## On the Theory of Metal-Insulator Transitions in Gated Semiconductors

Boris L. Altshuler<sup>1,2</sup> and Dmitrii L. Maslov<sup>3</sup>

<sup>1</sup>NEC Research Institute, 4 Independence Way, Princeton, NJ 08540

<sup>2</sup>Physics Department, Princeton University, Princeton, NJ 08544

<sup>3</sup>Department of Physics, University of Florida, P. O. Box 118440, Gainesville, Florida 32611

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 effects of magnetic and electric fields.

Recently, a metal-insulator transition has been observed in low-density two-dimensional electronic systems – first in Si 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  $\rho$ , 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]. On the contrary, the metallic phase is anomalous in at least three respects: i) the  $\rho(T)$  - dependence follows the exponential, i.e.,  $\rho(T) = \rho_0 + \rho_1 \exp(-T_0/T)$  rather than power-law form; ii)  $\rho$  drops by about an order of magnitude when T changes in the range comparable to the Fermi energy  $\varepsilon_F$  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 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 2DEG and of defects are much better known than in other systems.

A typical *n*-Si MOS structure consists of a metallic gate, SiO<sub>2</sub> layer, and p-type Si substrate. Strong enough, positive gate potential attracts electrons which form an inversion layer at the SiO<sub>2</sub>/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-ofthe-art devices, there are more than  $10^{12}$  hole traps per cm<sup>2</sup>, such as Si-Si weak bonds [19]. To introduce the idea of our mechanism, we assume all of the hole traps to be  $\star$ ) characterized by the same energy of the electron level  $\varepsilon_t$ , and  $\star\star$ ) located at the same distance z from the interface. We shall abandon assumption  $\star\star$ ) later on. Effects of a finite width of the trap band will be discussed elsewhere.

At T=0, the trap charge (and spin) state is determined by the chemical potential  $\mu$  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 cannot scatter them any more. Neutralization of the oxide charges reduces resistivity  $\rho$  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,  $\rho$  is rather high and depends weakly on both T and  $V_g$ . On the contrary, 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 [20]: Zeeman splitting favors spin 1/2 (charged) state with respect to the singlet (neutral) state of the defect.

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

Let us now abandon assumption  $\star\star$ ), 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  $\epsilon_{ox}$ ,  $\varepsilon_e$  can be written as

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

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  $\varepsilon_m$  at  $z = z_m$ , where

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

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

$$z_m = \bar{r}/\sqrt{8} = a_B r_s/\sqrt{8},\tag{3}$$

where  $a_B \equiv \bar{r}/r_s$  is the effective Bohr radius. Eq.(3) follows from the relation between the 2DEG concentration  $n_s = 1/(\pi \bar{r}^2)$  and the gate voltage:  $en_s = \epsilon_{ox} V_g/(4\pi d)$ [21]. 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 Si (001) surface [22]). 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.(1,4), a homogeneous distribution of traps leads to a distribution of charges which is peaked at  $z = z_m$ , the width of the peak being

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

For d = 2000 Å and  $eV_g = 1$  eV, we get  $\varepsilon_d = 1$  meV and  $\varepsilon_m = 63$  meV, 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  $V_q$  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 the energy of this state, respectively ( $\epsilon^*$  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  $z \sim a_B \sqrt{r_s}$  and  $\sqrt{r_s} \gg 1$ , the (transport) scattering cross-section  $\Sigma(\varepsilon, z)$  of such a dipole for electrons with energy  $\varepsilon$  can be evaluated classically:

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

The Drude formula for the resistivity can be written as

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

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

Eq. (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  $\bar{\varepsilon} \approx 2T$ .



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...07.  $T_a/\varepsilon_d = 0.04$ . Dot-and-dashed line indicates the transition. (b) Case B. v = -1.4...07.

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

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

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

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

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

$$s = \left[\varepsilon_m - \varepsilon_t + \mu(0)\right]/T; f(T) = e^{\left[\mu(T) - \mu(0)\right]/T}.$$
 (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

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

where  $T_a$  is determined by the acceptor concentration [20]. Although case B is more likely to occur in a real device [23], we shall concentrate mostly on case A which exhibits a clear metal-insulator transition even in a classical model; see below.



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 blow-up of the transition region.

