Double-dot charge transport in Si single-electron/hole transistors

L. P. Rokhinson,[a] L. J. Guo,[b] S. Y. Chou, and D. C. Tsui
Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544

(Received 11 October 1999; accepted for publication 24 January 2000)

We studied transport through ultrasmall Si quantum-dot transistors fabricated from silicon-on-insulator wafers. At high temperatures, $4<T<100$ K, the devices show single-electron or single-hole transport through the lithographically defined dot. At $T<4$ K, current through the devices is characterized by multidot transport. From the analysis of the transport in samples with double-dot characteristics, we conclude that extra dots are formed inside the thermally grown gate oxide which surrounds the lithographically defined dot.
[S0003-6951(00)03412-4]

Recent advances in miniaturization of Si metal–oxide–semiconductor field-effect transistors (MOSFETs) brought to light several issues related to the electrical transport in Si nanostructures. At low temperatures and low source-drain bias Si nanostructures do not follow regular MOSFET transconductance characteristics but show rather complex behavior, suggesting transport through multiply connected dots. Even in devices with no intentionally defined dots (like Si quantum wires[1] or point contacts)[2] Coulomb blockade oscillations were reported. In the case of quantum wires, formation of tunneling barriers is usually attributed to fluctuations of the thickness of the wire or of the gate oxide. However, formation of a dot in point contact samples is not quite consistent with such explanation. Recently in an elegant experiment with both n^+ and p^+ source/drain connected to the same Si point contact Ishikuro and Hiramoto[3] have shown that the confining potential in unintentionally created dots is similar for both holes and electrons. However, there is no clear picture where and how these dots are formed.

In this work we analyze the low temperature transport through ultrasmall lithographically defined Si quantum dots. While at high temperature $4<T<100$ K we observe single-electron tunneling through the lithographically defined dot, at $T<4$ K transport is found to be typical for a multidot system. We restrict ourselves to the analysis of samples with double-dot transport characteristics. From the data we extract electrostatic characteristics of both the lithographically defined and the extra dots. Remarkably, transport in some samples cannot be described by tunneling through two dots connected in sequence but rather reflects tunneling through dots connected in parallel to both source and drain. Taking into account the geometry of the samples we conclude that extra dots should be formed within the gate oxide. Transport in p- and n-type samples are similar, suggesting that the origin of the confining potential for electrons and holes in these extra dots is the same.

The samples are MOSFETs fabricated from a silicon-on-insulator wafer. The top silicon layer is patterned by an electron-beam lithography to form a small dot connected to wide source and drain regions, see schematic in Fig. 1(a). Next, the buried oxide is etched beneath the dot transforming it into a free-standing bridge. Subsequently, 40 or 50 nm of oxide are thermally grown which further reduce the size of the dot. Polysilicon gate is deposited over the bridge with the dot as well as over the adjacent regions of the source and drain. It is important to note that in this type of device the gate not only controls the potential of the dot but also changes the dot-source and dot-drain barriers. Finally, the uncovered regions of the source and drain are n-type or p-type doped. More details on samples preparation can be found in Ref. 4. Totally, about 30 hole and electron samples have been studied. Here we present data from two samples with hole (H5A) and electron (E5-7) field-induced channels.

A scanning electron microscopy (SEM) investigation of test samples, Fig. 1(b), reveals that the lithographically defined dot in the Si bridge is 10–40 nm in diameter and the distance between narrow regions of the bridge is ~ 70 nm. Taking into account the oxide thickness we estimate the gate capacitance to be 0.8–1.5 aF.

