12-26-2012

Suppressing Leakage by Localized Doping in Si Nanotransistor Channels

Jesse Maassen
McGill University; Birck Nanotechnology Center, Purdue University, jmaassen@purdue.edu

Hong Guo
McGill University

Maassen, Jesse and Guo, Hong, "Suppressing Leakage by Localized Doping in Si Nanotransistor Channels" (2012). *Birck and NCN Publications.* Paper 1087.
http://dx.doi.org/10.1103/PhysRevLett.109.266803

Suppressing Leakage by Localized Doping in Si Nanotransistor Channels

Jesse Maassen[1,2,*] and Hong Guo[1]

[1]Department of Physics and Centre for the Physics of Materials, McGill University, Montreal, Quebec H3A 2T8, Canada
[2]Network for Computational Nanotechnology, Purdue University, West Lafayette, Indiana 47907, USA

(Received 29 March 2012; published 26 December 2012)

By first principles atomistic analysis we demonstrate how controlled localized doping distributions in nanoscale Si transistors can suppress leakage currents. We consider dopants (B and P atoms) to be randomly confined to a ≈ 1 nm width doping region in the channel. If this region is located away from the electrodes, roughly 20% of the channel length L, the tunneling leakage is reduced $2\times$ compared to the case of uniform doping and shows little variation. Oppositely, we find the leakage current increases by orders of magnitude and may result in large device variability. We calculate the maximum and minimum conductance ratio that characterizes the tunnel leakage for various values of L. We conclude that doping engineering provides a possible approach to resolve the critical issue of leakage current in nanotransistors.

DOI: 10.1103/PhysRevLett.109.266803 PACS numbers: 73.63.−b, 31.15.A−, 73.22.−f, 73.23.Ad

An increasingly important aspect of nanoscale devices is the inevitable fluctuations that occur due to atomic disorder or randomness [1,2]. When system sizes are small enough (~ 10 nm) to reveal the discreteness or graininess of materials, the statistical fluctuations around the average properties become very important. Experimentally, one source of variation that has already become a serious problem for modern commercial devices is the effect of random dopant fluctuations (RDF). RDF results in random performance fluctuations due to the particular microscopic arrangement of the small number of dopant atoms in nanoscale devices. The primary consequence of RDF is fluctuations in the threshold voltage (gate voltage at which the device is ''ON'') that compromises device performance and functionality. It is impossible to design functional systems if physical properties fluctuate wildly and randomly from device to device. One solution is to precisely control the position of the dopants [3], as demonstrated in a single-dopant transistor [4]. However, incorporating disorder or randomness in nanoelectronics modeling remains of paramount importance [1].

In this work, instead of considering the detrimental effect of RDF on the ON-state performance as is common [5–8], we employ first principles transport modeling to investigate how highly controlled doping profiles can improve OFF-state characteristics. In particular, we focus on variations in leakage current in nanoscale Si transistor structures. As devices continue to shrink, leakage currents (source-drain current in the OFF state) are found to increase, due to quantum mechanical tunneling and reduced barrier heights, which significantly contributes to the rising power consumption of transistors [9]. Our results strongly suggest that careful control over the location of dopants in the device channel can reduce leakage currents, by orders of magnitude, and may minimize device variability. The importance of this effect is significantly dependent on channel length L, where the conductance is found to fluctuate by a factor of $2\times$ at $L = 6.5$ nm while a factor nearing $10^5\times$ is observed at $L = 15.2$ nm. In overview, we study leakage current as a function of varying doping locations as well as channel lengths and determine the interplay between thermionic and tunneling conduction.

Theoretical model.—Instead of considering completely random and uniform doping in the entire transistor channel, as is typical in RDF research, here we focus on how device characteristics can be *improved* by controlling and localizing the dopants in certain regions along the channel. Specifically, we fix the doping concentration and confine the dopants to a 1.1 nm region in the channel [see Fig. 1(a)]: in this confined region the dopants are randomly distributed. We consider Si n-p-n and p-n-p device structures with L ranging from 6.5 nm to 15.2 nm, where n and p designate electron and hole doped, respectively. The modeled device structures exclude the effect of a gate voltage, which is physically akin to Si transistors in the OFF state. The source and drain leads are uniformly and heavily doped to a doping concentration of 5×10^{19} cm^{-3}. For n-type (p-type) channels, we consider phosphorus (boron) dopant atoms with a concentration of 5×10^{18} cm^{-3}.

