Theory of Metal-Insulator Transitions in Gated Semiconductors

Boris L. Altshuler[1,2] and Dmitrii L. Maslov[3]
[1]*NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540*
[2]*Physics Department, Princeton University, Princeton, New Jersey 08544*
[3]*Department of Physics, University of Florida, P.O. Box 118440 Gainesville, Florida 32611-8440*

(Received 25 August 1998)

It is shown that recent experiments indicating a metal-insulator transition in 2D electron systems can be interpreted in terms of a simple model, in which the resistivity is controlled by scattering at charged hole traps located in the oxide layer. The gate voltage changes the number of charged traps which results in a sharp change in the resistivity. The observed exponential temperature dependence of the resistivity in the metallic phase of the transition follows from the temperature dependence of the trap occupation number. The model naturally describes the experimentally observed scaling properties of the transition and the effects of magnetic and electric fields. [S0031-9007(98)08103-4]

PACS numbers: 71.30.+h, 73.40.Qv

Recently, a metal-insulator transition has been observed in low-density two-dimensional (2D) electronic systems—first in Si metal-oxide-semiconductor (MOS) structures [1–4] and later in other heterostructures [5–9]. It has been found that, when the density of 2D electrons n_s is below some critical value n_s^c, cooling causes an increase of the resistivity ρ, while at $n_s > n_s^c$ the resistivity *decreases* with temperature T, i.e., the system exhibits an unexpected metallic behavior. The insulating phase has been found to be rather usual and easy to describe in terms of variable range hopping [10]. However, the metallic phase is anomalous in at least three respects. (i) $\rho(T)$ dependence follows the exponential, i.e., $\rho(T) = \rho_0 + \rho_1 \exp(-T_0/T)$, rather than power-law form; (ii) ρ drops by about an order of magnitude when T changes in the range comparable to the Fermi energy ε_F of 2D electrons; (iii) the metallic state is quenched by the magnetic field.

Here, we are not going to discuss existing attempts [11–18] to interpret these experiments. (We found in Refs. [11–18] no satisfactory physical explanation of the substantial drop in the resistivity in a narrow temperature interval in an obviously nonsuperconducting system.) Instead, we propose a simple mechanism which seems to naturally explain all of the peculiarities mentioned above. We believe that our general idea can be applied to all gated semiconductors. However, here we concentrate on Si MOS structures, where the important characteristics of a two–dimensional electron gas (2DEG) and of defects are much better known than in other systems.

A typical n-Si MOS structure consists of a metallic gate, SiO_2 layer, and p-type Si substrate. A strong enough, positive gate potential attracts electrons which form an inversion layer at the SiO_2/Si interface. It is known [19] that, due to the oxygen deficit in the oxide, there is a substantial concentration of defects close to the interface, which are capable of trapping charges. Even in state-of-the-art devices, there are more than 10^{12} hole traps per cm^2, such as Si-Si weak bonds [19]. To introduce the idea of our mechanism, we assume all of the hole traps to be (I) characterized by the same energy of the electron level ε_t and (II) located at the same distance z from the interface. We shall abandon assumption (II) later on. The effects of a finite width of the trap band will be discussed elsewhere [20].

At $T = 0$, the trap charge (and spin) state is determined by the chemical potential μ of the 2DEG. For $\varepsilon_t > \mu$, the electron level is empty, i.e., a hole is trapped. The trap has a charge $+e$ and thus *causes strong scattering of 2D electrons.* It is crucial for our theory that the charge state of a trap can be changed by varying the gate voltage V_g. Indeed, the bigger V_g, the smaller $\varepsilon_t = \varepsilon_t(V_g)$. At $V_g = V_g^*(z)$ determined from $\varepsilon_t(V_g^*) = \mu$, the trap captures an electron (i.e., emits a hole) and is neutralized. Being neutral and remote from 2D electrons, the defect can no longer scatter them. Neutralization of the oxide charges reduces resistivity ρ and thus causes an insulator-to-metal transition. When T is high ($\gg|\varepsilon_t - \mu|$), roughly half of the traps are charged. As a result, ρ is rather high and depends weakly on both T and V_g. However, for $|\mu - \varepsilon_t| \leq T$ the density of charged traps behaves as $\exp[(\varepsilon_t - \mu)/T]$, resulting in the exponential $\rho(T)$ dependence [feature (i)]. The transition takes place for both degenerate and nondegenerate 2DEGs [feature (ii)]. Finally, the magnetic field effect (iii) can be attributed to the spin freeze-out of holes [21]: Zeeman splitting favors a spin-1/2 (charged) state with respect to the singlet (neutral) state of the defect.

