaction energy of a collection of electrons cannot really be expressed solely in terms of the electron density n(r) which tells us the chances of finding an electron at r. We need the two-electron distribution $n_2(r,r')$ which tells us the chances of finding one electron at r and another at r'. If the motion of individual electrons were uncorrelated then $n_2(r,r') = n(r)\,n(r')$. But the electrons correlate their motion so as to lower their interaction energy making $n_2(r,r') < n(r)\,n(r')$. Physically this means that we can pack electrons more densely than we would otherwise expect, making the capacitance larger. Mathematically, this means that the "correct" $[U_{scf}]$ in Eq.(13) is somewhat smaller than what the Hartree approximation suggests.

We put "correct" within quotes as a reminder that the self-consistent field approach is essentially a one-electron approximation to a many-electron problem with no exact solutions, and what is "correct" for one property (say ground state structure) may not be "correct" for another (say current flow or optical absorption). There are different prescriptions commonly used to go beyond the Hartree approximation like Hartree-Fock (HF) or density functional theory (DFT), and the best choice can only be established through careful comparison with experiment.

Coulomb Blockade: Let me make one final point that is best explained by going back to the one-level model (Eqs.(4a,b)) and considering the equilibrium situation with $\mu_1 = \mu_2 = E_f$ so that $f_1 = f_2 \equiv f_0$ (see Eqs.(1),(2)) and from Eq.(4a), the number of electrons, $N_0 = 2f_0(\varepsilon)$, with the additional factor of two included to account for spin degeneracy. How does this number change with the gate voltage V_g (see Fig.3)? We could model this by letting the level ε float up or down by $-qV_g$ plus the self-consistent potential $U = U_0(N - N_0)$ (cf. Eq.(13)): $N = 2f_0(\varepsilon - qV_g + U)$.

The factor of two for spin-degeneracy can be taken more explicitly into account by writing separate equations for the up-spin ($\uparrow$) and down-spin ($\downarrow$) levels:

$$N_\uparrow = f_0(\varepsilon - qV_g + U_\uparrow) \ , \ N_\downarrow = f_0(\varepsilon - qV_g + U_\downarrow) \quad (14)$$

with $U_\uparrow = U_\downarrow = U_0\,(N_\uparrow + N_\downarrow - N_0)$ (15)

where U_0 is the single-electron charging energy.

If we solve Eqs.(14) and (15) iteratively (10) we get the result shown in Fig.4 (solid line with circles): as the gate voltage becomes more positive, the number of electrons increases smoothly from 0 to 2. But now suppose we modify Eq.(15) to recognize the fact that there is no self-interaction between electrons. An up-spin electron feels a potential due to the down-spin electron but not due to itself and vice versa:

$$U_\uparrow = U_0\,(2N_\downarrow - N_0) \ , \ U_\downarrow = U_0\,(2N_\uparrow - N_0) \quad (16)$$

If we solve Eqs.(14) and (16) self-consistently, the result we get is different as shown by the solid line in in Fig.4: the number of electrons changes abruptly from 0 to 1 and later from 1 to 2. This is an example of the single-electron charging effect or Coulomb blockade as it is often called.

If we look carefully at the solution we would find that when the number of electrons is 1, the two spin levels are NOT degenerate. One of the spin levels has a lower energy and is occupied while the other has a higher energy and is empty. Which spin has a lower energy depends on whether our starting guess for $U_\uparrow$ is smaller or larger than $U_\downarrow$. Indeed if we start with $U_\uparrow = U_\downarrow$, we would iterate to the same spin-degenerate solution that we obtained from Eqs.(14),(15). But an initial guess with $U_\uparrow \neq U_\downarrow$ will converge to a spin non-degenerate solution like the one shown. This is indeed a rather profound result showing that even with no magnetic field present, electron-electron interactions can spontaneously break the spin-symmetry and give rise to a little "magnet" whose spin levels are split.

