Quantum capacitance in nanoscale device modeling

D. L. John, a L. C. Castro, and D. L. Pulfrey b
Department of Electrical and Computer Engineering, University of British Columbia, Vancouver BC V6T 1Z4, Canada

(Received 12 May 2004; accepted 6 August 2004)

Expressions for the “quantum capacitance” are derived, and regimes are discussed in which this concept may be useful in modeling electronic devices. The degree of quantization is discussed for one- and two-dimensional systems, and it is found that two-dimensional (2D) metals and one-dimensional (1D) metallic carbon nanotubes have a truly quantized capacitance over a restricted bias range. For both 1D and 2D semiconductors, a continuous description of the capacitance is necessary. The particular case of carbon nanotube field-effect transistors (CNFETs) is discussed in the context of one-dimensional systems. The bias regime in which the quantum capacitance may be neglected when computing the energy band diagram, in order to assist in the development of compact CNFET models, is found to correspond only to the trivial case where there is essentially no charge, and a solution to Laplace’s equation is sufficient for determining a CNFET’s energy band diagram. For fully turned-on devices, then, models must include this capacitance in order to properly capture the device behavior. Finally, the relationship between the transconductance of a CNFET and this capacitance is revealed. © 2004 American Institute of Physics.

[DOI: 10.1063/1.1803614]

I. INTRODUCTION

The concept of “quantum capacitance” was used by Luryi 1 in order to develop an equivalent circuit model for devices that incorporate a highly conducting two-dimensional (2D) electron gas. Recently, this term has also been used in the modeling of one-dimensional (1D) systems, such as carbon nanotube (CN) devices. 2,3 Here, we derive expressions for this capacitance in one and two dimensions, showing the degree to which it is quantized in each case.

Our discussion focuses primarily on the 1D case, for which we use the carbon nanotube field-effect transistor (CNFET) as the model device, although the results apply equally well to other types of 1D semiconductors. The 2D case has been discussed in Ref. 1, and is included here only to illustrate key differences.

Equilibrium expressions are derived, and these are extended to cover two extremes in the nonequilibrium characteristic, namely, phase-coherent and phase-incoherent transport. In the former case, the wave function is allowed to interfere with itself, and may produce resonances depending on the structure of the device. This results in the charge, and the quantum capacitance, becoming strong functions of the length of the semiconductor. In the latter case, this type of resonance is not allowed, and the quantum capacitance is more uniform. Finally, we show how the quantum capacitance affects the transconductance of a CNFET, where the Landauer expression can be used to compute the current. 4

II. EQUILIBRIUM QUANTUM CAPACITANCE

In order to derive analytical expressions, it is assumed that our device is in quasi-equilibrium, and that the carrier distribution functions are rigidly shifted by the local electrostatic potential. If the density of states (DOS) is symmetric with respect to the Fermi level $E_F$, as in graphene, then we can write the charge density $Q$, due to electrons and holes in the semiconductor, as

$$Q = q \int_0^\infty g(E) \left[ f \left( E + \frac{E_G}{2} + qV_a \right) - f \left( E + \frac{E_G}{2} - qV_a \right) \right] dE,$$

where $q$ is the magnitude of the electronic charge, $E$ is the energy, $g(E)$ is the 1D or 2D DOS, $f(E)$ is the Fermi-Dirac distribution function, $V_a$ is the local electrostatic potential, $E_G$ is the band gap, and $E_F$ is taken to be midgap when $V_a = 0$. The quantum capacitance $C_Q$ is defined as

$$C_Q = \frac{\partial Q}{\partial V_a},$$

and has units of $F/m^2$ and $F/m$ in the 2D and 1D cases, respectively.