It should be noted that here we neglect quantum interference of 2D electrons and thus do not attempt to describe the insulating phase. However, we see that, even in the classical case, $d\rho/dT$ can change sign due to the $\mu(T)$-dependence.

Let us now abandon assumption (II), i.e., take into account a broad distribution of distances z. In order to understand why such a distribution does not smear the transition, we consider the electrostatic energy of an electron in the oxide $\varepsilon_e(z)$. Given the total oxide thickness

d and its dielectric constant ϵ_{ox}, ε_e can be written as

$$-\varepsilon_e(z) = eV_g z/d + e^2/(2\epsilon_{ox} z) . \qquad (1)$$

Here the two terms represent the external electric field and the image force from the 2DEG, respectively (charges induced in the gate can be neglected provided that $z \ll d$). $\varepsilon_e(z)$ reaches its maximum ε_m at $z = z_m$, where

$$\varepsilon_m = -2\sqrt{eV_g \varepsilon_d} , \qquad z_m = d\sqrt{\varepsilon_d/eV_g} , \qquad \varepsilon_d \equiv e^2/(2\epsilon_{ox} d) . \qquad (2)$$

z_m can also be expressed through the mean distance $\bar{r}$ between 2D electrons:

$$z_m = \bar{r}/\sqrt{8} = a_B r_s/\sqrt{8} , \qquad (3)$$

where $a_B \equiv \bar{r}/r_s$ is the effective Bohr radius. Equation (3) follows from the relation between the 2DEG concentration $n_s = 1/\pi\bar{r}^2$ and the gate voltage: $en_s = \epsilon_{ox} V_g/4\pi d$ [22]. In order to have a meaning in a macroscopic theory, z_m has to exceed the screening radius of 2DEG (equal to $a_B/4$ for a Si(001) surface [23]). Therefore, in low-density devices ($r_s \gg 1/\sqrt{2}$), this length scale is quite legitimate.

Assuming that the double (hole) occupancy of a trap is impossible, the probability of a trap to be charged is

$$P_+(z) = \left[\frac{1}{2C} \exp\left(\frac{\mu - \varepsilon_e(z) - \varepsilon_t}{T} \right) + 1 \right]^{-1} , \qquad (4)$$

where $C = 1$. According to Eqs. (1) and (4), a homogeneous distribution of traps leads to a distribution of charges which is peaked at $z = z_m$, the width of the peak being

$$\delta z = d[T^2 \varepsilon_d (eV_g)^{-3}]^{1/4} = z_m [T^2 (eV_g \varepsilon_d)^{-1}]^{1/4} . \qquad (5)$$

For $d = 2000$ Å and $eV_g = 1$ eV, we get $\varepsilon_d = 1$ meV and $\varepsilon_m = 63$ meV, respectively, so that $\delta z/z_m \simeq \sqrt{T(K)}/18 \ll 1$, since $T \leq 5$ K. At $T = 5$ K, $z_m \approx 63$ Å and $\delta z \approx 8$ Å. This sharpness of the distribution peak in Eq. (4) manifests itself in a sharp metal-insulator transition, as V_g is varied.

How does a positive charge, separated by a distance $z \gg a_B$ from the 2DEG, affect the resistivity? It turns out that a bound localized state is formed with $\xi = z^{3/4} a_B^{1/4} < z$ and $\varepsilon_b = -e^2/(\epsilon^* z)$ being the localization length and energy of this state, respectively (ϵ^* is the effective dielectric constant of the 2DEG). The trap and bound electron form a dipole, which is oriented perpendicular to the 2DEG plane. For $z \simeq a_B r_s$ and $r_s \gg 1$, the (transport) scattering cross section $\Sigma(\varepsilon, z)$ of such a dipole for electrons with energy ε can be evaluated classically:

$$\Sigma(\varepsilon, z) = 2.74 (e^2 z^2/2\epsilon^* \varepsilon)^{1/3} . \qquad (6)$$

The Drude formula for the resistivity can be written as

$$\rho = (N_t/e^2 n_s)\sqrt{2m^* \bar{\varepsilon}}\, \Sigma(\bar{\varepsilon}, z_m) \int_0^d dz\, P_+(z) \left(\frac{z}{z_m} \right)^{2/3} , \qquad (7)$$

where N_t is the total volume concentration of the traps, m^* is the effective mass of the electrons, and $\bar{\varepsilon}$ has a meaning of their effective energy, which can be expressed through the 2DEG Fermi energy ε_F via

$$\bar{\varepsilon} = \varepsilon_F \left[\int_0^\infty \frac{d\varepsilon}{4T} \left(\frac{\varepsilon}{\varepsilon_F} \right)^{5/6} \cosh^{-2}\left(\frac{\varepsilon - \varepsilon_F}{2T} \right) \right]^{-6} . \qquad (8)$$