When do we need to worry about this single-electron charging effect? Only when the single electron charging energy U_0 exceeds the thermal energy and the energy broadening $\gamma_{1,2}$. Roughly speaking, if an electronic wavefunction is localized in a sphere of diameter of R in a medium with a dielectric constant ε, then the corresponding charging energy is given by $U_0 = q^2/4\pi\varepsilon R$. Charging energies of several meV's require wavefunctions localized in very small regions of space (say R < 10 nm). This can arise in semiconductor devices in the context of say impurity levels or floating gate memory devices. In such cases a spin degenerate prescription for the self-consistent field could miss important physics. It is then necessary at least to use a spin non-degenerate prescription and evaluate the need for more elaborate schemes (9).

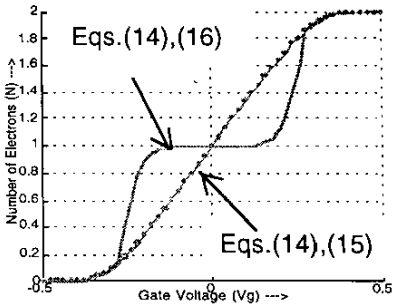

Fig.4. Number of electrons, $N = N_\uparrow + N_\downarrow$ as a function of the gate voltage calculated from Eq. (14) and Eq.(15) OR Eq.(16) assuming $E_f = \varepsilon$ and $U_0 = 0.25$ eV.

CONCLUDING REMARKS

Quantum transport is an active field where exciting new experiments are being reported utilizing many different material systems including semiconductors, carbon nanotubes and organic molecules. We have presented an introductory tutorial description of the NEGF formalism which provides a solid unified framework for developing and refining transport models as important new effects are identified.

This article is dedicated to the late Prof. Phil Bagwell with whom I had many illuminating discussions on quantum transport. This work is supported by the National Science Foundation, the Army Research Office and the Semiconductor Research Corporation.

REFERENCES

(1) http://www-hpc.jpl.nasa.gov/PEP/gekco/nemo/nemo.html.
(2) A. Svizhenko, M. Anantram, T. Govindan, B. Biegel, and R. Venugopal, "Nano-transistor modeling: two dimensional Green's function method" *J. Appl. Phys.*, vol. 91, pp. 2343, 2002; Z. Ren, R. Venugopal, S. Datta,, M.S. Lundstrom, D. Jovanovic, and J.G. Fossum, "The Ballistic Nanotransistor: A Simulation Study," *Int. Electron Dev. Meeting, Tech. Digest*, pp. 715-718, Dec. 10-13, 2000.
(3) M.P. Anantram,"Which nanowire couples better electrically to a metal contact: armchair or zigzag nanotube? Appl. Phys. Lett. **78**, 2055 (2001)..
(4) P.S. Damle, A.W. Ghosh and S. Datta,"Unified description of molecular conduction: From molecules to metallic wires," *Phys.Rev.*.vol. **B64**: pp. 201403 (R) (2001).
(5) M.Paulsson, F. Zahid and S. Datta, "Resistance of a molecule", Chapter in *Nanoscience, Engineering and Technology Handbook*, eds. W. Goddard, D. Brenner, S. Lyshevski and G. Iafrate, CRC Press (2002), to be published.
(6) S. Datta, "Nanoscale device modeling: the Green's Function method," *Superlattices and Microstructures*, vol. 28, pp. 253 (2000).
(7) The derivation of the NEGF equations typically involves the second quantized formalism, though simpler derivations based on the one-particle Schrodinger equation can be worked out: M. Paulsson, *One-particle NEGF*, unpublished notes; S. Datta, *Quantum Phenomena: From Atoms to Transistors* (unpublished)
(8) This seminal idea can be traced to M. Buttiker, *Phys. Rev.* B33, 3020 (1986).
(9) See for example, A. Scholze, A. Schenk and W. Fichtner, "Single-Electron Device Simulation", Special Issue on Computational Electronics, IEEE Transactions on Electron Devices vol.47, 1811 (2000).
(10) This is easily done with a MATLAB program on a PC – the author will be glad to share his program with interested readers.