Our calculation is based on atomic first principles where density functional theory (DFT) is carried out within the Keldysh nonequilibrium Green's function formalism [10]. NEGF-DFT is the state-of-the-art approach for parameter-free atomistic modeling of quantum transport properties. To deal with doping at the atomic level, we apply the coherent potential approximation (CPA) for the open device structure [11]. The NEGF-DFT-CPA formalism is implemented in the NANODSIM first principles quantum transport package [12]. This technique is designed for quantum transport calculations of nanoscale devices with random disorder, without using any phenomenological parameters. It allows for arbitrary doping concentrations, includes the chemical interactions of the dopant atoms and extracts the average properties in a single calculation.

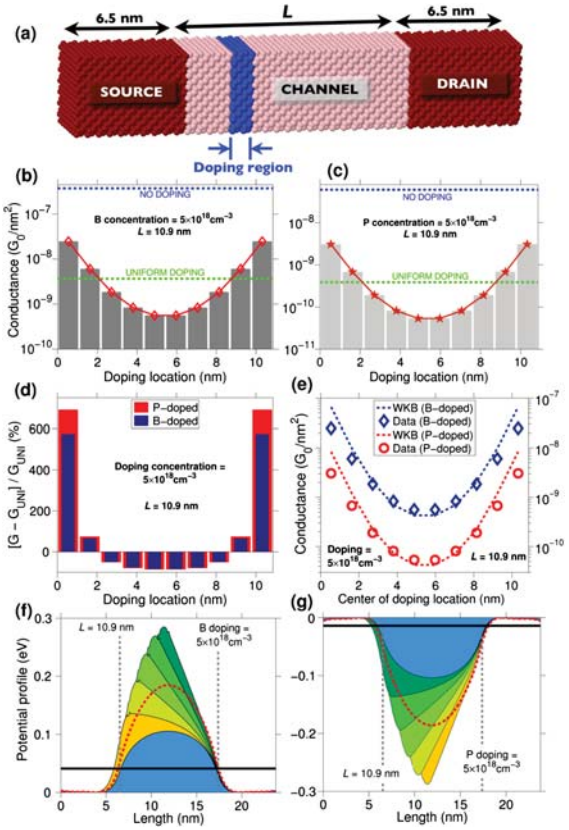

FIG. 1 (color online). (a) Diagram of the simulated system. The doping length (indicated with arrows) is set to 1.1 nm for this study. (b, c) Conductance per area versus doping location for B- and P-doped channels for $L = 10.9$ nm and $C_d = 5 \times 10^{18}$ cm^{-3}. (d) Relative conductance variation around the case of uniform doping versus doping location. (e) Comparison between *ab initio* G and WKB conductance per area [using the self-consistent potential shown in panels (f, g)]. The amplitude of the WKB G is adjusted to the simulated data. The conductivity effective mass for silicon was used, $m_e = 0.26$ and $m_h = 0.36$ [20]. (f, g) Potential profile at each point along the simulation box. From front to back the dopants are located in layers starting next to the source and ending in the center of the channel. The area (dashed) curve at the forefront corresponds to the potential profile of the undoped (uniformly doped) channel and the solid horizontal lines indicate E_F.

Moreover, NEGF-DFT-CPA naturally and self-consistently computes the energy levels and occupations of the disorder sites, as well as the scattering properties of the potentials due to impurities which depend on screening determined through the self-consistent electronic density. The approximation made when utilizing CPA is the assumption that the disorder sites are statistically uncorrelated (for more on CPA, see Ref. [13]). Employing other self-consistent DFT techniques for this problem is intractable, as one would need to simulate 10^4 atoms with a single dopant. The modified Becke-Johnson exchange potential [14] is employed to correctly predict the experimental band gap of Si (= 1.11 eV). The simulated device possesses an infinite cross section, as we employ periodic boundary conditions in the directions perpendicular to transport. In this study we consider the effect of RDF along the transport direction, and the cross section of the supercell is equal to $a_{\mathrm{Si}} \times a_{\mathrm{Si}}$ where $a_{\mathrm{Si}} = 5.43$ Å is the lattice constant. To ensure a smooth matching of the potential at the simulation box edges, 6.5 nm of buffer layers are used for the source and drain [as shown in Fig. 1(a)]. All other numerical details can be found in Ref. [15].