Equation (8) interpolates between two limits: $\bar{\varepsilon} \approx T/\Gamma^6(11/6) \approx 1.44T$ for $T \gg \varepsilon_F$, while in the opposite limit $\bar{\varepsilon} = \varepsilon_F$. For $T = \varepsilon_F$, the effective energy $\bar{\varepsilon} \approx 2T$.

In the saddle-point approximation, Eq. (7) reduces to

$$\rho = (h/e^2)\rho_0 \mathcal{R}(V_g, T) , \qquad (9a)$$

$$\rho_0 = 0.46\sqrt{r_s}\,(\epsilon^*/2\epsilon_{ox})^{1/6} (N_t \bar{r}/\pi n_s)\,(\bar{r}/d)^{2/3} , \qquad (9b)$$

$$\mathcal{R}(V_g, T) = \left(\frac{T^3 \bar{\varepsilon}}{\varepsilon_d^4} \right)^{1/6} \int_0^\infty \frac{dx}{[f(T)/2]\exp(x^2 + s) + 1} . \qquad (9c)$$

In Eq. (9c), we took into account the $\mu(T)$-dependence:

$$s = [\mu(0) - \varepsilon_m - \varepsilon_t]/T , \qquad f(T) = e^{[\mu(T) - \mu(0)]/T} . \qquad (10)$$

We have to consider two distinct cases: (A) Chemical potentials of the 2DEG and of the Si substrate coincide; (B) the 2DEG is disconnected from the substrate. A straightforward calculation gives

$$f_A(T) = (T/T_a)^{3/4} , \qquad f_B(T) = 1 - \exp(-\varepsilon_F/T) , \qquad (11)$$

where T_a is determined by the acceptor concentration [21]. Although case B is more likely to occur in a real device [24], we shall concentrate mostly on case A which exhibits a clear metal-insulator transition even in a classical model (see below).

The exponential part of the $\mathcal{R}(T)$-dependence disappears when s, Eq. (10), vanishes. This happens at $V_g = V_g^c = [\varepsilon_t - \mu(0)]^2/4e\varepsilon_d$ and thus V_g^c defines the transition point. Away from this point $\mathcal{R}$ behaves as

$$\mathcal{R}_{A,B} \left(\frac{T^3 \bar{\varepsilon}}{\varepsilon_d^4} \right)^{1/6} \times \begin{cases} \sqrt{\pi}\,\Omega , & \text{for } \Omega \ll 1 , \\ \ln^{1/2} \Omega , & \text{for } \Omega \gg 1 . \end{cases} \qquad (12)$$

where $\Omega \equiv [2/f(T)]e^{-s}$. The distance from the transition can be measured by $\delta \equiv (V_g - V_g^c)/V_g^c$. Provided $\delta^2 \ll 4T^2/(\varepsilon_d e V_g)$, variable s in Eq. (10) acquires a scaling form

$$s \approx \sqrt{\varepsilon_d e V_g^c}\,(\delta/T) \equiv v/t , \qquad t \equiv T/\varepsilon_d , \qquad v \equiv \delta\sqrt{eV_g^c/\varepsilon_d} . \qquad (13)$$

The $\mathcal{R}(T)$ dependence in the scaling region is shown in Fig. 1. For $v \gg t$, the system is in the "metallic" phase characterized by $\mathcal{R}_{A,B}$ exponentially decreasing with t. Because of the $\mu(T)$ dependence, $d\mathcal{R}_A/dt$ changes sign at some v slightly bigger than zero, exhibiting thus a metal-to-insulator transition. For larger negative v [not