In most of our samples (with both n and p channel) we see clear Coulomb blockade oscillations with a period $\Delta V_{g1}=100-160$ mV up to ~ 100 K. A typical charge addition spectra is plotted in Figs. 2 and 3 for samples H5A and E5-7. In H5A the spectrum is almost periodic as a function of the gate voltage V_g at $T>4$ K with the period $\Delta V_{g1} \approx 130$ mV. Assuming that each peak corresponds to an addition of one hole into the dot we calculate the gate capacitance $C_{g1}=e/\Delta V_{g1}=1.2$ aF, which is within the error bars for the capacitance estimated from the sample geometry. The line shape of an individual peak can be described[5] by $G \propto \cosh^{-2}[(V_g-V_g^i)/2.5\alpha k_B T]$, where V_g^i is the peak position and coefficient $\alpha=C_{\text{total}}/eC_g$ relates the change in the V_g to

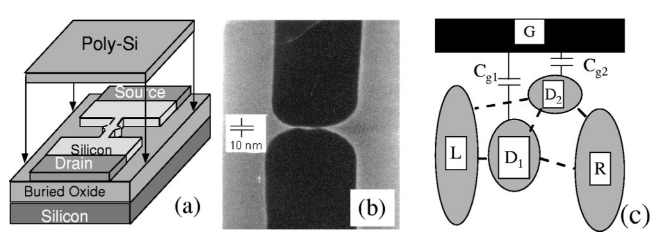

FIG. 1. (a) Schematic of the device structure, (b) SEM micrograph of a device, and (c) schematic view of two dots D_1 and D_2 connected to source and drain contacts L and R. G represents a gate electrode and C_{g1} and C_{g2} are gate capacitances. Dashed lines represent possible tunneling barriers.

[a]Electronic mail: leonid@ee.princeton.edu
[b]Present address: Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109.

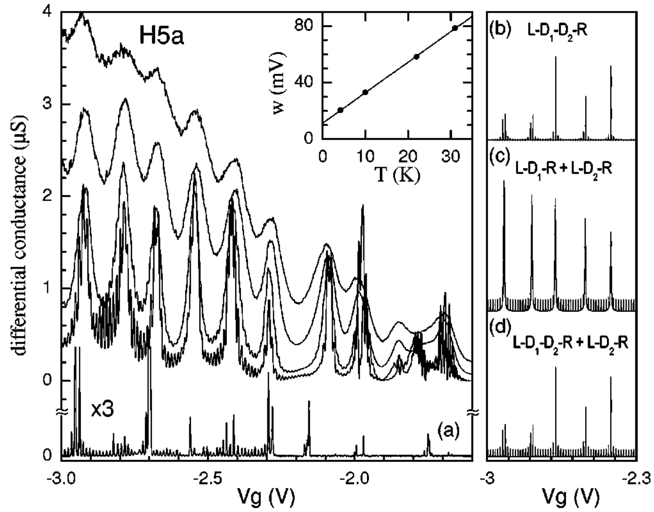

FIG. 2. (a) Differential conductance in the hole quantum dot sample H5A is shown as a function of the gate voltage V_g for $T=31$, 22, 10, 4.2, and 0.3 K (from top to bottom). The trace at the lowest temperature 0.3 K has been taken in a separate cooldown. In the inset peak width w vs T is plotted for peaks between $-3.0<V_g<-2.2$ V. (b)–(d) Modeling of the total conductance at $T=0.3$ K assuming that the two dots are connected (b) in series, (c) in parallel, or (d) mixed.

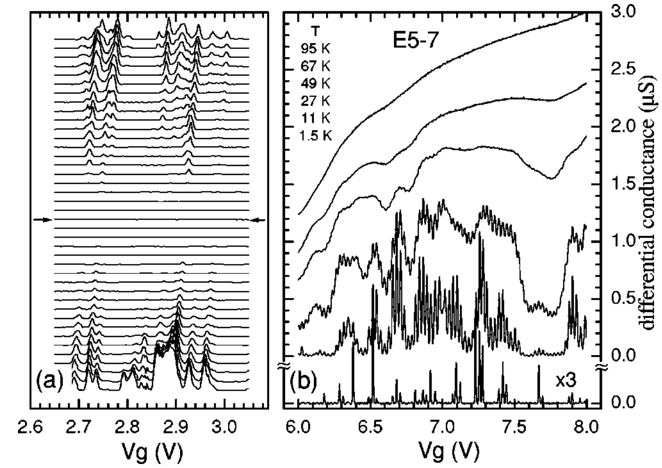

FIG. 3. Differential conductance in an electron quantum dot sample E5-7 is plotted as a function of the gate voltage V_g for (a) different dc source-drain bias V_b and (b) different temperatures. In (a) each curve is measured at different V_b from -20 (bottom curve) to 20 mV (top curve) at $T=1.5$ K. Arrows indicate the curve with $V_b=0$. All curves are offset by 0.5 μS. Data in (b) is taken at zero bias. The excitation voltage is 100 μV.