Localized doping.—To begin, we keep L fixed to 10.9 nm. In Figs. 1(b) and 1(c), we plot the OFF-state conductance per area ($V = 0$) versus doping location for B-doped (n-p-n) and P-doped (p-n-p) channels. Note that throughout this article, reference to the conductance G implies a conductance per area. Depending on the doping location, large variations in conduction are observed. The largest (smallest) tunneling probability corresponds to the dopants nearest (farthest) to the electrodes. The ratio of maximum conductance over minimum conductance equals ≈ 44 for B doped and ≈ 57 for P doped. It is noteworthy that conduction is almost entirely ballistic, as the diffusive contribution to the transmission—which we also calculated by including the vertex correction to the transmission formula [11], is roughly 0.01% of the total.

We also plot the conductance G obtained by simulating a uniform doping profile and a channel without doping. As expected, the uniform doping conductance G_{UNI} is in between the maximum and minimum conductance values (G_{MAX} and G_{MIN}) obtained via localized doping. In Fig. 1(d), we show the relative conductance variation around G_{UNI} for both B- and P-doped channels. The conductance deviates most significantly ($\sim 600\%$) from G_{UNI} when the dopants are located in the first nanometer next to the source or drain. Importantly, Fig. 1(d) also indicates that one can achieve, on average, a 70% reduction in tunneling current (relative to G_{UNI}) as long as the dopants are located anywhere except the first 2 nm of the source or drain.

To better understand the behavior of G with varying doping location, in Figs. 1(f) and 1(g) we show the potential profile at every point along the channel. With the exception of the area curve at the forefront, each area plot corresponds to the potential with the dopants moving from the first layer (next to the source) to the middle layer. The absolute value of the potential increases significantly as the dopants move towards the center of the channel. By grouping the dopants in a narrow section of the channel, instead of uniformly doping (red dashed line), a larger potential barrier is obtained since the concentrated doping charge enhances the local work function difference between the lead and the channel. When the dopants are near the lead-channel interface the holes (electrons) originating from the B (P) dopants partially combine with the excess electrons (holes) donated from the n-type (p-type) lead.

To confirm that the conduction variations shown in Figs. 1(b) and 1(c) originate from the particular shape of

the potential profile, we compare against the Wentzel-Kramers-Brillouin (WKB) tunneling model [16]. The expression for WKB tunneling depends solely on the barrier shape:

$$T \sim e^{-2\gamma}, \qquad \gamma = \frac{1}{\hbar}\int \sqrt{2m^*|V(z) - E_F|}dz, \quad (1)$$

where $V(z)$ is the potential profile, m^* is the conductivity effective mass, and $h = 2\pi\hbar$ is Planck's constant. The integration occurs only in the region where $V(z) - E_F > 0$, i.e., where carriers must tunnel through the barrier. Figure 1(e) compares the WKB conductance versus the simulated data using a proportionality constant as a parameter. The agreement is good, which confirms that the variations in G are solely due to the particular shape of the tunneling barrier.

In Figs. 1(b) and 1(c), the conductance of the B-doped channel is roughly a decade larger than that of the P-doped channel. The main reason is the difference in E_F relative to the band edge for the conduction ε_c and valence ε_v bands. With the leads doped to 5×10^{19} cm^{-3}, we have $E_F - \varepsilon_c = 0.041$ eV and $|E_F - \varepsilon_v| = 0.014$ eV, due to the particular curvature of the bands. From Eq. (1) for WKB tunneling, it is clear that a small change in $|V(z) - E_F|$ will influence the tunneling probability exponentially.

Varying channel length.—The effect of channel length on leakage current with regard to localized doping is found to be very illuminating. Figures 2(a) and 2(c) present G versus doping location as a function of L, for B- and P-doped channels. Two trends are clear: (i) the average G decreases with increasing L and (ii) the variations in G are more pronounced with increasing L. Regarding (i), G increases significantly for small L because the potential barrier height begins to drop when $L < 2l_s$, where l_s is the screening length. If the channel is very short, incomplete screening of the lead-channel interface prevents the potential from attaining its maximum value. For $L = 6.5$ nm, the barrier height is almost small enough to allow carriers to travel above the barrier. Simultaneously, the barrier width always decreases linearly with L, which further enhances tunneling probabilities when $L < 15$ nm.

In order to highlight point (ii), Fig. 2(b) shows both G_{MAX} and G_{MIN} versus L. We find that the conductance drops very rapidly and the disparity between G_{MAX} and G_{MIN} increases with L. Figure 2(d) presents the ratio of $G_{\mathrm{MAX}}/G_{\mathrm{MIN}}$ versus L. This plot indicates the importance of doping location; the larger the ratio, the more significant the effect. For $L = 6.5$ nm the G ratio is ≈ 2, thus changing the position of the dopants is only somewhat important. However, for $L = 15.2$ nm the ratio approaches 10^5, indicating that control over the position of the dopants can be crucial for device performance and variability. The source of this effect is the result of a more rapidly increasing potential when the dopants are near the center of the channel versus next to the electrodes.