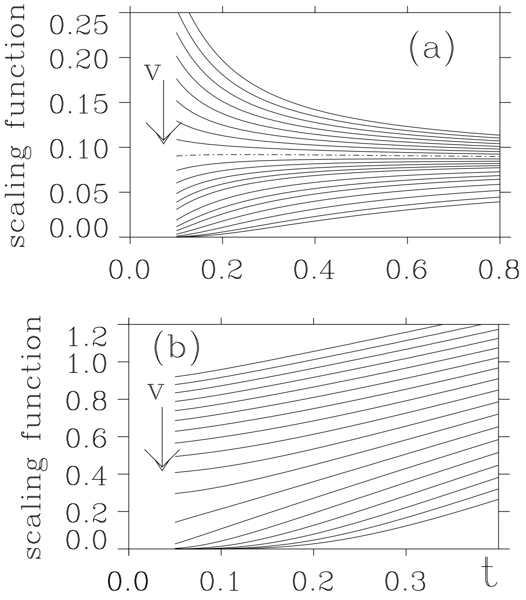

FIG. 1. Scaling function $\mathcal{R}$ [Eq. (9c)] vs dimensionless temperature t for several dimensionless gate voltages v [Eq. (13)]. v increases in the direction of the arrow. $\varepsilon_F/\varepsilon_d = 0.25$. (a) Case A: $v = -0.2-0.7$. $T_a/\varepsilon_d = 0.04$. Dot-dashed line indicates the transition. (b) Case B: $v = -1.4-0.7$.

shown in Fig. 1(a)], the $\mathcal{R}_A(t)$ dependence saturates. At $eV_g^c = 1$ eV, $\varepsilon_d = 1$ meV, and $T = 5$ K, we predict critical behavior for $|\delta| \leq 0.01$. This is consistent with experiments [1–4].

In case B, there are two distinct regions: exponentially decreasing and t-independent $\mathcal{R}_B$, the crossover between the two occurring for $|v| \simeq t$. Quantum interference effects should result in localization, converting T-independent ρ into an exponentially diverging one; the metal-insulator transition in this case will be discussed elsewhere [20].

At the transition, $\mathcal{R}_A \approx 0.1$ and $\mathcal{R}_B \approx 1$ (cf. Fig. 1). At the same time, $n_s \simeq 10^{11}$ cm^{-2} and $\bar{r} \simeq 100$ Å for $d = 2 \times 10^{-5}$ cm and $eV_g = 1$ eV. Estimating $N_t\bar{r} \simeq 10^{12}$ cm^{-2} and $r_s \simeq 10$, we obtain for the resistivities at the transition $\rho_A^c \simeq 0.1h/e^2$ and $\rho_B^c \simeq h/e^2$. These values are within the experimentally observed range [3(b)].

As Fig. 2 shows, the transition between the insulating and metallic phases in case A is very well defined, despite the fact that $\mathcal{R}$ does not solely depend on the scaling variable v/t. Closer inspection of the transition region (Fig. 2, inset) reveals, however, that the transition occurs over a finite range of v rather than at a single point.

Figure 3 depicts the (approximate) data collapse for $\mathcal{R}$ plotted as a function of $t/|v| = T/T_0$ with $T_0 = |\delta|\sqrt{\varepsilon_d eV_g^c}$. The inset of Fig. 3 demonstrates the "duality" feature, i.e., the symmetry between the resistivity ρ in the insulating phase and the conductivity σ in the metallic phase. Experimentally, a similar collapse was achieved for $T_0 \propto |\delta|^a$ [in the quantum phase transition theory (QPTT), $a = \nu z$]. In all of the experiments, except Ref. [8], a is close to 1, i.e., to our prediction.

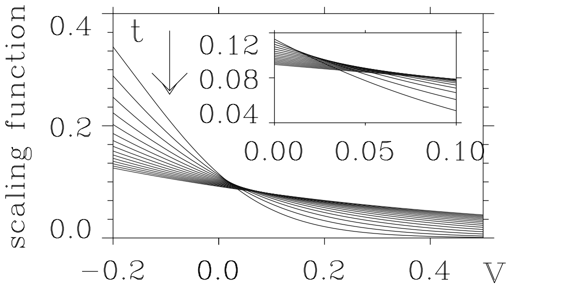

FIG. 2. Case A: Scaling function $\mathcal{R}$ [Eq. (9c)] vs dimensionless gate voltage v for dimensionless temperatures $t = 0.1-0.6$ [Eq. (13)]. t increases in the direction of the arrow. Inset: A blowup of the transition region.