the shift of the energy levels in the dot relative to the Fermi energy in the contacts. This expression is valid if both coupling to the leads Γ and single-particle level spacing ΔE are small: $\Gamma, \Delta E \ll k_B T \ll e^2/C_{\text{total}}$. We fit the data for H5A with $\Sigma_i \cosh^{-2}[(V_g - V_g^i)/w]$ in the range $-3.0<V_g<-2.2$ V and the extracted w is plotted in the inset in Fig. 2. From the linear fit $w=11.3+2.2$ T (mV) we find the coefficient $\alpha=10$ (mV/meV), thus the Coulomb energy is ≈ 13 meV and the total capacitance $C_{\text{total}}=12.3$ aF. The main contribution to C_{total} comes from dot-to-lead capacitances (an estimated self-capacitance is a few attofarads). The extrapolated value of w at zero temperature provides an estimate for the level broadening $\Gamma \approx 1$ meV.

At $T<4$ K oscillations with another period, much smaller than ΔV_{g1}, appear as a function of V_g. The small period is in the range $\Delta V_{g2}=8-25$ mV in different devices ($\Delta V_{g2}=11.8$ mV for the sample in Fig. 2). This small period is due to a single-hole tunneling through a second dot and the corresponding gate capacitance $C_{g2}=e/\Delta V_{g2}=6-20$ aF. However, there is no intentionally defined second dot in our devices. Later we first analyze the experimental results and then discuss where the second dot can be formed.

At low temperatures and small gate voltages (close to the turn-on of the device at high temperatures) current is either totally suppressed, as in E5-7 at $V_g<3.5$ V, Fig. 3(a), or there are sharp peaks with no apparent periodicity, as in H5A at $V_g>-2.3$ V, Fig. 2. Both suppression of the current and ''stochastic Coulomb blockade'' [6] are typical signatures of tunneling through two sequentially connected dots. The nonzero conductance can be restored either by raising the temperature (Fig. 2) or by increasing the source-drain bias V_b [Fig. 3(a)]. In both cases, G is modulated with ΔV_{g1} and ΔV_{g2}, consistent with sequential tunneling. We conclude that in these regime the two dots are connected in series $L-D_1-D_2-R$ [see schematic in Fig. 1(c)].

At larger gate voltages ($V_g>6$ V for E5-7 and $V_g<-2.3$ V for H5A) current is not suppressed even at the lowest temperatures. However, the G pattern is different in the H5A and E5-7 samples. In H5A, the oscillations with ΔV_{g2} have approximately the same amplitude (except for the sharp peaks which are separated by approximately ΔV_{g1}), while in E5-7 the amplitude of the fast oscillations is modulated by ΔV_{g1}. Also, the dependence of the amplitude of the fast modulations on the average conductance $\langle G \rangle$ is different: in H5A the amplitude is almost $\langle G \rangle$ independent, while in E5-7 it is larger for larger $\langle G \rangle$.

Nonvanishing periodic conductance at low temperatures requires that the transport is governed by the Coulomb blockade through only one dot D_2. That can be achieved either if both barriers between the contacts and the D_2 become transparent enough to allow substantial tunneling or if the strong coupling between the main dot D_1 and one of the leads results in a nonvanishing density of states in the dot at $T=0$. If we neglect coupling between the dots, in the former case the total conductance is approximately the sum of two conductances, $G_{\text{parallel}} \approx G_1+G_2$, where G_1 is conductance through the main dot $L-D_1-R$ and G_2 is conductance through the second dot $L-D_2-R$. This case is modeled in Fig. 2(c) using experimentally determined parameters of sample H5A. From the analysis of high-temperature transport we found that the zero-temperature broadening of D_1 peaks $\alpha\Gamma \approx 10\,\text{mV} \approx \Delta V_{g2} \ll \Delta V_{g1}=130$ mV and that G should be exponentially suppressed between D_1 peaks at $T=0.3$ K if the dots are connected in series $L-D_1-D_2-R$, Fig. 2(b). The best description of the low temperature transport at $-3.0<V_g<-2.3$ V in H5A is achieved if we assume that there are two conducting paths in parallel: through the extra dot $L-D_2-R$ and through both dots together $L-D_1-D_2-R$, Fig. 2(d).