A. Two dimensions

In the two-dimensional case, if we employ the effective-mass approximation with parabolic bands, the DOS is given by

$$g(E) = \frac{m}{\pi \hbar^2} \nu(E),$$

where $\nu(E)$ is the number of contributing bands at a given energy, $m$ is the effective mass, and $\hbar$ is Dirac’s constant. If we combine this with Eqs. (1) and (2), and exchange the order of differentiation and integration, we get
C_Q = \frac{mv^2}{4\pi^2kT} \int_0^\infty \nu(E) \left[ \text{sech}^2 \left( \frac{E + Eg/2 - qV_a}{2kT} \right) \right. \\
+ \left. \text{sech}^2 \left( \frac{E + Eg/2 + qV_a}{2kT} \right) \right] dE, 
\tag{4}

where k is Boltzmann's constant and T is temperature. If \nu is a constant, we can perform the integration to get

\[ C_Q = \frac{vma^2}{2\pi^2} \left[ 2 - \frac{\sinh(Eg/2kT)}{\cosh(\frac{Eg/2 - qV_a}{2kT}) \cosh(\frac{Eg/2 + qV_a}{2kT})} \right], \]

which reduces to

\[ C_Q = \frac{vma^2}{\pi a^2}, \tag{5} \]

when \( Eg=0 \) in agreement with Ref. 1, where metallic properties were assumed. Note that this function is quantized in the metallic case, but continuous for a semiconductor. For \( Eg \) greater than about 15\( kT \), however, the function makes a rapid transition from a small value to that given by Eq. (6) when \( V_a \) crosses \( Eg/2 \), and is thus effectively quantized.

**B. One dimension**

In the one-dimensional, effective-mass case, we have

\[ g(E) = \frac{\nu(E)}{\pi h} \sqrt{\frac{2m}{E}}. \tag{7} \]

The explicit energy dependence of this DOS complicates the evaluation of our integral for \( C_Q \). The approach suggested in Ref. 2, i.e., using the fact that the derivative of \( f(E) \) is peaked about \( Ef \) in order to approximate this integral using a Sommerfeld expansion,\(^5\) cannot be done in general, due to the presence of singularities in the 1D DOS.

The capacitance is given by

\[ C_Q = \frac{q^2}{2kTh} \int_0^\infty \frac{\nu(E)}{\sqrt{E}} \left[ \text{sech}^2 \left( \frac{E + Eg/2 - qV_a}{2kT} \right) \right. \\
+ \left. \text{sech}^2 \left( \frac{E + Eg/2 + qV_a}{2kT} \right) \right] dE, \tag{8} \]

where \( h \) is Planck's constant. For sufficiently large \( |V_a| \), we can completely neglect one of the \( \text{sech}^2(\cdot) \) terms. As a simple example, if \( V_a=0.1 \text{ V} \) for a material with \( Eg\approx 1 \text{ eV} \), the contribution to the integral from the first term is roughly four orders of magnitude greater than the second. This approximation is equivalent to neglecting hole charge for positive \( V_a \), and electron charge for negative \( V_a \).

The solid line in Fig. 1 shows the equilibrium \( C_Q \) as a function of \( V_a \) for a semiconductor with two valence and conduction bands: at 0.2 and 0.6 eV away from the Fermi level. An effective mass of 0.06\( m_0 \) is assumed, where \( m_0 \) is the free-electron mass. The van Hove singularities, at each band edge, result in corresponding peaks in \( C_Q \).

For a linear energy–wavevector relation, such as that near the Fermi level in graphene or a metallic CN, the DOS is constant. This is the case considered by Burke,\(^3\) and is valid when \( V_a \) is such that \( f(E) \) is approximately zero before the first van Hove singularity is encountered in the integral. Since the higher energy bands are not relevant to the integration under such a condition, \( \nu \) is constant, and the DOS is given by

\[ g(E) = \frac{2\nu}{hv_F}, \tag{9} \]

where \( v_F \) is the Fermi velocity. The result is

\[ C_Q = \frac{2\nu q^2}{hv_F}, \tag{10} \]

which agrees with the expression quoted in Ref. 3.

Note that in Eq. (8) \( C_Q \) does not manifest itself as a multiple of some discrete amount, so “quantum capacitance” is not an appropriate description for a 1D semiconductor, unlike in the metallic 2D and metallic 1D CN cases, where the capacitance is truly quantized.