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. The asymptotic behavior of  $\mathcal{R}$  away from this point is

$$\mathcal{R}_{A,B} = \left(\frac{T^3 \bar{\varepsilon}}{\varepsilon_d^4}\right)^{1/6} \times \begin{cases} \sqrt{\pi} \Omega, \text{ for } \Omega \ll 1;\\ \ln^{1/2} \Omega, \text{ for } \Omega \gg 1. \end{cases}$$
(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  $\delta^2 \ll 4T^2/(\varepsilon_d eV_g)$ , variable s in Eq.(10) acquires a scaling form

$$s \approx \sqrt{\varepsilon_d e V_g^c}(\delta/T) \equiv v/t; t \equiv T/\varepsilon_d; v \equiv \delta \sqrt{e V_g^c/\varepsilon_d}.$$
 (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. Due to 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 shown in Fig. 1(a)], the  $\mathcal{R}_A(t)$ -dependence saturates. At  $eV_q^c = 1$  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 occuring for  $|v| \simeq t$ . Quantum interference effects should result in localization, converting T-independent  $\rho$  into an exponentially diverging one; the metal-insulator transition in this case will be discussed elsewhere.

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<sup>-2</sup> and  $\bar{r} \simeq 100$  Å for  $d = 2 \cdot 10^{-5}$  cm and  $eV_g = 1$  eV. Estimating  $N_t \bar{r} \simeq 10^{12}$  cm<sup>-2</sup> 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.



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  $\delta$  [Eq. (13)]. Solid downward:  $\rho^*(\delta)$ . Solid upward:  $\sigma^*(\delta)$ . Dashed:  $\sigma^*(-\delta)$ .

Figure 3 depicts the (approximate) data collapse for  $\mathcal{R}$  plotted as a function of  $t/|v| = T/T_0$  with  $T_0 = |\delta| \sqrt{\varepsilon_d e V_g^c}$ . The inset of Fig. 3 demonstrates the "duality" feature, i.e., the symmetry between the resistivity  $\rho$  in the insulating phase and the conductivity  $\sigma$  in the metallic one. 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.

An additional insight comes from nonohmic measurements. In Ref. [1]c, the dependence of  $\rho$  on the sourcedrain 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,

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

Here D(T) is the diffusion constant of electrons at temperature T, and  $\tau_{eph}(T) \propto T^{-p}$  is the relaxation time of the electron temperature, which we assume to be determined by electron-phonon scattering [24]. 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 a scaling-like:  $\rho = G(\mathcal{E}/\mathcal{E}_0)$ , where  $\mathcal{E}_0 = T_0^{\alpha} \propto |\delta|^{\alpha a}$ with  $\alpha = 1 + p/2$ , and where function G is obtained by solving Eq. (14) for given F. If p = 3 (as it is the case for good metals),  $\alpha = 2.5$ . 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 regime is realized under the conditions of Ref. [1]c, if

$$\tau_{eph} > 0.1 \text{ms} / \left[ \mathcal{E}(\text{mV/cm}) \right]^2.$$
 (15)

Strong heating of a 2DEG has recently been observed in a Si MOSFET device [25] similar in its parameters to that used in Refs. [1–4]



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. 1a but for  $E_Z = 0.15\varepsilon_d$ .

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  $P_+$  to find 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 benefited greatly from discussions with M. E. Gershenson and V. M. Pudalov. We would also like to thank I. L. Aleiner, L. I. Glazman, J. Graybeal, S. M. Kravchenko, K. Muttalib, M. Reizer, and O. A. Starykh for their valuable comments. The work of D. L. M. was supported by NSF DMR-970338.

- [1] a) S. V. Kravchenko et al., Phys. Rev. B 50, 8039 (1994);
  b) *ibid*, 51, 7038 (1995); c) Phys. Rev. Lett. 77, 4938 (1996).
- [2] D. Simonian, S. V. Kravchenko, and M. P. Sarachik, Phys. Rev. B 55, R13421 (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) cond-mat/98071077.
- [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. Zawadzkii, Phys. Rev. B 56, R12764 (1997).
- [7] J. Lam, M. D'Iorio, D. Brown, and H. Lafontaine, Phys. Rev. B 56 R12741 (1997).
- [8] M. Y. Simmons et al., Phys. Rev. Lett. 80, 1292 (1998).
- [9] Y. Hanein et al., Phys. Rev. Lett. 80, 1288 (1998); condmat/9805113.
- [10] B. I. Shklovskii and A. L. Efros, *Electronic Properties of Doped Semiconductors* (Springer-Verlag, Berlin, Heidelberg, 1984).
- [11] V. Dobroslavljević, 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., cond-mat/9709168.
- [14] D. Belitz and T. R. Kirkpatrick, cond-mat/9705023.
- [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, Heidelberg, 1997).
- [20] R. A. Smith, *Semiconductors* (Cambridge University Press, Cambridge, 1978).
- [21] We incorporate the threshold voltage into  $V_g$ .
- [22] T. Ando, A. B. Fowler, F. Stern, Rev. Mod. Phys. 54, 437 (1982).
- [23] V. M. Pudalov, private communication.
- [24] It is not quite clear how the authors of Ref. [16] estimated heating without specifying the relaxation mechanism.

[25] R. J. Zieve, D. E. Prober, and R. G. Wheeler, Phys. Rev. B 57, 2443 (1998).