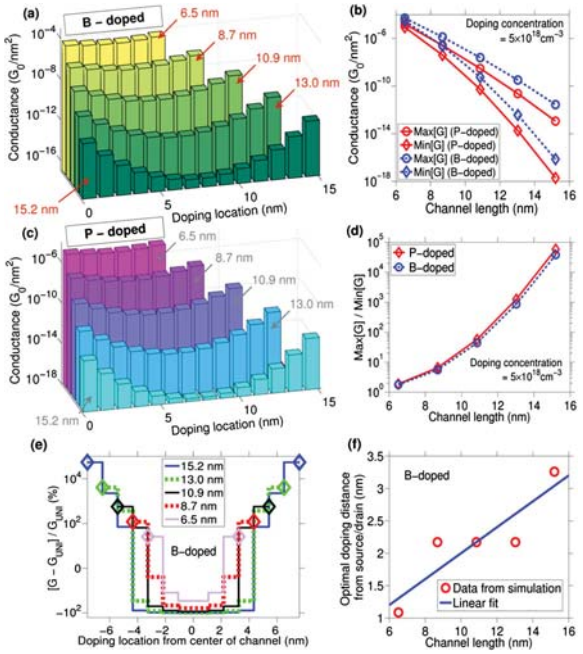

FIG. 2 (color online). (a)–(c) Conductance versus doping location as a function of channel length, for B- and P-doped channels. (b) Maximum G and minimum G [taken from panels (a) and (c)] versus channel length L. (d) Ratio of Max$[G]$/Min$[G]$ versus L. (e) Relative conductance variation around the case of uniform doping versus doping location as a function of L. (f) Optimal doping distance from the source or drain versus L. The optimal distance corresponds to the doping layer where the relative conductance variation in (e) becomes negative. C_d is fixed to 5×10^{18} cm^{-3}.

Last, in Fig. 2(e), we show the relative conductance variation around the case of uniform doping. For short channels the positive and negative fluctuations are roughly equal, while for longer channels the positive values grow larger and the negative values appear to saturate. Thus, for larger L, control over doping distributions can decrease the leakage current by a factor of 2 compared to G_{UNI}. However, a lack of control can lead to leakage currents up to $\sim 10^2$–$10^3 \times G_{\mathrm{UNI}}$. Figure 2(e) also indicates the regions where doping is to be avoided, namely in the vicinity of the source or drain. In Fig. 2(f), we plot the distance from the source or drain at which G is less than G_{UNI}. Since the doping layers are 1.1 nm in length, the data points appear as discrete steps corresponding to 1, 2, and 3 doping layers. From a linear fit, the optimal doping distance from the source or drain is found to be roughly 20% of L.

Tunneling versus thermionic conduction.—Up to this point, we have only considered conduction at E_F [$G = G_0 T(E_F)$], corresponding to experiments performed at low temperature and small bias. In the following, we study how the effect of temperature influences the transport characteristics. Starting from the expression of the current [17],

$$I = \frac{2e}{h}\int_{-\infty}^{+\infty} T(\varepsilon)[f_L(\varepsilon) - f_R(\varepsilon)]d\varepsilon, \quad (2)$$

where $f_{L,R} = f(\varepsilon - \mu_{L,R})$ is the Fermi-Dirac occupation function and $T(\varepsilon)$ is the transmission. In the limit of vanishing temperature T and applied voltage V, one finds $f_L - f_R \approx \delta(\varepsilon - E_F)eV$ which yields $I/V = 2e^2/hT(E_F) = G_0T(E_F)$. Here, we are interested in the case of $V \to 0$, but finite T. By expanding the integrand of Eq. (2) to first order in V, we find

$$G = I/V = \int_{-\infty}^{+\infty} G(\varepsilon)d\varepsilon = \int_{-\infty}^{+\infty} G_0\beta T(\varepsilon)e^{\beta(\varepsilon - E_F)}f^2(\varepsilon - E_F)d\varepsilon, \quad (3)$$

where $\beta \equiv (k_BT)^{-1}$, k_B is Boltzmann's constant. This relation describes how thermally excited carriers contribute to the current. Using this expression, we can determine both the tunneling (carriers going through the potential barrier) and thermionic (carriers excited above the potential barrier) contributions to the current.