**III. GENERAL CONSIDERATIONS**

We can now extend our discussion to include the non-equilibrium behavior for a general, 1D, intrinsic semiconductor. All of the numerical results are based on the methods described in Refs. 6 and 7, which consider the cases when transport in the 1D semiconductor is either coherent or incoherent, respectively. While these methods were developed in order to describe CNFETs, their use of the effective-mass approximation allows them to be used for any device, and bias, where the semiconductor is described well by this approximation.

For phase-incoherent transport, we utilize a flux-balancing approach\(^7,8\) to describe the charge in an end-contacted semiconductor. If we consider only the electrons that are far away from the contacts, i.e., in the midlength region, Equation (1) becomes\(^8\)
\[ Q = -\frac{q}{2} \int_{-\infty}^{\infty} \frac{g(E) T^* (E)}{2} \left[ f \left( E + V_a - qV_0 \right) \left( \frac{2}{T_L(E)} - 1 \right) \right. \]
\[ + f \left( E + \frac{E_G}{2} + q(V_{\text{bias}} - V_0) \right) \left( \frac{2}{T_R(E)} - 1 \right) \bigg] dE, \]  

where \( V_{\text{bias}} \) is the potential difference between the end contacts, \( T_L(E) \) and \( T_R(E) \) are the transmission probabilities at the left and right contacts, respectively, \( T^*(E) = T_L(T_R) + T_R - T_L \) is the composite transmission probability for the entire system, and \( V_0 \) is evaluated in the midlength region. A similar expression holds for holes.

The first term in Eq. (11) resembles the equilibrium case, so we expect a similar form for that contribution to \( C_Q \). The peak for each contributing band will occur at the same \( V_a \), but the overall magnitude will be smaller due to the multiplication by the transmission function. The second term is also similar except that these peaks will now be shifted by \( V_{\text{bias}} \). This is depicted by the dashed curve in Fig. 1, where the case illustrated by the solid curve has been driven from equilibrium by \( V_{\text{bias}} = 0.2 \) V. Note the splitting of each large peak into two smaller peaks: one at the same point and the other shifted by \( V_{\text{bias}} \). Of course, the numerical value of the nonequilibrium capacitance depends on the exact geometry considered, as it will influence both \( V_a \) and the transmission probabilities in Eq. (11), but the trends shown here are general and geometry independent.

For the coherent, nonequilibrium case, it is instructive to consider a metal-contacted device, in which the band discontinuities at the metal-semiconductor interfaces are sufficient to allow significant quantum-mechanical reflection of carriers even above the barrier. Further, we restrict our attention to short devices since the importance of coherence effects is diminished as the device length is increased. Due to the phase coherence, then, we have a structure very much like a quantum well, even for devices where tunneling through the contact barriers is not important. For our device, we expect quasibound states to emerge at the approximate energies

\[ E_n \approx \frac{n^2 \pi^2 \hbar^2}{2mL^2}, \]

where \( L \) is the semiconductor length. For \( m = 0.06m_0 \) such as in a (16,0) CN, \( E_n = 6.3(n/L)^2 \) eV, where \( L \) is in nanometers. This may be compared with the result for metallic CNs, where the linear energy–wavevector relationship yields a \( 1/L \) dependence. Figure 2(a) displays \( C_Q \) as a function of position and \( V_a \) for this choice of \( m \). The maxima, indicated as brighter patches, show a dependence on \( V_a \) that reveals the population of quasibound states. Moreover, the maxima in position clearly show the characteristic modes expected from our simple square-well analogy. Note that the peak-splitting occurs for coherent transport as well, as shown in Fig. 2(b), where the peaks have been split by \( V_{\text{bias}} = 0.1 \) V.