An additional insight comes from non-Ohmic measurements. In Ref. [1(c)], the dependence of ρ on the source-drain voltage $V_{SD} = \mathcal{E}L$ (where $\mathcal{E}$ is the electric field, and L is the source-drain distance) was also found to be a scaling one, $\rho = \rho(\mathcal{E}/\mathcal{E}_0)$ with $\mathcal{E}_0 \propto |\delta|^b$. We believe that this $\mathcal{E}$ dependence can be attributed to simple heating. Indeed, the effective temperature of electrons T^* is determined by the energy balance. For strong enough electric field, i.e, when $T^* \gg T$, and for 2D electrons,

$$e\mathcal{E}\sqrt{D(T^*)\tau_{\mathrm{eph}}(T^*)} = (\pi/\sqrt{6})T^*. \qquad (14)$$

Here $D(T)$ is the diffusion constant of electrons at temperature T, and $\tau_{\mathrm{eph}}(T) \propto T^{-p}$ is the relaxation time of the electron temperature, which we assume to be determined by electron-phonon scattering [25]. One can check that, if $\rho(T)$ and $D(T)$ obey a scaling law $1/D(T) \propto \rho(T) = F(T/T_0)$, the $\mathcal{E}$ dependence of the resistivity is also scalinglike: $\rho = G(\mathcal{E}/\mathcal{E}_0)$, where $\mathcal{E}_0 = T_0^\alpha \propto |\delta|^{\alpha a}$ with $\alpha = 1 + p/2$ and function G is obtained by solving Eq. (14) for a given F. If $p = 3$ (as is the case for good metals), $\alpha = 2.5$. The experimental value of $\alpha = b/a$ [1] is $\simeq$2.25. This discrepancy can easily be explained by p being smaller than 3. On the other hand, QPTT predicts $\alpha = 1 + z^{-1}$, i.e., $\alpha = 2$ at $z = 1$. One can check that the strong heating

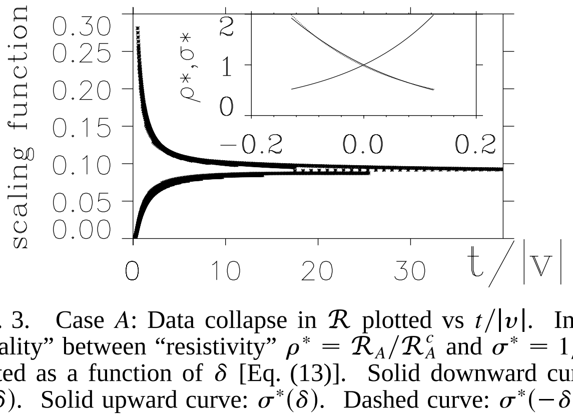

FIG. 3. Case A: Data collapse in $\mathcal{R}$ plotted vs $t/|v|$. Inset: "Duality" between "resistivity" $\rho^* = \mathcal{R}_A/\mathcal{R}_A^c$ and $\sigma^* = 1/\rho^*$ plotted as a function of δ [Eq. (13)]. Solid downward curve: $\rho^*(\delta)$. Solid upward curve: $\sigma^*(\delta)$. Dashed curve: $\sigma^*(-\delta)$.

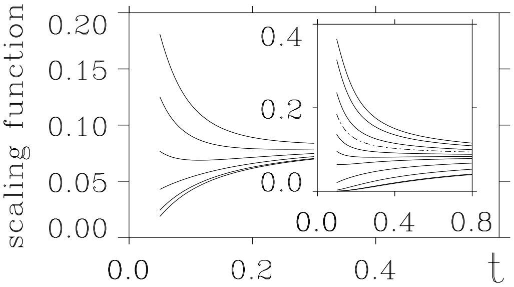

FIG. 4. Case A: Scaling function $\mathcal{R}$ vs temperature t for Zeeman splittings $E_Z = (0, 0.04, 0.08, 0.12, 0.16, 0.2) \times \varepsilon_d$ (E_Z increases from the bottom to the top curves). $v = 0.1$. Inset: The same as in Fig. 1(a) but for $E_Z = 0.15\varepsilon_d$.

regime is realized under the conditions of Ref. [1(c)], if $\tau_{\mathrm{eph}} > 0.1$ ms$/[\mathcal{E}(\mathrm{mV/cm})]^2$. Strong heating of a 2DEG has recently been observed in a Si metal-oxide-semiconductor-field-effect transistor device [26], similar in its parameters to that used in Refs. [1–4].

We now turn to the effect of a magnetic field. Consider a hole trap, e.g., a Si-Si weak bond [19]. Such a trap can find itself in one of the three states with energies E_i: $i = 1, 2, 3$. For $i = 1$, two electrons occupy the bond. This is supposed to be a neutral ($Q = 0$) singlet ($S = 0$) state. State 2 (3) has one spin down (up) electron on the bond. Accordingly, $Q = +1$ and $S = 1/2$ for both states 2 and 3. A magnetic field splits the doublet: $E_1 - E_{2(3)} = \varepsilon_t \pm E_Z$, where E_Z is the Zeeman splitting. As a result, at given T and V_g probability P_+ of finding a trap in a $Q = +1$ state increases with E_Z, i.e., with the magnetic field: One should substitute $C = \cosh(E_Z/T)$ instead of 1 into Eq. (4). This results in a magnetoresistivity demonstrated in Fig. 4.