In Figs. 3(a) and 3(b), the transmission versus energy (relative to E_F) is presented for P and B doping. As a comparison, we show the case of doping in the first and middle layers, which yield the highest and lowest $T(\varepsilon)$ values. The barrier height energies indicate the separation between carriers traveling through the barrier (tunneling) and above the barrier (thermionic). Tunneling transport is found to vary exponentially while the ballistic transport of the thermionic transmission shows a linear distribution, as expected for 3D semiconductors [18].

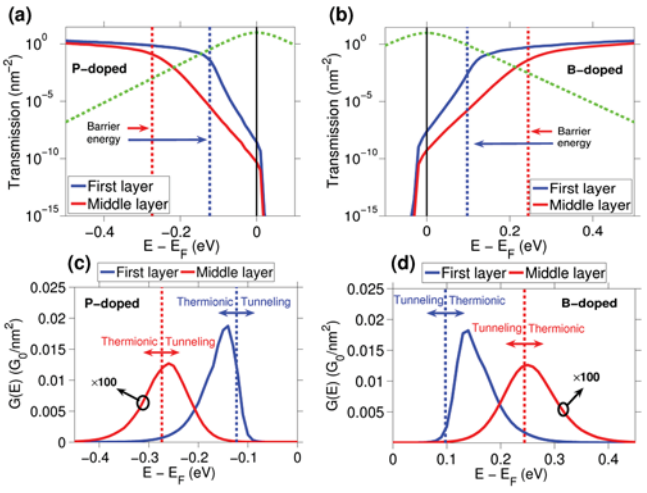

FIG. 3 (color online). (a, b) Transmission versus energy (relative to E_F) for P- and B-doped channels, which correspond to p- and n-type leads, respectively. The blue (red) curve corresponds to doping in the first (middle) layer. The barrier height energy is also indicated. The occupation function for transport [the integrand of Eq. (3) divided by $G_0T(\varepsilon)$] at $T = 300$ K is shown as a dashed green line. (c, d) Energy-dependent conductance $G(\varepsilon)$ [Eq. (3)] versus energy for P- and B-doped channels. Note that $G(\varepsilon)$ for the middle layer was multiplied by 100, for clarity. $L = 10.9$ nm and $C_d = 5 \times 10^{18}$ cm^{-3}.

Figures 3(c) and 3(d) plot the energy-dependent conductance [$G(\varepsilon)$, from Eq. (3)] versus energy for P and B doping. At $T = 300$ K, we observe that G for middle layer (ML) doping is still much smaller than that of first layer (FL) doping. Moreover, the average ε is pushed further from E_F as doping moves from FL to ML. The small barrier height of FL doping can easily be surpassed by excited carriers, thus leading to a large thermionic contribution (P, 89% and B, 98%). The opposite behavior is observed for ML, which results in roughly equal contributions of tunneling (P, 58% and B, 42%) and thermionic (P, 42% and B, 58%) conduction. The larger thermionic contribution for B doping originates from the larger $E_F - \varepsilon_c$ value, compared to P doping, which decreases the effective barrier height (as discussed above).

Interestingly, G increases by 10^5–10^6 going from $T = 0$ to 300 K (see Table I). This occurs because $T(\varepsilon)$ increases exponentially as ϵ moves away from E_F. Importantly, at room temperature the variations in G due to doping location are further enhanced compared to $T = 0$. Table I shows that the ratio $G(\mathrm{FL})/G(\mathrm{ML})$ reaches $\approx 10^2$, roughly double that at zero T. Thus, the effect of localized doping is even more pronounced for common device temperatures.

Summary.—We have investigated the effect of localized doping on leakage current in short channel Si transistor structures using first principles transport simulations. Assuming the dopants (B or P) are confined to 1.1 nm regions along the channel, our results predict large conductance fluctuations as a function of doping location. The effect becomes more pronounced with increasing L, with $G_{\mathrm{MAX}}/G_{\mathrm{MIN}}$ ranging from 2 ($L = 6.5$ nm) to 10^5 ($L = 15.2$ nm). By placing dopants near the center of the channel, leakage current is always below that of uniform doping and varies only slightly; however, dramatic increases occur when doping in the vicinity (roughly 20% of L) of the leads. Last, our calculations demonstrate that the variations in G are even more pronounced when going from $T = 0 \to 300$ K. In summary, if nanometer control over dopant locations is beyond our current

TABLE I. Conductance of P- and B-doped channels with doping localized in the first and middle layer for $T = 300$ K and $T = 0$ K. Ratio of first layer-middle layer conductance is shown. $L = 10.9$ nm and $C_d = 5 \times 10^{18}$ cm^{-3}. FL, first layer and ML, middle layer.