The main difference between the coherent and incoherent cases is the presence of the quasibound states. These serve to increase the number of \( C_Q \) peaks, since each quasibound state behaves like an energy band, and they also give rise to a strong spatial dependence. While Fig. 2 shows only a single-band, coherent result, inclusion of multiple bands would cause \( C_Q \) to exhibit peaks corresponding to each band, and to each quasibound state.

\[ C_Q(t) \approx C_{\text{ins}} \left\{ -\frac{q^2}{2} \int_{-\infty}^{\infty} \frac{g(E) T^* (E)}{2} \left[ f \left( E + V_a - qV_0 \right) \left( \frac{2}{T_L(E)} - 1 \right) \right. \]
\[ + f \left( E + \frac{E_G}{2} + q(V_{\text{bias}} - V_0) \right) \left( \frac{2}{T_R(E)} - 1 \right) \bigg] dE, \]  

where \( V_{\text{bias}} \) is the potential difference between the end contacts, \( T_L(E) \) and \( T_R(E) \) are the transmission probabilities at the left and right contacts, respectively, \( T^*(E) = T_L(T_R) + T_R - T_L \) is the composite transmission probability for the entire system, and \( V_0 \) is evaluated in the midlength region. A similar expression holds for holes.

The first term in Eq. (11) resembles the equilibrium case, so we expect a similar form for that contribution to \( C_Q \). The peak for each contributing band will occur at the same \( V_a \), but the overall magnitude will be smaller due to the multiplication by the transmission function. The second term is also similar except that these peaks will now be shifted by \( V_{\text{bias}} \). This is depicted by the dashed curve in Fig. 1, where the case illustrated by the solid curve has been driven from equilibrium by \( V_{\text{bias}} = 0.2 \) V. Note the splitting of each large peak into two smaller peaks: one at the same point and the other shifted by \( V_{\text{bias}} \). Of course, the numerical value of the nonequilibrium capacitance depends on the exact geometry considered, as it will influence both \( V_a \) and the transmission probabilities in Eq. (11), but the trends shown here are general and geometry independent.

For the coherent, nonequilibrium case, it is instructive to consider a metal-contacted device, in which the band discontinuities at the metal-semiconductor interfaces are sufficient to allow significant quantum-mechanical reflection of carriers even above the barrier. Further, we restrict our attention to short devices since the importance of coherence effects is diminished as the device length is increased. Due to the phase coherence, then, we have a structure very much like a quantum well, even for devices where tunneling through the contact barriers is not important. For our device, we expect quasibound states to emerge at the approximate energies

\[ E_n \approx \frac{n^2 \pi^2 \hbar^2}{2mL^2}, \]

where \( L \) is the semiconductor length. For \( m = 0.06m_0 \) such as in a (16,0) CN, \( E_n = 6.3(n/L)^2 \) eV, where \( L \) is in nanometers. This may be compared with the result for metallic CNs, where the linear energy–wavevector relationship yields a \( 1/L \) dependence. Figure 2(a) displays \( C_Q \) as a function of position and \( V_a \) for this choice of \( m \). The maxima, indicated as brighter patches, show a dependence on \( V_a \) that reveals the population of quasibound states. Moreover, the maxima in position clearly show the characteristic modes expected from our simple square-well analogy. Note that the peak-splitting occurs for coherent transport as well, as shown in Fig. 2(b), where the peaks have been split by \( V_{\text{bias}} = 0.1 \) V.

The main difference between the coherent and incoherent cases is the presence of the quasibound states. These serve to increase the number of \( C_Q \) peaks, since each quasibound state behaves like an energy band, and they also give rise to a strong spatial dependence. While Fig. 2 shows only a single-band, coherent result, inclusion of multiple bands would cause \( C_Q \) to exhibit peaks corresponding to each band, and to each quasibound state.

IV. APPLICATION: CNFETS

We now elaborate on the above in the context of CNFET modeling. In particular, for the purpose of developing compact models, it would be useful to ascertain the conditions under which the quantum capacitance is small in comparison with that due to the insulator geometry, a regime previously described as the “quantum capacitance limit.” To this end, we examine a coaxial CNFET, and treat \( V_a \) as the potential, with respect to the source contact, on the surface of the CN in the midlength region. \( C_Q \) can be considered to be in series with the insulator capacitance \( C_{\text{ins}} \); however, the ratio of these capacitances is related not to \( V_a \) and the gate-source voltage \( V_{GS} \), but to \( \partial V_{GS}/\partial V_a \). If the charge accumulation were linear over some bias range, as might be deemed appropriate at the local extrema of \( C_Q \), we could relate this ratio directly to the potentials.