In the latter case, the dots are connected in series $L-D_1-D_2-R$. At high V_g the barrier between L and D_1 is reduced giving rise to a large level broadening Γ. The total conductance is $G_{\text{series}} \approx G_{\text{BW}} G_2/(G_{\text{BW}}+G_2)$, where G_2 is the Coulomb blockade conductance through D_2 alone and $G_{\text{BW}}=2e^2/h\,\Gamma^2/(\Gamma^2+\delta E^2)$ is the Breit–Wigner conductance through D_1 and $\delta E=(V_g-V_g^i)/\alpha$. In this case G_{series}

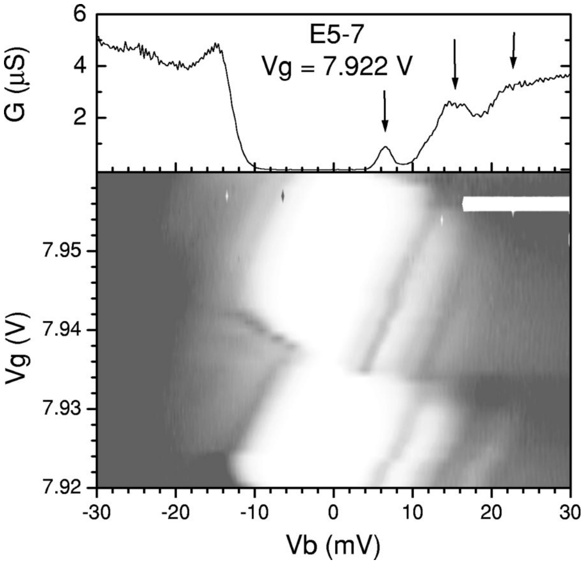

FIG. 4. Differential conductance on a gray scale as a function of both V_g and V_b. A single trace at $V_g = 7.922$ is shown at the top. Arrows indicate onset of the tunneling of 1, 2, and 3 electrons simultaneously.

is following G_{BW} and is modulated by G_2. Moreover, if we assume that the amplitude of G_2 is not a strong function of V_g, the amplitude of G_{series} modulation will be a function of G_{BW}, namely the larger G_{BW} the larger the amplitude of the modulation of the total conductance. This model of two dots in series with one being strongly coupled to the leads is in qualitative agreement with the data from sample E5-7.

Nonequilibrium transport through E5-7 is shown in Fig. 4 with a single G vs V_b trace at a fixed V_g shown at the top of the figure. White diamond-shaped Coulomb blockade regions are clearly seen on the gray-scale plot. Peaks in G at positive bias are due to asymmetry in the tunneling barriers:[7] at negative biases tunneling to the dot is slower than tunneling off the dot and only one extra electron occupies the dot at any given time, thus only one peak, corresponding to the onset of the current, is observed (we have not seen any features due to the size quantization, which is not surprising if we take into account the large number of electrons in this dot). At positive biases current is limited by the time the electron spends in the dot before it tunnels out. In this regime an extra step in the current–voltage characteristic (and a corresponding peak in its derivative G) is observed every time one more electron can tunnel into the dot. These peaks, marked with arrows, are separated by the charging energy $U_c = e\Delta V_b = 8$ meV.

Electrostatic parameters of the D_2 dot can be readily extracted from Fig. 4. The source, drain and gate capacitances are 8.5, 2.7, and 6.4 aF and the corresponding charging energy is ≈ 9 meV. The charging energy of ≈ 11 meV is obtained by analyzing Fermi–Dirac broadening of the conductance peaks as a function of temperature and the period of oscillations. The fact that it requires the application of $V_b = 10$ mV to lift the Coulomb blockade means that in the Coulomb blockade regime all the bias is applied across the second dot, consistent with large conductance through D_1.