We conclude by emphasizing that this paper suggests a mechanism of the sharp resistance drop with temperature and an illustration, rather than a theory, of the metal-insulator transition. The assumptions/predictions which we are proposing at this stage to check experimentally are the existence of hole traps in the proper energy interval with short enough times of charge transfer between these traps and a 2DEG. One can think about several scenarios of the transition which allow for inhomogeneous broadening of trap energy levels and involve quantum interference effects. These more realistic pictures will be discussed elsewhere [20].

We benefited greatly from discussions with D. Cobden, M. E. Gershenson, S. V. Kravchenko, and V. M. Pudalov. We would also like to thank I. L. Aleiner, L. I. Glazman, J. Graybeal, K. Muttalib, M. Reizer, and O. A. Starykh for their valuable comments. The work at Princeton University was supported by ARO MURI DAAG55-98-1-0270. D. L. M. acknowledges financial support from NSF DMR-970338.

[1] (a) S. V. Kravchenko *et al.*, Phys. Rev. B **50**, 8039 (1994); (b) **51**, 7038 (1995); (c) Phys. Rev. Lett. **77**, 4938 (1996).
[2] D. Simonian, S. V. Kravchenko, M. P. Sarachik, and V. M. Pudalov, Phys. Rev. B **55**, R13 421 (1997); Phys. Rev. Lett. **79**, 2304 (1997).
[3] (a) V. M. Pudalov, G. Brunthaler, A. Prinz, and G. Bauer, JETP Lett. **65**, 932 (1997); (b) *ibid.* **68**, 442 (1998).
[4] D. Popović, A. B. Fowler, and S. Washburn, Phys. Rev. Lett. **79**, 1543 (1997).
[5] K. Ismail *et al.*, cond-mat/9707061.
[6] P. T. Coleridge, R. L. Williams, Y. Feng, and P. Zawadzki, Phys. Rev. B **56**, R12 764 (1997).
[7] J. Lam, M. D'Iorio, D. Brown, and H. Lafontaine, Phys. Rev. B **56**, R12 741(1997).
[8] M. Y. Simmons *et al.*, Phys. Rev. Lett. **80**, 1292 (1998).
[9] Y. Hanein *et al.*, Phys. Rev. Lett. **80**, 1288 (1998); Phys. Rev. B **58**, R7520 (1998).
[10] B. I. Shklovskii and A. L. Efros, *Electronic Properties of Doped Semiconductors* (Springer-Verlag, Berlin, 1984).
[11] V. Dobrosavljević, E. Abrahams, E. Miranda, and S. Chakravarty, Phys. Rev. Lett. **79**, 455 (1997).
[12] C. Castellani, C. Di Castro, and P. A. Lee, Phys. Rev. B **57**, R9381 (1998).
[13] P. Phillips *et al.*, Nature (London) **395**, 253 (1998).
[14] D. Belitz and T. R. Kirkpatrick, Phys. Rev. B **58**, 8214 (1998).
[15] F.-C. Zhang and T. M. Rice, cond-mat/9708050.
[16] Song He and X. C. Xie, Phys. Rev. Lett. **80**, 3324 (1998).
[17] S. Chakravarty, S. Kivelson, C. Nayak, and K. Völker, cond-mat/9805383.
[18] V. M. Pudalov, JETP Lett. **66**, 175 (1997).
[19] T. Hori, *Gate Dielectrics and MOS ULSIs* (Springer-Verlag, Berlin, 1997).
[20] B. L. Altshuler and D. L. Maslov (unpublished).
[21] R. A. Smith, *Semiconductors* (Cambridge University Press, Cambridge, England, 1978).
[22] We incorporate the threshold voltage into V_g.
[23] T. Ando, A. B. Fowler, and F. Stern, Rev. Mod. Phys. **54**, 437 (1982).
[24] V. M. Pudalov (private communication).
[25] It is not quite clear how the authors of Ref. [16] estimated heating without specifying the relaxation mechanism.
[26] R. J. Zieve, D. E. Prober, and R. G. Wheeler, Phys. Rev. B **57**, 2443 (1998).