Knowledge of the “\( C_Q \)” limit is beneficial since a relatively low \( C_Q \) implies that changes in \( V_a \) will closely track changes in \( V_{GS} \), obviating the need to calculate \( C_Q \) when computing the energy band diagram. Note, however, that \( C_Q \) cannot be neglected when considering performance metrics that depend on the total capacitance, i.e., the propagation delay may be dominated by \( C_Q \) in this limit. We find that \( C_Q \ll C_{\text{ins}} \) only when \( Q \) is small enough as to allow for the employment of a Laplace solution for the position-dependent potential: eliminating the need for a cumbersome self-consistent Schrödiger-Poisson solution. The difference between these solutions is illustrated in Fig. 3 for a coaxial CNFET with an insulator thickness and CN radius of 2.5 and 0.6 nm, respectively, and an end contact work function that is 0.6 eV less than that of the CN. Figures 3(a) and 3(b) correspond to the off and turn-on states, respectively. While equilibrium band diagrams are shown for simplicity, similar trends prevail with the application of a drain-source voltage.
For a device dominated by thermionic emission, such as the one depicted here, the discrepancy shown in (a), close to the contacts, will not significantly affect the current calculation, while in (b), the error would clearly be much greater. For a device dominated by tunneling, i.e., if the energy bands had the opposite curvature, a similar discrepancy would result in a large error in the current calculations due to the exponential dependence of the tunneling probability on the barrier shape.

Now, we seek to theoretically quantify the condition under which \( C_Q \lesssim C_{\text{ins}} \). From Fig. 1, we see that the first local maxima is on the order of 0.3 nF/m. For this peak to be insignificant, we would require \( C_{\text{ins}} \) to be orders of magnitude higher than this. For a 2 nm thick dielectric in a coaxial device, we would require a relative permittivity of \( \sim 530 \) in order to give two orders of magnitude difference between \( C_{\text{ins}} \) and \( C_Q \). Reports of solid, high-permittivity dielectrics for CN devices \(^{13-16}\) have quoted values only as high as 175 for the relative permittivity, \(^{16}\) so we conclude that, for realistic dielectrics, we can expect to only marginally enter the \( C_Q \) limit, and that the first \( C_Q \) peak will be significant. If we consider an electrolytically gated CNFET, \(^{17}\) we could perhaps achieve a relative permittivity of 80, and an effective thickness of 1 nm, as considered in Ref. 2, but this would yield \( C_{\text{ins}} \approx 25C_Q \), and would, again, only marginally be entering this limit.

For a short-channel, phase-coherent device, the requirement for negligible \( C_Q \) is that the Fermi levels for the injecting contacts should be far away from \( E_1 = 6.3/L^2 \) eV. If we consider positive applied voltages to the gate and drain, this would imply that \( qV_a \) should be more than about \( 5kT \) below \( E_G/2 + E_1 \). For the long-channel or phase-incoherent cases, this condition is given by \( E_1 = 0 \), corresponding to the conduction band edge. The relative importance of \( C_Q \), computed in the midlength region of the device, is depicted in Fig. 4 for a phase-incoherent device as a function of \( V_{GS} \) and the drain-source voltage \( V_{DS} \), where we note that \( V_{DS} \) corresponds to \( V_{\text{bias}} \), and that \( V_a \) is influenced by both \( V_{DS} \) and \( V_{GS} \). Here, the aforementioned peak splitting for nonzero \( V_{DS} \) is clearly evident in the diverging bright lines. Only for low bias voltages can \( C_Q \) be neglected, as shown by the black region in the center of the figure. However, this figure also reveals the regions where it becomes approximately constant, i.e., the bias ranges where the series capacitance relationship can be used to estimate \( V_a \) from \( V_{GS} \).\(^{11}\) Note, though, that this is a single-band calculation, and these regions may not be as prevalent when higher transverse modes are considered.