Where does the second dot reside? One possibility is that the silicon bridge, containing the lithographically defined dot, breaks up at low temperatures as a result of the depletion due to variations of the bridge thickness and fluctuations in the thickness of the gate oxide, or due to the field induced by ionized impurities. However, in this case C_{g2} should be less than C_{g1}. In fact, if we assume that the thickness of the thermally grown oxide is uniform, the gate capacitance of the largest possible dot in the channel cannot be larger than 1.5 aF. Also, if at low temperatures the main dot would split into two or more dots we should see the change in the period of the large oscillations,[8] inconsistent with our observations.

Another possibility is that the dot is formed in the contact region adjacent to the bridge. Given that the oxide thickness is 40 nm, the second dot diameter should be ≈ 100 nm. We measured two devices which have 30 nm wide and 500 nm long channels, fabricated using the same technique as the dot devices. Both samples show regular MOSFET characteristics down to 50 mK. Thus, it is unlikely that a dot is formed in the wide contact regions of the device. Even if such a dot was formed occasionally in some device by, for example, randomly distributed impurities, it is unlikely that dots of approximately the same size would be formed in all samples. Another argument against such a scenario is that if the second dot is formed inside one of the contact regions, it cannot be coupled to the other contact to provide a parallel conduction channel, as in sample H5A.

Thus, the second dot should reside within the gate oxide, which surrounds the lithographically defined dot. Some traps can create confining potential in both conduction and valence bands, for example P_b center has energy levels at $E_c - 0.3$ and $E_v + 0.3$ eV. Several samples show a hysteresis during large gate voltage scans accompanied by sudden switching. This behavior can be attributed to the charging discharging of traps in the oxide. If such a trap happens to be in a tunneling distance from both the lithographically defined dot and a contact, or the trap is extended from one contact to the other, it may appear as a second dot in the conductance.

To summarize our results, we performed an extensive study of a large number of Si quantum dots. We found that all devices show multidot transport characteristics at low temperatures. From the data analysis, we arrived at the conclusion that at least double-dot behavior is caused not by the depletion of the silicon channel but by additional transport through traps within the oxide.

The authors acknowledge the support from ARO, ONR, and DARPA.

[1] Y. Nakajima, Y. Takahashi, S. Horiguchi, K. Iwadate, H. Namatsu, K. Kurihara, and M. Tabe, Jpn. J. Appl. Phys., Part 1 **34**, 1309 (1995); T. Hiramoto, H. Ishikuro, T. Fujii, G. Hashiguchi, and T. Ikoma, *ibid.* **36**, 4139 (1997); H. Ishikuro, T. Fujii, T. Saraya, G. Hashiguchi, T. Hiramoto, and T. Ikoma, Appl. Phys. Lett. **68**, 3585 (1996); R. A. Smith and A. H. Ahmed, *ibid.* **71**, 3838 (1997).
[2] H. Ishikuro and T. Hiramoto, Appl. Phys. Lett. **71**, 3691 (1997).
[3] H. Ishikuro and T. Hiramoto, Appl. Phys. Lett. **74**, 1126 (1999).
[4] E. Leobandung, L. Guo, Y. Wang, and S. Y. Chou, Appl. Phys. Lett. **67**, 938 (1995).
[5] I. O. Kulik and R. I. Shekhter, Zh. Eksp. Teor. Fiz. **68**, 623 (1975) [Sov. Phys. JETP **41**, 308 (1975)]; C. W. J. Beenakker, Phys. Rev. B **44**, 1646 (1991).
[6] I. M. Ruzin, V. Chandrasekhar, E. I. Levin, and L. I. Glazman, Phys. Rev. B **45**, 13469 (1992).
[7] B. Su, V. J. Goldman, and J. E. Cunningham, Phys. Rev. B **46**, 7644 (1992).
[8] F. R. Waugh, M. J. Mar, R. M. Westervelt, K. L. Campman, and A. C. Gossard, Phys. Rev. Lett. **75**, 705 (1995).