	FL	ML
$G(G_0/\mathrm{nm}^2)$	1.4×10^{-3} (P)	1.4×10^{-5} (P)
$[T = 300$ K]	1.4×10^{-3} (B)	1.5×10^{-5} (B)
$G(\mathrm{FL})/G(\mathrm{ML})$	101 (P)	
$[T = 300$ K]	93 (B)	
$G(G_0/\mathrm{nm}^2)$	3.1×10^{-9} (P)	5.4×10^{-11} (P)
$[T = 0$ K]	2.5×10^{-8} (B)	5.6×10^{-10} (B)
$G(\mathrm{FL})/G(\mathrm{ML})$	57 (P)	
$[T = 0$ K]	44 (B)	

capability, one can reduce leakage current and perhaps lessen device variability by minimizing the probability of dopants appearing next to the source and drain. We note in passing that apparently a similar strategy is being considered commercially [19].

We gratefully acknowledge the financial support of NSERC of Canada and FQRNT of Quebec. We thank CLUMEQ and RQCHP for providing computation facilities.

*jmaassen@purdue.edu

[1] International Technology Roadmap for Semiconductors, 2009, http://public.itrs.net/.
[2] A. Brown, A. Asenov, and J. Watling, IEEE Trans. Nanotechnol. **1**, 195 (2002).
[3] M. Hori, T. Shinada, Y. Ono, A. Komatsubara, K. Kumagai, T. Tanii, T. Endoh, and I. Ohdomari, Appl. Phys. Lett. **99**, 062103 (2011).
[4] G. P. Lansbergen, R. Rahman, C. J. Wellard, I. Woo, J. Caro, N. Collaert, S. Biesemans, G. Klimeck, L. C. L. Hollenberg, and S. Rogge, Nat. Phys. **4**, 656 (2008).
[5] A. Asenov, IEEE Trans. Electron Devices **45**, 2505 (1998).
[6] A. Martinez, J. Barker, A. Svizhenko, M. Anantram, and A. Asenov, IEEE Trans. Nanotechnol. **6**, 438 (2007).
[7] T. Markussen, R. Rurali, A.-P. Jauho, and M. Brandbyge, Phys. Rev. Lett. **99**, 076803 (2007).
[8] M. Amato, S. Ossicini, and R. Rurali, Nano Lett. **12**, 2717 (2012).
[9] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, Proc. IEEE **91**, 305 (2003).
[10] J. Taylor, H. Guo, and J. Wang, Phys. Rev. B **63**, 245407 (2001).
[11] Y. Ke, K. Xia, and H. Guo, Phys. Rev. Lett. **100**, 166805 (2008).
[12] Nanoacademic Technologies, http://nanoacademic.ca/.
[13] I. Turek *et al.*, *Electronic Structure of Disordered Alloys, Surfaces and Interfaces* (Kluwer Academic Publishers, Boston, 1997).
[14] F. Tran and P. Blaha, Phys. Rev. Lett. **102**, 226401 (2009).
[15] At equilibrium $V = 0$, the complex energy contour $r_C e^{i\theta}$ (used to integrate the Green's function for the electronic density) is cut into two segments $\theta = [\pi \rightarrow \pi/6, \pi/6 \rightarrow 0]$ where $d_C = 2r_c = 1.25\,\mathrm{Ry}$ is the contour diameter. An additional energy integration is performed $\pm 15 k_B T$ around E_F along the real-energy axis, where $T = 300$ K in this case. For each of the three integration segments the number of energy and $k_\perp$ points are [10,10,20] and $[5, 13, 15]^2$, respectively. For transmission calculations, the number of $k_\perp$ is 45×45.
[16] D. J. Griffiths, *Introduction to Quantum Mechanics* (Prentice Hall, Toronto, 1995).
[17] S. Datta, *Quantum Transport: Atom to Transistor* (Cambridge University Press, New York, 2005).
[18] R. Kim, S. Datta, and M. S. Lundstrom, J. Appl. Phys. **105**, 034506 (2009).
[19] http://spectrum.ieee.org/semiconductors/devices/startup-seeks-new-life-for-planar-transistors.
[20] E. Barta, Infrared Phys. **17**, 111 (1977).