Finally, we consider the influence of \( C_Q \) on the transconductance for our model device, which has a doubly degenerate lowest band. If we employ the Landauer equation\(^ {4}\) for transport in two conducting channels, the current is

\[
I = \frac{4q}{h} \left[ \int_{E_{C}}^{\infty} T_s(E)[f(E) - f(E + qV_{DS})]dE - \int_{E_{V}}^{\infty} T_p(E)[f(E) - f(E - qV_{DS})]dE \right],
\]

(13)

where \( E_C = E_G/2 - qV_a \) is the spatially constant conduction band edge in the midlength region of a long-channel device, \( E_V = E_C - E_G \) is the valence band edge, and \( T_s(E) \) and \( T_p(E) \) are the transmission probabilities for electrons and holes, respectively, from one end contact to the other. The transconductance is defined as

\[
g_m = \frac{\partial I}{\partial V_{GS}},
\]

(14)

which yields

\[
g_m = \frac{4q^2}{h} \left[ \frac{C_{\text{ins}}}{C_Q + C_{\text{ins}}} \right] \left[ T_s(E_C)[f(E_C) - f(E_C + qV_{DS})] - T_p(E_V)[f(E_V) - f(E_V - qV_{DS})] \right].
\]

(15)

Note that, if we assume only electron transport with \( C_Q \ll C_{\text{ins}} \), low temperature, and high \( V_{DS} \), this expression re-
Contributes to the classic Landauer result\(^4\) for two conducting channels,
\[
g_m = \frac{4\pi^2}{\hbar^2} n, \tag{16}
\]
which is the ultimate transconductance in this case.\(^7,10\) Figure 5(a) shows the theoretical transconductance, from Eq. (15), for our model device, while (b) and (c) show the energy-distribution term (in curly braces) and the capacitance ratio term (in square brackets), respectively. The decrease in \(g_m\) at high \(V_{GD}\) is due primarily to the decreasing difference in the contact distribution functions as, for example, \(E_C\) becomes closer to \(qV_D\). However, the exact magnitude of \(g_m\) is dependent on the capacitance ratio. Further, \(C_Q\) will be responsible for additional oscillations in \(g_m\), if higher bands or quasi-bound states are considered in the calculation. Such transconductance features have been observed experimentally in Ref. 18, and have also been described in Ref. 19.

V. CONCLUSION

From this theoretical study on the charge-voltage relationship in one- and two-dimensional systems, it can be concluded that (1) the “quantum capacitance” occurs in discrete quanta for 2D and 1D metals if \(V_n\) is such that the Fermi level falls in a linear portion of the energy–wavevector relationship; (2) for 2D semiconductors, this capacitance is approximately quantized if the band gap is greater than about \(15kT\), and varies continuously otherwise; (3) for long, 1D systems with parabolic bands, and with \(V_n\) such that these bands contribute to the charge density, the equilibrium capacitance exhibits maxima that are related to the number of contributing bands at a given energy; (4) application of a bias to a 1D semiconductor causes each equilibrium capacitance peak to split into two smaller peaks, with one remaining at the equilibrium position, and the other shifting by the applied bias; (5) the potential in the midlength region of a 1D semiconductor cannot be computed, in general, from potential division due to two capacitors in series due to the nonlinearity of \(C_Q\); (6) for short, phase-coherent structures, the quasi-bound states cause the capacitance peaks to occur at higher local electrostatic potentials, with additional maxima corresponding to the occupation of these states; (7) for a CNFET, it is unlikely that the insulator capacitance can become high enough to allow the quantum capacitance to be neglected in energy band calculations, except in cases where the accumulated charge is low enough that the solution to Laplace’s equation is sufficient for the calculation, or if extremely high permittivity dielectrics are used; (8) the quantum capacitance has a significant effect on the transconductance, and should be considered when modeling CNFETs.

ACKNOWLEDGMENTS

This work was supported, in part, by the Natural Sciences and Engineering Research Council of Canada, and by the Institute for Computing, Information and Cognitive Systems at the University of British Columbia. D.L.J. would also like to acknowledge S. Rotkin for a stimulating discussion.